计算机科学
COMPUTER SCIENCE

基于知识标注平台的水利枢纽工程知识图谱构建及应用

张军珲, 昝红英, 欧佳乐, 阎子悦, 张坤丽

引用本文

张军珲, 昝红英, 欧佳乐, 阎子悦, 张坤丽. 基于知识标注平台的水利枢纽工程知识图谱构建及应用[J]. 计算机科学, 2024, 51(11): 255-264.

ZHANG Junhui, ZAN Hongying, OU Jiale, YAN Ziyue, ZHANG Kunli. Knowledge Annotation Platform-based Knowledge Graph Construction and Application for Water Conservancy Hub Projects [J]. Computer Science, 2024, 51(11): 255-264.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100079

基于知识标注平台的水利枢纽工程知识图谱构建及应用

张军珲[1,2]　昝红英[1]　欧佳乐[1]　阎子悦[1]　张坤丽[1]
1 郑州大学计算机与人工智能学院　郑州 450001
2 黄河勘测规划设计研究院有限公司　郑州 450003
(zhang_jh@yrec.cn)

摘　要　大量水利异构数据的产生,为领域知识图谱的构建及应用提供了场景,但也导致了水利知识图谱构建过程的差异。针对现有水利知识图谱构建流程复杂的问题,提出了一套有效的基于知识标注平台的水利知识图谱构建流程。以小浪底水利枢纽工程知识的智能应用为例,使用该枢纽的工程数据,应用提出的流程在水利领域构建水利枢纽工程知识图谱(Water Conservancy Hub Project Knowledge Graph,WCHP-KG)。首先以小浪底水利枢纽工程为中心,依据行业术语标准和现有词汇表,制定了概念分类和关系描述体系,形成了 WCHP-KG 的模式层。通过 BiLSTM-CRF 和序列标注模型,在水利专家的指导下,使用知识标注平台对非结构化文本进行了半自动标注和人工校对,实现了知识融合,进而构建了 WCHP-KG 的数据层。结果表明 WCHP-KG 涵盖了 43 种水利实体以及 110 种实体关系。经过实践验证,构建的 WCHP-KG 为小浪底水利枢纽工程的相关应用提供了有力的结构化知识基础,为工程决策和管理提供了可靠的参考依据,进而证明了所提构建流程的有效性。未来将进一步扩展 WCHP-KG 和完善水利知识图谱的构建流程,以适应更多的应用场景和领域需求。

关键词:异构数据;领域知识图谱;知识图谱构建;水利枢纽;知识标注平台
中图分类号　TP391

Knowledge Annotation Platform-based Knowledge Graph Construction and Application for Water Conservancy Hub Projects

ZHANG Junhui[1,2],ZAN Hongying[1],OU Jiale[1],YAN Ziyue[1] and ZHANG Kunli[1]
1 College of Computer and Artificial Intelligence,Zhengzhou University,Zhengzhou 450001,China
2 Yellow River Engineering Consulting Co.,Ltd,Zhengzhou 450003,China

Abstract　The generation of a significant volume of heterogeneous data in water resources has facilitated the creation and utilization of domain knowledge graphs,but it has led to discrepancies in the construction processes of these graphs. To address the complexities involved in building water resources knowledge graphs,an efficient approach based on a knowledge annotation platform is proposed. Taking the intelligent application of knowledge in Xiaolangdi Water Conservancy Hub project as an example,using the engineering data of the hub,the proposed method is applied to construct a water conservancy hub project knowledge graph(WCHP-KG) in the field of water conservancy. Firstly,focusing on the Xiaolangdi Water Conservancy Hub project,a construction for conceptual classification and relationship description is established based on industry terminology and existing vocabularies,forming the pattern layer of WCHP-KG. Through BiLSTM-CRF and sequence labeling models,under the guidance of water conservancy experts,a knowledge annotation platform is used to semi-automatically annotate and manually proofread unstructured texts,achieving knowledge fusion and constructing the data layer of WCHP-KG. Results indicate that WCHP-KG covers 43 water conservancy entities and 110 entity relationships. Through practical validation,the proposed WCHP-KG provides a solid structured knowledge base for applications related to the Xiaolangdi Water Conservancy Hub project,and provides a reliable reference for engineering decision-making and management,validating the efficacy of the proposed construction method. In the future,WCHP-KG will be further expanded and the construction process will be improved to meet the needs of more application scenarios and fields.

Keywords　Heterogeneous data,Domain knowledge graph,Knowledge graph construction,Water conservancy hub,Knowledge annotation platform

到稿日期:2023-11-13　返修日期:2024-04-28
基金项目:国家社会科学基金重大项目(21&ZD338)
This work was supported by the Major Program of the National Social Science Foundation of China(21&ZD338).
通信作者:昝红英(iehyzan@zzu.edu.cn)

1 引言

Google公司于2012年5月正式提出了知识图谱(Knowledge Graph,KG)的概念,旨在为下一代智能搜索引擎提供服务[1]。该概念是通用的语义知识描述框架,类似于语义网络。其中,结点表示实体或概念,边表示它们之间的语义关系[2]。这项技术能从大量文本中提取结构化知识,支持多个领域的下游任务,借助领域知识推动智能发展。在智慧水利建设领域,知识图谱也有着广泛应用。一方面,通过构建水利知识库深度挖掘水利数据的关联并预测其规律;另一方面,通过图谱检索和推理检索[3],充分利用储存于水利知识图谱中的信息来创建水利智能问答系统。

自20世纪70年代以来,我国便将智能水利建设和水利信息资源整合视为重要任务[4]。2019年12月,水利部发布了全国水利一张图的新版本[5],以促进水利信息资源的内部整合与共享,推动水利业务的协同发展,广泛推广智能应用。2021年水利部发布了《智慧水利建设指导意见》和《"十四五"智慧水利建设实施方案》,旨在积极推进智慧水利建设,充分利用积累的水利数据提升管理水平。同时,研究者们深入研究了知识图谱在水利领域的应用,通过知识图谱对已有的领域知识进行结构化的整合。此外,还提出利用知识图谱技术整合水利信息资源,实现智能数据检索和构建智能问答系统。这些研究和实践展现了水利领域知识图谱应用的广泛前景,但不同的水利知识图谱构建形式也使得彼此间的互用存在壁垒。

