知识图谱在 BIM 模型审查中的应用研究

刘毅[1]，吴浪韬[1]，梁雄[2]，罗征[2]，胡振中[1]

（1. 清华大学土木工程系，北京，100084）

（2. 广联达科技股份有限公司，北京，100193）

【摘 要】随着我国建筑行业的高速发展，建筑领域在各个方面积累了大量的原始数据，如何从这些琐碎的数据中获取结构化的知识成为了行业智能化所面临的新问题，而知识图谱是一个很好的解决方案。在构建知识图谱的过程中，本研究分别应用和评估了 BiLSTM-CRF 和 ResCNN 模型，结果表明它们具有良好的性能。针对机电设备规范中的属性约束条件，本研究提出了一种基于知识图谱的提取方法，该方法对于辅助 BIM 模型的审查具有重要意义。

【关键词】知识图谱；机电设备；约束条件；实体提取；关系抽取

1 引言

随着我国建筑行业的高速发展，建筑领域在技术规范、学术文献、项目档案和社区讨论等各个方面中积累了丰富的数据。如果能够从这些数据中提取出关键信息，将对建筑设计、施工和运维的自动化实现起到极大的帮助。例如，把规范中的设计限制信息提取出来，就可以减少设计人员对规范的查询，进而提高设计效率。

知识图谱是结构化的知识库[1]，它通过符号形式来描述物理世界中的概念及其相互关系。知识图谱的基本单元是“实体－关系－实体”或者“实体－属性－属性值”三元组，实体与实体之间通过关系相互联结，构成一个网状的知识结构。

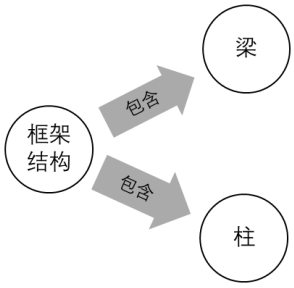

图 1 知识图谱示例

【基金项目】清华大学-广联达 BIM 联合研究中心资助项目

【作者简介】胡振中(1983-)，男，清华大学土木系副教授。主要研究方向为土木与海洋工程信息技术、建筑信息模型（BIM）和数字防灾技术。E-mail：huzhenzhong@tsinghua.edu.cn

知识图谱从海量的数据中提取出实体和关系信息，并通过图的方式进行表示，实现了知识的结构化和可视化，从而使知识的获取更加准确和快速[错误!未找到引用源。]。知识库的构建方法通常有以下两种：人工构建的方法和自动（或半自动）构建的方法。其中，人工构建的方法要求工作者对该领域有深入的了解，并且在整个构建过程中都要参与，存在着效率低、成本高、可扩展性差的问题，因此自动（或半自动）构建的方法正在逐步取代人工构建的方法[3]。

随着 NLP（Natural Language Processing，自然语言处理）技术的不断发展并日渐成熟，知识图谱构建的自动化程度不断提高，其应用范围也越来越广，特别是在互联网行业中出现了大量的应用。近年来国内也出现了搜狗知立方、百度知心等知识图谱的应用，如知立方可以通过已有语义网来推理获取新的知识以补充数据，并实现对用户查询内容的句法分析和语义理解等[4]。

2 研究现状

知识提取是从原始数据中提取所需信息的过程，也是构建知识图谱的主要步骤，涉及的关键技术包括实体提取、关系抽取和属性抽取，下面分别对它们进行介绍。

实体提取也称为 NER（Named entity recognition，命名实体识别），是指识别出文本中具有特定意义的命名实体如专业实体词汇。早期的实体提取多采用人工编写规则的方法，主要从文本中识别地名、人名、时间等特定的实体[5]，如 Rau 通过结合人工编写规则与启发式算法的方式实现了能够自动提取公司名称的原型系统[6]。后面渐渐出现了统计机器学习和深度学习的方法，一般通过标注数据对模型进行参数训练，然后利用训练好的模型提取新数据集中的实体，如 Liu 等通过条件随机场模型以及 K-最近邻算法来识别 Twitter 文本中的实体[7]。

关系抽取也称为 RE（Relation extraction，关系抽取）。经过实体提取之后，得到的只是实体内容和类别信息，还需要获取实体与实体之间的关联信息，这样才能构建出完整的知识图谱。关系抽取的方法与实体提取类似，包括人工编写规则和机器学习方法，如刘克彬等通过知网的知识库来构造核函数，在开放数据集上进行关系抽取并得到了 88%的准确率[8]。

属性抽取是指从原始数据中抽取出实体的相关属性信息，如人物的出生时间、出生地点等。属性抽取问题可以归类到广义上的关系抽取问题，即把属性作为实体和属性值之间的一种关系，这样的话就可以通过关系抽取的方法进行属性抽取，如郭剑毅等通过支持向量机实现了对人物属性的抽取[9]。其方法与前两者类似，由于百科类网站上有大量的半结构化数据，可以通过这些数据来训练模型，再利用模型对非结构化数据进行属性抽取[10]。