黄河小浪底水利枢纽工程是黄河干流上的一座集减淤、防洪、防凌、供水灌溉、发电等为一体的大型综合性水利工程,是治理开发黄河的关键性工程。面对这一关键性工程,智慧水利建设旨在从丰富的水利数据中提取并分类信息,构建知识体系从而为智能监控与智慧水利问答系统提供基础支持。本文基于小浪底水利枢纽的实际情况,从头梳理了水利知识图谱的构建方法,并提出了一套有效的基于知识标注平台的构建流程,将其应用于本文提出的面向小浪底的水利枢纽工程知识图谱(Water Conservancy Hub Project Knowledge Graph,WCHP-KG)构建。首先通过详细分析小浪底水利枢纽工程在业务规则与工程安全领域的内容,进而提炼出适用的知识图谱实体体系。然后通过多轮人工标注,系统性地将非结构化水利数据逐步转化为结构化标注,从而逐步完善水利知识体系。本文的主要贡献如下:

1)提出了一套有效的基于知识标注平台的水利知识图谱构建流程。

2)基于小浪底水利枢纽工程数据,应用提出的构建流程构建了WCHP-KG,将核心领域数据深度融合,构建起适用于业务管理和工程监测的知识推理与问答能力的水利枢纽工程知识图谱。

本文第2章介绍了背景知识及研究工作;第3章介绍了基于知识标注平台的水利知识图谱构建流程中使用到的方法与模型;第4章描述了该方法具体应用下的WCHP-KG构建过程;第5章描述了实体抽取模型的相关实验;第6章详述了WCHP-KG的实际应用;最后总结全文并展望未来。

2 相关工作

2.1 水利知识背景

在水利领域中,知识的涵盖范围极为广泛,涉及多种数据源的丰富信息。这些数据类型包括结构化数据(如水利业务数据),以及半结构化和非结构化数据(如与水利学科相关的知识文本以及来自互联网的数据)。

Duan等[6]对水利领域的知识进行了全面而深入的探究,部分分析结果如表1所列。该研究将水利领域的知识分为两种:水利事实知识和水利认识知识。水利事实知识是关于水利的各种实体、属性及它们之间关系的综合性描述,可用于准确描述河流、水利设施、管理机构等实际情况,为了解水利事实提供了基础。水利认识知识是水利综合知识的框架性描述,建立在不同水利学科的主题基础上,包括水利领域的概念和方法,为理解和解释水利现象提供了理论支持,也构成了水利事实知识的运作基础。基于这两类水利数据,研究者们积极开发信息服务平台,以实现有效组织、精准管理和全面水利信息服务。例如利用气象和水文观测数据,创建了水文模拟预测系统[7],该系统能预测河流水位变化和洪水风险,为防洪减灾和水资源调配提供科学依据。此外,基于水质数据的信息服务平台也在迅速发展[8],通过监测溶解氧、浊度、PH值等参数,实时监测水体健康状况进行水质风险评估,保障饮用水安全,保护环境。此外,通过遥感技术和地理信息系统数据建立的水资源遥感监测平台[9],基于卫星图像和地理信息数据全面监测水域分布和水资源利用。

表1 水利知识的描述

Table 1 Description of water conservancy knowledge

知识类别	描述对象	概念
水利事实类知识	自然对象	水利研究与管理中的天然对象
	工程对象	为控制和调配自然界的水体而修建的各类工程设施
	社会对象	涉水组织机构及进行水利研究和管理的人
水利认识类知识	安全鉴定	各类水利工程安全检查的结果判定,对判定结果进行详细的总结
	风险隐患	水利枢纽各项工程中存在的风险隐患整理汇总
	安全会商	工程施工中进行的聚集磋商,一般为双方或多方共同商量

为了有效组织水利知识并提供全面高效的水利知识服务,本文在前人研究的基础上结合多学科知识实现了对不同类型水利知识的关联,通过提出的基于知识标注平台的知识图谱构建流程,以小浪底水利枢纽工程的业务和安全检测数据为基础,最终构建了面向该水利枢纽工程的知识图谱,为小浪底水利枢纽区域的水资源智能管理和水利知识服务奠定了基础;同时,该方法的提出也为水利知识图谱的构建提供了一套有效的流程。

2.2 知识抽取

2.2.1 实体抽取

实体抽取(Entity Extraction),也称为实体识别,指从

文本中识别和提取出具体类型的命名实体。在实体识别的最初研究中，一般采用基于规则与词典的方法从文本中识别提取命名实体，常依靠领域专家指导及人工构造的规则及词表[10]。该方法构造的规则缺乏灵活性，且当新数据加入时往往需要重构部分内容，但在指定领域中，抽取的实体具备较高的准确性。

基于机器学习的实体识别采用监督学习方法训练模型，以识别实体。Liu 等[11]使用 K-最近邻（K-Nearest Neighbor，KNN）与线性条件随机场（Conditional Random Field，CRF）对推特数据进行实体识别，之后 KNN 和 CRF 逐渐成为实体识别模型的两种主要方法。例如，结合了 BiLSTM 和 CRF 优势的 BiLSTM-CRF[12]，能很好地处理序列数据中的标签依赖关系；Character-Based BiLSTM-CRF[13] 则是在 BiLSTM-CRF 的基础上利用了上下文信息与字符级别的表示。

2.2.2 关系抽取

关系抽取（Relation Extraction）指从文本中识别并提取出实体之间的关系或者实体与属性值之间的关系。目前在深度学习领域，关系抽取方法经历了多次演进和创新[14]。根据可用的标注数据集的规模，可以将关系抽取方法分为三大类别：监督学习、半监督学习，以及远程监督学习。

监督学习方法是最传统和直接的关系抽取方法之一。在监督学习中，模型使用带有标签的训练数据来学习实体对之间的关系，这些标签表示实体对之间的具体关系类型。监督学习方法主要基于卷积神经网络（CNN）[15]、循环神经网络（RNN）[16]以及长短期记忆（LSTM）[17]网络等深度学习模型及相关变体。

半监督学习方法尝试充分利用有标签数据和无标签数据来提升关系抽取模型的性能。这些方法通常包括半监督的迁移学习[18-19]和自监督学习策略[20-21]。利用大规模的无标签数据，半监督学习方法可以在数据稀缺的情况下训练出更鲁棒的关系抽取模型。

远程监督是一种创新的方法，它使用已知的实体-关系对作为监督信号，然后通过双向递归神经网络，如双向长短时记忆（Bidirectional Long Short-Term Memory，BiLSTM）[22]网络，来提取未标记文本中实体对之间的关系，可以克服标记数据不足的问题，但需要解决负例采样和类别不平衡等问题。

2.3 水利知识图谱构建

面向水利知识领域的知识图谱构建属于垂直知识图谱构建，即针对特定领域（行业）进行的知识图谱构建，具有鲜明的领域特色。

学者们对水利领域知识图谱的应用研究可分为两个主要类别：一类是整合专业知识，旨在更深入地分析水利领域知识，以便学者们进行领域知识的推进和跨学科研究；另一类则关注实际应用，以知识图谱技术整合水利领域知识，以期实现水利知识的智能应用，如智能问答系统和安全缺陷的定位等。

第一类水利知识图谱应用的研究重点在于对知识进行分析。例如，Chen 等[23]利用 Cite-Space 软件分析了我国近十年来有关水资源研究的科技论文，并制作了国内水资源研究的知识图谱；Jin 等[24]对 512 篇水资源承载力研究文献进行了分析，并创建了水资源研究热点知识图谱，可视化地展示了该领域正在不断更新的研究内容；Mao 等[25]基于知识图谱的结构性和时间性指标，探索了水生态与水环境领域专业知识图谱的构建；Li 等[26]利用知识图谱技术研究了我国节水灌溉技术的演进；Liu 等[27]以 4 263 篇文献为原始数据，应用知识图谱工具深入分析了再生水问题的发展趋势与研究热点。第二类水利知识图谱应用的研究更侧重于知识图谱在知识智能化方面的运用。例如，Wang 等[28]为解决传统水污染溯源方法工作量大的问题，构建了水污染关系知识图谱，优化了水污染扩散模型的训练，实现了对污染物溯源的正向模拟。

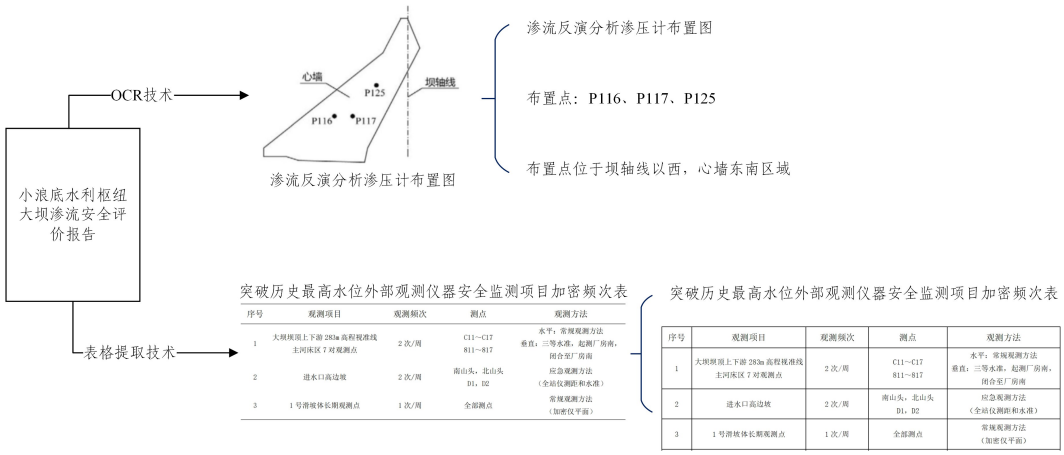
上述研究涵盖的水利知识主要针对水利行业的核心领域，包括对河流、湖泊等水利要素的知识图谱构建。第二类研究侧重于实际应用，但其中的知识图谱应用较为有限，仅能在模型优化方面发挥作用。本文构建的水利枢纽工程知识图谱属于第二类水利知识图谱应用研究，着重于对水利枢纽工程这一特定主体及相关内容进行知识图谱构建（这在目前的水利知识图谱研究中较为缺乏）。本文在前人研究成果的基础上，通过更深入的专业领域整合，将枢纽工程领域的专业知识融入其中，相比更广泛的水利知识图谱，能够更准确地获取相关知识。此外，在将知识图谱应用于工程智能化的同时，也保留了知识图谱本身作为知识库的特性，能够与其他领域的知识图谱进行交叉融合。

3 研究方法

3.1 OCR 和表格提取

在构建领域知识图谱时会使用多种数据类型，这些数据类型并不局限于文本，还包括图表数据，这有助于将领域知识以结构化的方式存储和展示。图表数据具有独特性，因此，其处理方式与传统文本数据的处理方式存在不同，但这种差异有利于我们更好地利用图像内容，进一步丰富知识图谱的内容和价值。

在处理图片内容时，主要将其分为两部分：图片的归属路径和图片内容。归属路径包括图片名称和所属文档，可将其视为一个实体。可使用 OCR 技术[29]来提取图片内容，进行内容识别和提取图片及其文本信息，将其转化为结构化数据，有助于丰富和完善知识图谱。例如，在评价水利工程安全的文档中，包含许多工程结构描述图片，仅用文本记录无法覆盖全部信息，但通过 OCR 技术便可以提取相关段落和图片内容，并将这些数据与其他实体进行关联。表格也是信息的重要来源之一。为了高效处理表格数据，可采用表格提取技术[30-31]，该技术能有序地整理表格信息并导入 Excel 表格中，例如在处理水利工程业务规则的文档时，其中包含大量的可读性差的表格图片。单纯的图片提取对后续的信息抽取会造成极大的困难，此时使用表格提取技术进行自动提取、整合和转化表格数据，可以提高数据处理的准确性和效率。在后续的数据分析中，该技术也将拼接表格内容用以提供简洁的数据结构。OCR 技术和表格提取技术的使用示例如图 1 所示。各类图表实体的命名以及相关的基础信息被分别储存于 Neo4j 图数据库和 MySQL 关系型数据库中。这两个数据库间的关联由图表的唯一名称实现。



序号	观测项目	观测频次	测点	观测方法
1	大坝坝顶上下游 283m 高程视准线主河床区 7 对观测点	2 次/周	C11~C17 811~817	水平：常规观测方法 垂直：三等水准，起测厂房南，闭合至厂房南
2	进水口高边坡	2 次/周	南山头，北山头 D1，D2	应急观测方法 （全站仪测距和水准）
3	1 号滑坡体长期观测点	1 次/周	全部测点	常规观测方法 （加密仅平面）

图 1 OCR 技术和表格提取技术的使用示例

Fig. 1 Examples of using OCR technology and table extraction technology

3.2 半自动标注

半自动标注是使用词典库中的双向最大匹配算法对语料进行初步标注，然后由人工进行确认和校对的过程。为了高效且一致地标注，可采用专门针对医疗文本开发的实体和关系标注平台[32]。通过重新配置该知识标注平台，使其适用于某一领域的知识结构，形成定制的实体和关系标注平台。

本文在构建 WCHP-KG 时便配置了这个标注平台，使其适用于水利工程领域的业务规则和工程安全方面，其中词典库的数据基于对水利数据的收集得到。经过半自动试标注与水利工程专业人员的确认后，构建了可供选择的实体和关系体系。在标注水利工程语料时，只能从预定义的实体和关系类型中进行选择。使用不同颜色区分实体概念，并使用连线表示实体之间的关系。标注流程可分为预标注和正式标注阶段。

预标注阶段指在正式对数据进行标注之前的准备阶段，如图 2 所示，其中包括数据的背景调研、标注规范的确定、标注平台的部署，以及试标注等一系列流程。这一阶段的主要目的是熟悉数据，通过领域知识筛查、简化数据，为标注平台提供清晰而简洁的实体和关系，以便进行后续的自动标注和半自动标注。试标注流程的存在是为了验证筛查后的标注规范的适用性，以确保后续标注过程的稳定性。

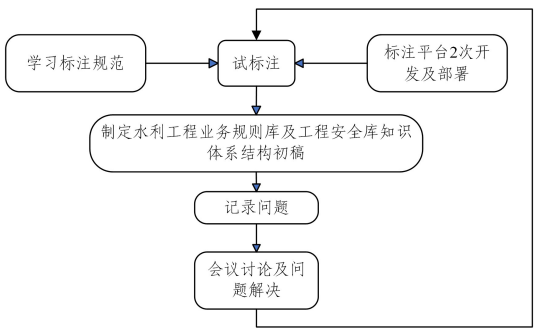


图 2 预标注流程

Fig. 2 Process of pre-labeling

正式标注阶段是对数据进行大规模标注的阶段，如图 3 所示。为确保标注数据的准确性和一致性，我们采用了多轮交叉标注的策略。在处理非结构化数据方面采用了基于规则的方法进行实体识别，随后经过人工校对，将经过整合和去重处理的数据转化为结构化的三元组，作为图谱数据的有益补充。知识标注平台会呈现一标任务的情况以及后续交叉检查的文件说明，不同阶段的标注任务以 N 标文件的形式分割开来，上一阶段的任务完成后才会生成下阶段任务。

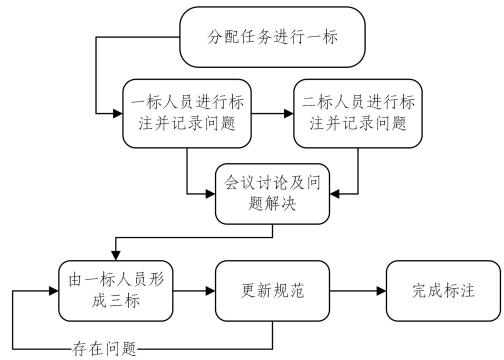


图 3 正式标注流程

Fig. 3 Process of formal labeling

在两个标注阶段，不恰当的标注经常会导致实体选择错误、关系混乱等问题。因此在进行实体标注时，需要遵循特定的标注规则。标注规则可简单概括为两类：单实体的标注规则和多实体的标注规则。下文以工程安全领域的半自动标注为例，详细阐述了这两类标注规则的应用。

第一类是单实体的标注细则，其可以分为两个方面，即“相似单实体的区分”与“复合单实体的统一标注”。前者主要指句子中存在的实体被主观判断为类型 A，但实际上是类型 B 的情况，如例句 1，类似于“引水发电系统”这样的建筑名词有时会被视为设备，但实体类型实际为“建筑”。后者涉及预先规定的实体类型，需要统一标注多个实体，如例句 2 描述了坝体总渗流量，包括整体概况和当前情况，融合了“状态变化”实体类型的“状态”和“变化”，展现了它们的统一性。

例句 1 小浪底水利枢纽工程等别为Ⅰ等，主要建筑物为 1 级建筑物。枢纽工程由拦河大坝、泄洪排沙系统和引水发电系统 3 部分组成。（下划线为实体）

例句 2 主坝及两岸渗流量呈逐年降低趋势，但在 2011 年底至 2012 年坝基和左岸渗流量较前几年有明显增加，主要与库水位保持在 260.00 m 以上运行较长时间有关，应密切关注渗流量变化。（下划线为实体）

第二类是多实体的标注细则，同样可以分为两个方面，即

“同一句子中存在多个实体”和“实体嵌套”。前者类似于区分相似但独立实体的标准，但与“相似单实体的区分”标准不同的是，后者可能存在实体 A 和实体 B 共存的情况，这可能导致实体 A 的类型发生变化。如例句 3 中，单独考虑“2012 年底”时，可将其归类为“时间周期”实体类型；然而句子后半部分描述了小浪底水位和渗流关系历史，这使得“2012 年底”成为对历史描述的“状语”，纳入整体“历史”实体类型。后者指由多个实体组合而成的复合实体，通常出现在标题中。如例句 4 展示了报告标题，其中明确包含“2014 年”，但解读时应将其视为整体，而不单独视为“时间周期”实体。

例句 3　2012 年底，水位蓄至 270.10 m 时，渗流量与当年 250.00 m 时差不多，由于蓄水时间短，其渗流量变化还需进一步观察。（下划线为实体）

例句 4　《2014 年小浪底水利框纽大坝渗流安全评价报告》（下划线为实体）

3.3　实体及关系抽取模型

实体识别和关系抽取是领域知识图谱构建的核心所在，BiLSTM-CRF 和 Lattice-LSTM 模型是目前深度学习领域中备受关注的主流实体识别模型，能够有效地提升模型的准确性。本文主要以所使用的知识标注平台中采用的 BiLSTM-CRF 模型为主要的实体识别模型，该模型在构建 WCHP-KG 中的实体识别应用如图 4 所示。

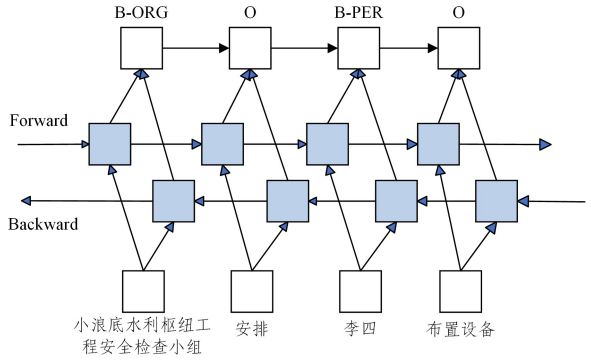


图 4　BiLSTM-CRF 模型

Fig. 4　BiLSTM-CRF model

BiLSTM-CRF 模型是将一个双向 LSTM 层的输出拼接到一个 CRF 网络的输入构成的，它能够使用序列的“历史”和“未来”信息，这些额外的信息能够给序列标注任务带来可观的性能提升。假设数据集中包含人物（Person，以下简称 PER）和机构（Organization，以下简称 ORG）两种实体类型，并假设采用 BIO 标注体系，则存在 5 种实体标签：B-PER，I-PER，B-ORG，I-ORG，以及 O（非实体）。

模型运行的伪代码如算法 1 所示。首先模型运行双向 LSTM-CRF 层的前向过程，利用 LSTM 来捕捉序列数据的历史信息并计算位置分数以确定存在实体的位置。之后运行 CRF 层的前向和后向过程，计算每个转移状态边的梯度。最后利用反向传播算法计算输出到输入各模块的损失值，然后更新模型参数。随着模型逐渐收敛，CRF 层可以提供句子级的标注信息。

算法 1　BiLSTM-CRF 模型

```
1. for each epoch do
   // suggest setting it to 32 to ensure model convergence
2.    for each batch do
      // batch size needs to be set appropriately
3.       1)bidirectional LSTM-CRF model forward pass:
4.          forward pass for forward state LSTM
5.          forward pass for backward state LSTM
            // obtain annotation information at the sentence level
6.       2)CRF layer forward and backward pass
         // capture historical information of sequence data
7.       3)bidirectional LSTM-CRF model backward pass:
8.          forward pass for forward state LSTM
9.          forward pass for backward state LSTM
10.         4)update parameters
11.   end
12. end
```

因为 PCNN 模型[18]能有效地捕捉实体间的特征分布，有助于提高关系抽取的准确性，因此本文在 WCHP-KG 的构建中选择 PCNN 模型作为关系抽取的主模型。PCNN 实体关系抽取模型由 4 个模块组成：向量表示、卷积、分段最大池化和 Softmax 输出，具体结构如图 5 所示。

在已确定实体的情况下，可以将实体间关系抽取看作分类问题。PCNN 首先使用预训练的词向量和位置特征（位置向量）将文本转换为低维向量表示。传统 CNN 最大池化方法被优化为分段池化。在池化层，卷积输出分为实体 *Sh* 之前的信息 *Ch*、实体间的信息 *Cp* 和实体 *St* 之后的信息 *Ct*，然后分别进行最大池化。随后将 *n* 个卷积滤波器的结果拼接，经过 tanh 函数得到 PCNN 层的输出，传递至 Softmax 层进行关系分类。

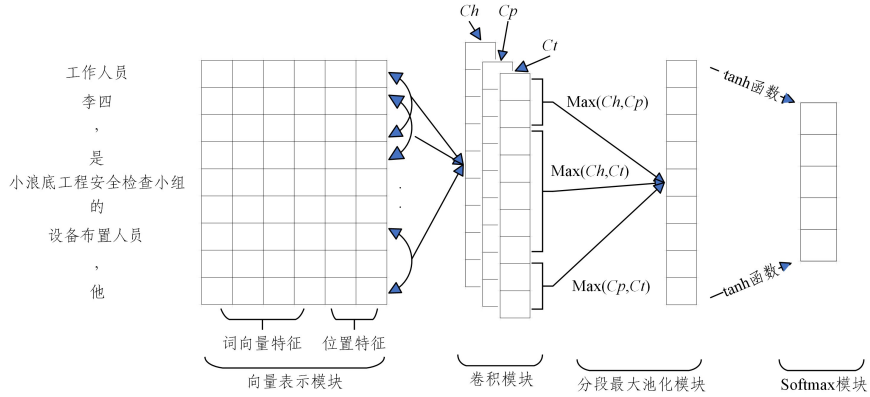


图 5　PCNN 模型

Fig. 5　PCNN model

4 构建过程

4.1 WCHP-KG 构建流程

知识图谱的构建可分为两种方式：自顶向下和自底向上[33]。本文在 WCHP-KG 的构建中采用了自顶向下的方式，借鉴国内水利枢纽相关领域分类体系和术语，以图谱的形式呈现构建流程。具体而言，该方法首先确定模式层，然后在此基础上构筑数据层，构建过程如图 6 所示。

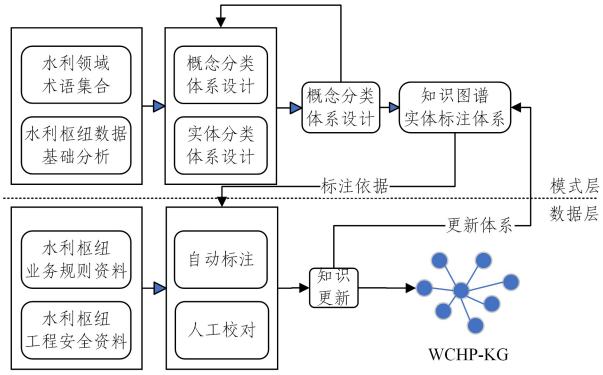


图 6 WCHP-KG 构建流程

Fig. 6 WCHP-KG construction process

在模式层面，首先对小浪底水利枢纽工程资料进行系统整理并制定了构建计划。通过人工试标注和详尽案例分析建立了关系分类体系，并在水利枢纽工作人员的协助下制定了 WCHP-KG 实体标注体系。然后结合小浪底水利枢纽的实际应用背景，在模式层基础上灵活采用半自动或自动方法，对搜集整理的半结构化和非结构化文本中的实体及其关系进行标注。

在数据层面，WCHP-KG 使用三元组来描述实体及实体关系，同时，为了全面准确地描述知识，也会在三元组的基础上加入对每一组的约束或属性，将原本的三元组扩展为六元组。S_h 为实体 1，S_{h_pro} 为实体 1 的属性，S_t 为实体 2，S_{t_pro} 为实体 2 的属性，P 为实体 1 和实体 2 的关系，$P_{\_pro}$ 为关系的属性，组成形式为〈S_h，P，S_t〉的三元组或形式为〈S_h，S_{h_pro}，P，$P_{\_pro}$，S_t，S_{t_pro}〉的六元组。

4.2 WCHP-KG 模式层构建

本文所构建的 WCHP-KG 主要涵盖了水利领域内业务规则与工程安全这两个重要方面的知识。本文在论述上重点关注这两个知识领域的关键内容。表 2 中列出了部分标注关系的定义。需要注意的是，两个领域所涉及的具体内容和关系还可以进一步细分为多个子类，而其中未提及的子关系则不再一一详述。

表 2 WCHP-KG 部分标注关系定义

Table 2 Partial definition of annotation relationship in WCHP-KG

领域	关系	子关系	关系描述
业务规则	检修规章-范围	适用范围	指该检修规章适用于哪个具体设备的检修及维护工作
		规范范围	指该检修规章规定了某个设备的维护、检修以及技术要求等具体内容
	同义词	—	术语包括中文术语及其英文表达，如“浮充电”=“Floating charge”
工程安全	工程-参数	标准	对工程的统一规定，比如“防洪标准”等
		容量	工程的相关容量，有“年供水量”等
	工程-位置	—	位置是对各种地理位置的描述，包括工程所处地区、工程上下游等

在业务规则方面，以检修章程为主要描述对象建立了实体之间的层级和关联关系，形成了综合的知识描述体系，包括概念分类和关系分类。其中包含 21 类水利枢纽业务规则实体，涵盖要素如标题、规范引用文件、术语和设备名。同时，考虑了自然语言处理属性，如同义词、编码，以及与其他概念（包括检修章程）之间的 44 类关系。

在工程安全方面，以工程为核心，建立了一个知识描述体系，主要实体是工程，次实体涵盖多种相关工作。体系中定义了 22 个水利枢纽工程安全实体类型，包括文件名称、工程、工作、会议等。通过拓展和关联水利枢纽工程的基本属性，结合其在安全监测等方面的工作，共建立了 66 种关系类型。

4.3 WCHP-KG 数据层构建

WCHP-KG 的数据层构建共分为两个关键步骤。在数据层构建的首阶段，主要获取了多样化的水利工程数据。这些数据包括数量化信息和工程运营管理规定以及详实的安全检测记录。基于这些数据构建了 WCHP-KG，并重点利用业务规则和工程安全方面的水利领域知识来实现构建。通过广泛的信息搜集，确保了数据的全面准确。数据层构建的第二阶段是半自动标注和模型自动抽取的应用，这一步骤至关重要。通过专业算法和半自动标注技术，能够从数据中识别特定信息并进行标注，显著提升数据质量和准确性。同时，先进的模型自动抽取技术也能从大规模数据中快速提取有价值的信息，进一步丰富数据层内容。

在业务规则方面，通过分析运行规程和相关资料建立了业务规则的实体体系，并维护了一系列重要的规程文件。这些文件涵盖水工建筑物、息系统以及通信系统等多个方面，其中特别关注检修章程。从多个文本来源中提取水利工程业务规则库的关键内容后，采用半自动标注和自动抽取的方式处理半结构化数据，而非结构化数据则需要通过人工标注处理。最终在业务规则知识数据上获得了 43 本文档的数据，总字数达 87.2 万，如表 3 所列。

表 3 业务规则水利数据来源

Table 3 Water resources data source of business rules

名称	描述	语料规模/本	字数/万
分析报告	资料分析报告	1	5.9
维护规程	水轮发电机组运行规程	14	32.7
	水情水调运行维护规程	2	3.6
	水工建筑物运行维护规程	7	11.9
	通信系统运行维护规程	9	21.7
	信息系统运行维护规程	6	6.3
	其他	4	5.1

在工程安全方面，依赖水库信息和安全检测资料详尽地获取了实体体系的信息。这些信息涵盖了问题描述、会议记录、定时检测汇报等内容，包括工程风险隐患、隐患事故案例等多个方面。通过对多源文本的深入分析，提取了关于水利

工程安全的关键信息。半结构化数据和非结构化数据的处理同业务规则数据的处理相同，使用半自动标注、自动抽取以及人工标注。最终在工程安全知识数据上获得了 33 本文档的数据，总字数约 97.3 万，如表 4 所列。

表 4　工程安全水利数据来源

Table 4　Water conservancy data source of engineering safety

名称	描述	语料规模/本	字数/万
案例隐患	工程风险隐患	2	3.4
	隐患事故案例	2	3.1
	事件处置案例	1	2.1
检查鉴定	工程安全会商	3	6.1
	工程安全鉴定	15	49.2
	专项安全检查	10	33.4

5　实验

5.1　数据集介绍

本文的实验数据来自于在 WCHP-KG 数据层构建阶段使用的业务规则和工程安全知识文档。我们随机选取了其中的 50 份文档，经过无用信息筛查和去重处理，最终总计得到约 5 万条句子。在经过半自动标注处理后，获得了 12 494 个实体关系对，并将这些数据的 80%用作训练数据，20%用作验证数据进行实验。

5.2　对比实验

本文通过对比实验来探究不同实体识别模型的性能及其在应用中的最佳参数选择。在实体标注规范中，由于实体间关系通常较为简单且主要为包含关系，因此关系抽取实验结果偏高，缺乏深入的分析意义，故本文选择略去关系抽取模型的实验部分。

对比实验所选取的实体抽取模型包括具备大规模网络结构的 ADV-SEQ LABEL 模型、单层 LSTM 和 CRF 的 SEQ LABELING 模型，以及应用了双向 LSTM 和 CRF 的 BiLSTM-CRF 模型。同时，考虑到水利文档上下文的紧密性，使用预训练模型 BERT 对传入的文本数据进行预处理，将 BERT 处理后的序列信息传入 BiLSTM-CRF 模型进行实体识别，即结合了 BERT 使用的 BERT-BiLSTM-CRF 模型。在参数设置上，实验主要涉及学习率与 epoch 的调整。

5.3　评价指标

本文采用实体识别实验中常用的 3 个评价指标：精确率、召回率和 F1 值。其中 F1 值被用作主要参考指标，其他两个指标则被视为辅助评估指标。

精确率指模型正确预测为正占全部预测为正的比例，如式(1)所示：

$$Precision=\frac{TP}{TP+FP} \tag{1}$$

其中，TP 表示预测为正的实体中实际为正的数量，FP 表示预测为正的实体中实际为负的数量。

召回率指模型正确预测为正占全部实际为正的比例，如式(2)所示：

$$Recall=\frac{TP}{TP+FN} \tag{2}$$

其中，TP 表示预测为正的实体中实际为正的数量，FN 表示预测为假的实体中实际为负的数量。

F1 值是精确率和召回率的调和平均数，其综合考虑了分类器的精确性和完整性，如式(3)所示：

$$F1=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{3}$$

5.4　实验结果及分析

不同的实体识别模型的实验结果如表 5 所列。本文采用相同的参数设置(dropout 率为 0.5，学习率为 0.001)对每个实体识别模型进行了 32 个 epoch 的训练，并记录了在验证集上当损失值不再下降时的各项评价指标，将其作为实验结果。为了排除实验中的偶然因素，本文进行了 5 次重复实验，并取其平均值作为最终结果。

表 5　不同实体识别模型的实验结果

Table 5　Experimental results of various entity recognition models

Model	*F1*	*Precision*	*Recall*
ADV-SEQ LABEL	0.576	0.582	0.571
SEQ LABELING	0.546	0.548	0.544
BILSTM-CRF	0.571	0.574	0.568
BERT-BILSTM-CRF	**0.683**	**0.686**	**0.681**

注：表中最佳的结果用粗体标出。

可以观察到，4 种模型的实体抽取结果均不理想，这一现象与水利文档中存在的大量设备名称及工程任务代号有一定关联。例如，部分设备名称以水库名称作为设备前缀，导致模型更容易将其识别为“水库”实体，而非“设备”实体。ADV-SEQ LABEL 相较于 SEQ LABELING 和 BILSTM-CRF 表现更佳，主要原因在于大规模的网络结构存在的多层 CNN 卷积层相较于单向或者双向 LSTM 能够捕捉到更多且更精确的序列历史信息。BERT-BILSTM-CRF 的性能优于 ADV-SEQ LABEL 的原因在于其对 BERT 的应用。由于 BERT 具有强大的信息记忆能力和抽取能力，因此模型将其置于最底层来对文本信息的上下文进行抽取和预处理，然后将文本序列送入 BILSTM-CRF 的 LSTM 层，这样可以减少水利文档中的大量“噪声”信息带来的干扰。

为了进一步分析 BERT-BILSTM-CRF 模型的最佳参数，本文对其在不同学习率(0.001，0.005 和 0.01)以及不同训练轮次(16 个 epoch 和 32 个 epoch)下进行了实验。总共设置了 6 组实验，每组实验均重复 3 次，并取平均值作为实验结果。实验结果如表 6 所列。

表 6　不同参数下的 BERT-BILSTM-CRF 的实验结果

Table 6　Experimental results of BERT-BILSTM-CRF with different parameters

Settings	*F1*	*Precision*	*Recall*
0.001+16	0.591	0.593	0.589
0.001+32	**0.683**	**0.686**	**0.681**
0.005+16	0.535	0.537	0.534
0.005+32	0.564	0.569	0.559
0.01+16	0.482	0.488	0.476
0.01+32	0.498	0.495	0.501

注：表中最佳的结果用粗体标出。

由表 6 可以观察到，随着 epoch 轮次的增加，在 3 个不同的学习率下，模型的性能都有所改善。考虑到本文实验受限于实验时间和算力，可以将 32 个 epoch 视为最适合本文模型

的训练轮次。此外，随着学习率的成倍增加，实验结果逐渐下降。这是因为过大的学习率导致模型在训练初期迅速越过最优点，使得模型无法充分学习数据的特征，更多地偏向于对“水库”这类出现频次较多的实体的学习。

6 图谱展示及应用

6.1 WCHP-KG 图谱展示

标注过程中共有 2 名专业人员和 26 名标注人员参与了标注工作。为了呈现 WCHP-KG 中各个概念间的关联，我们设计了一个知识图谱可视化展示平台。WCHP-KG 的部分展示界面如图 7 所示。

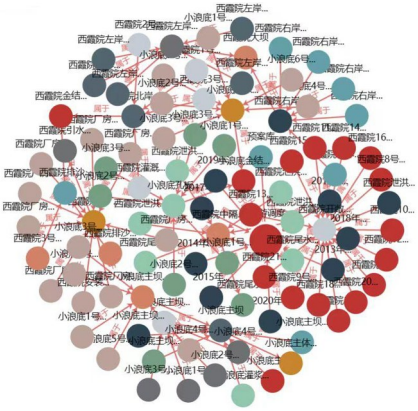


图 7 WCHP-KG 展示

Fig. 7 WCHP-KG display

6.2 WCHP-KG 图谱应用

基于水利枢纽工程的业务需求和流域管理的特点，本文进行了知识图谱应用的设计，包括知识图谱内容的查询服务以及针对小浪底知识图谱的智能问答功能。

在深入分析黄河流域的基础上，我们通过知识图谱内容查询服务快速准确地检索信息，为小浪底水利决策提供支持。同时，智能问答功能以人性化的方式回应用户问题，为广大使用者提供高效信息交流。这些应用旨在将先进知识图谱技术与小浪底水利需求相结合，推动该领域的发展。

6.2.1 知识图谱查询服务

知识图谱查询服务能够以可视化的方式展示经过标注及抽取的实体内容、基础属性和关系，为用户提供查询服务，涵盖了水利工程业务规则以及工程安全的规程关系展示。以业务规则领域的应用为例，水利枢纽的运行维护规程中包括总则、险情报告以及组织保障等内容。图 8 通过实例展示了现场应急指挥部职责的分解。利用知识图谱查询服务，不仅可以获取现场应急指挥部职责的属性信息，还可以快速梳理出相关的负责人和责任范围。

图 8 WCHP-KG 查询服务实例

Fig. 8 Instance of WCHP-KG query service

6.2.2 知识图谱智能问答

根据水利工程业务规则以及工程安全规程的要求设置了小浪底知识图谱智能问答系统，如图 9 所示。该系统涵盖了以下主要类型的问答：

1）统计类：用于统计特定流域的汛期等关键信息。例如提问“小浪底调水调沙的开始时间是什么？”的问题，用户可以迅速获取有关小浪底调水调沙的详细信息。

2）文件规则类：旨在呈现与水利工程相关的法规文件和规章制度。例如提问“我国水资源的管理体制是怎样的？”的问题，智能问答系统可根据后台管理人员上传的相关文件，为用户提供关于我国水资源管理体制的详细解答。这种方式为决策者提供了必要的技术支持，以便更好地了解法规和制度。

通过此智能问答系统，用户能够更加便捷地获取水利领域的关键信息，有助于作出更明智的决策。

图 9 WCHP-KG 智能问答

Fig. 9 WCHP-KG intelligent Q&A

结束语 本文梳理了 WCHP-KG 的构建过程、方法使用以及实际应用，并提出了一套有效的基于知识标注平台的水利知识图谱构建流程。以面向水利枢纽工程的知识图谱构建为例，首先是模式层和数据层的基础构建。在模式层方面，通过整合多个来源的水利工程业务规则库和工程安全库文本，并在专业人员指导下设计了完备的知识图谱描述体系。在

数据层方面，通过规则和机器学习方法成功地抽取了实体及其关系，从而构建了面向水利枢纽工程的知识图谱的知识本体。在知识融合阶段，对人工和自动标注的三元组进行深入分析，并经过专业人员的人工检查，确保了知识图谱的准确性和一致性。

然而在 WCHP-KG 的构建过程中，对于大规模数据的处理仍存在困难，最后图谱对于数据所包含的海量信息的使用也并不十分完全，这也反映了本文提出的基于知识标注平台的水利知识图谱构建流程的局限性。为了优化构建流程，下一步计划采用半监督的知识融合方法，以对大规模知识进行筛选和整合，同时借助大模型的优势，在水利知识图谱构建时对领域信息进行更具广度和深度的使用，为更全面的应用和分析提供丰富的知识资源。

参 考 文 献

[1] HANG T T，FENG J，LU J M. Knowledge Graph Construction Techniques：Classification，Investigation，and Future Directions [J]. Computer Science，2021，48(2)：175-189.

[2] WANG M，WANG H F，LI B H，et al. Overview of Key Technologies for the New Generation Knowledge Graph[J]. Computer Research and Development，2022，59(9)：1947-1965.

[3] FENG J，XU X，LU J M. Construction and Application of Water Conservancy Information Knowledge Graph[J]. Computer and Modernization，2019(9)：35-40.

[4] WANG L. The application of knowledge graph technology in the joint scheduling of water conservancy projects in the Beijiang River Basin[J]. Heilongjiang Water Conservancy Technology，2021，49(12)：187-190.

[5] CAI Y，XIE W J，CHENG Y L，et al. A review of key technology research on a national water conservancy map[J]. Journal of Water Resources，2020，51(6)：685-694.

[6] DUAN H，HAN K，ZHAO H L，et al. Research on the construction of a comprehensive knowledge graph for water conservancy [J]. Journal of Water Resources，2021，52(8)：948-958.

[7] WANG L，SUN W J. Comparative Study on Simulation of Rainfall Characteristics between HEC-HMS and Vflo Based on DEM Data：A Case Study of Miyun District，Beijing[J]. Journal of Environmental Science，2019，39(10)：3559-3565.

[8] XIAO J Q，ZHOU X，PAN Y，et al. Application of GA-BP optimized TS fuzzy neural network water quality monitoring and evaluation system prediction model－taking the Taihu Lake as an example[J]. Journal of Southwest University(Natural Science Edition)，2019，41(12)：110-119.

[9] LIU M，FU B L，HE H C，et al. Water surface monitoring and water quality parameter inversion of the Li River based on multi temporal active and passive remote sensing(2016-2020)[J]. Lake Science，2021，33(3)：687-705.

[10] HUILIN S Z W. Overview on the advance of the research on named entity recognition[J]. Data Analysis and Knowledge Discovery，2010，26(6)：42-47.

[11] LIU X，ZHANG S，WEI F，et al. Recognizing named entities in tweets[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies. 2011：359-367.

[12] XU K，ZHOU Z，HAO T，et al. A bidirectional LSTM and conditional random fields approach to medical named entity recognition[C]// Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Springer International Publishing，2018：355-365.

[13] JIA Y，MA X. Attention in character-based BiLSTM-CRF for Chinese named entity recognition[C]// Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence. 2019：1-4.

[14] KUMAR S. A survey of deep learning methods for relation extraction[J]. arXiv：1705. 03645，2017.

[15] 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.

[16] SOCHER R，HUVAL B，MANNING C D，et al. Semantic compositionality through recursive matrix-vector spaces[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012：1201-1211.

[17] XU Y，MOU L，LI G，et al. Classifying relations via long short term memory networks along shortest dependency paths[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015：1785-1794.

[18] 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.

[19] SURDEANU M，TIBSHIRANI J，NALLAPATI R，et al. Multi-instance multi-label learning for relation extraction[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012：455-465.

[20] HE H，SUN X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[21] VASHISHTH S，JOSHI R，PRAYAGA S S，et al. RESIDE：Improving Distantly-Supervised Neural Relation Extraction using Side Information[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018：1257-1266.

[22] MIWA M，BANSAL M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). Association for Computational Linguistics，2016.

[23] CHEN S Y，LU D D，CHENG H M. Knowledge Graph Analysis

of Water Resources Research in China Based on Technology Text Mining[J]. Hydrology, 2019, 39(2): 61-66.

[24] JIN J L, CHEN P F, CHEN M L, et al. Bibliometric analysis of water resource carrying capacity research based on knowledge graph[J]. Water Resources Protection, 2019, 35(6): 14-24.

[25] MAO W S, ZHAO H L, JIANG Y Z, et al. Construction and application of a knowledge graph for domestic water ecological environment research based on bibliometrics[J]. Journal of Water Resources, 2019, 50(11): 1400-1416.

[26] LI Z Q, HU H. In recent years, research progress on water-saving irrigation technology in China — a scientific knowledge graph analysis based on bibliometrics[J]. Water Saving Irrigation, 2015(8): 104-109.

[27] LIU X J, YANG X, FU H L. The Development Trend and Research Hotspot Analysis of Regenerated Water Research: A Graph Quantification Study Based on CiteSpace[J]. Resources and Environment in Arid Areas, 2019(4): 68-75.

[28] WANG X L, XUE X P, SUN R F. A Tracing Method for Sudden Water Pollution Events Based on Particle Swarm Optimization and Knowledge Graph[J]. Hydroelectric Power, 2020, 46(2): 17-21.

[29] WANG J, TANG J, YANG M, et al. Improving OCR-based image captioning by incorporating geometrical relationship[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1306-1315.

[30] CHEN W, ZHA H, CHEN Z, et al. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data [C]// Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 1026-1036.

[31] ZHU F, LEI W, HUANG Y, et al. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3277-3287.

[32] ZHANG K L, ZHAO X, GUAN T F, et al. Construction and application of entity and relationship labeling platform for medical texts[J]. J. Chin. Inf. Process, 2020, 34(6): 36-44.

[33] HOU M W, WEI R, LU L, et al. A review of knowledge graph research and its application in the medical field[J]. Computer Research and Development, 2018, 55(12): 2587-2599.

ZHANG Junhui, born in 1986, Ph. D, senior engineer, is a member of CCF (No. 07619G). Her main research interests include natural language processing and so on.

ZAN Hongying, born in 1966, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 08671S). Her main research interests include machine translation, Q&A, abstract and machine learning.

(责任编辑:何杨)