建筑领域中已经出现了构建知识图谱的尝试，如吴浪韬采用 BiLSTM-CRF 模型进行实体提取，使用 CNN 模型进行关系抽取，在 MEP 领域构建了一个以三元组方式储存的关系数据库原型，并通过图数据库 Neo4j 实现了数据的检索和可视化展示[11]；吴雪峰同样采用 BiLSTM-CRF 模型进行实体提取，并进一步进行了关系抽取和属性抽取，最终通过图数据库 Neo4j 对获取的知识进行存储，构建了一个煤矿巷道支护领域的知识图谱[12]。

3 实体提取和关系抽取

实体提取和关系抽取是构建知识图谱的关键步骤，其中实体提取是从文本中提取出有意义的实体，而关系抽取则是抽出实体之间具有的关系。本章主要介绍我们在实体提取以及关系抽取中选用的方法、工作流程以及评估结果。

3.1 实体提取

实体提取的目的是从文本中提取出实体的内容和类别信息，其中类别一般是预先定义好的。BiLSTM-CRF（双向长短期记忆-条件随机场）是一种结合了深度学习和统计机器学习的方法，也是目前比较常用的实体提取方法。BiLSTM-CRF 模型的作用是为句子中的每个字打上一个标签，其中 BiLSTM 层是长短时记忆神经网络，用来给每个字的每种可能标签打出一个评分，而 CRF 层是一个条件随机场模型，用来选择其中一组最为合理的标签作为结果。

数据标注是实体提取中最为基础的工作，只有通过一定量数据的训练，才能让模型“学会”从句子中提取出实体及其类别。目前在某些领域已经有一些公开可用的标注数据集，但是建筑领域仍然十分缺乏这样的数据。考虑到人工标注数据的工作量较大，我们在建筑机电领域相关规范中通过正则匹配和简单人工筛查等方式整理了 7 类实体（部位、场所、设备、属性、行为、系统、修饰）。生成自动标注数据的思路是：将这些实体词语在原始数据中进行匹配，如果匹配上则认为句子中的该词语表示该实体且属于对应的分类。

我们将前 80%的数据作为训练集，后 20%的数据作为测试集。研究中先采用部分数据集进行试训练和评估，并结合一些已有研究的经验，最终调整主要超参数如下表。

表 1 BiLSTM-CRF 模型超参数设置

超参数	字向量维度	隐藏层向量维度	学习率	权重衰减	训练轮数
值	300	300	0.001	0.0001	15

完成以上工作之后，将训练集的数据传入到模型中进行训练，之后便得到了最终的模型。训练好的模型在测试集中取得了 96.21%的精确率和 95.19%的召回率。模型预测结果的精确率和召回率都在 95%以上，说明 BiLSTM-CRF 模型具有很好的性能，且能够较好的用于建筑领域的实体提取。

3.2 关系抽取

关系抽取的目的是从文本中抽取出实体与实体之间的关系，因此通常要在实体提取的基础上进行。我们选择了 ResCNN（深度残差卷积神经网络）模型进行关系抽取部分的研究，它是一种用于弱监督关系提取的深度残差卷积神经网络[13]。ResCNN 模型的作用是在句子中捕捉实体对的关联信息，从而给出实体对最可能的关系类别。此外，由于 ResCNN 模型在判断实体对关系类型时参考了所有相关句子的信息，因此能够极大的减少噪声数据带来的影响。

同样由于建筑领域中关系标注数据的缺乏，我们尝试通过某种方式来自动生成标注数

据，以完成 ResCNN 模型的训练。在设计实体的 7 个类别时，我们考虑了类别对关系抽取的辅助作用，即通过两个实体的类别信息来辅助校验它们的实际关系，例如在“A 具有属性 B”这样一条关系中，实体 B 的实体类别应当为属性，否则这条关系很可能是错误的。由于实际关系的分类较为复杂，为了简化自动标注的流程，我们定义了一种抽象的关系：处于同一句子中的两个不同类别的实体即认为是有关系的。此外，如果一个词对在句子里共同出现的次数很少，那么这两个实体词语有可能是偶然出现在同一个句子里的，我们认为一方面它们的关系不明显，另一方面应用的机会也很小，因此需要对这些词对进行过滤。基于 Apriori 算法，我们以出现次数大于 10 作为频繁项集的条件，并把共同出现次数大于 10 的所有词对过滤出来作为有关系的实体对。

标注数据集同样被划分为 80%的训练集和 20%的测试集，模型的主要超参数最终调整如下表。

表 2　ResCNN 模型超参数设置

超参数	词向量维度	句子最大词数	训练轮数	Batch Size	Dropout
值	100	100	30	64	0.5

其中，设置 Dropout 参数是为了避免模型出现过拟合，其值为 0.5 意味着每次迭代中每个单元有 50%的概率被丢弃。将训练集的数据传入到模型中进行训练，完成后即得到了最终模型。模型最终取得了 95.72%的精确率和 95.72%的召回率，说明该模型在建筑领域的关系抽取方面具有良好的性能。

4　基于知识图谱的 BIM 模型审查

通过实体提取和关系抽取，我们得到了机电领域中的一些实体和关系，它们构成了一个简单的知识图谱。在我们定义的关系之中，“设备_属性”关系是一种比较直观的关系，因为当设备实体 A 和属性实体 B 经常出现在同一句子中时，那么属性 B 很可能是针对设备 A 而言的，或者说“A 具有属性 B”。例如，对于“变压器”这一设备，查询与其具有“设备_属性”关系的实体可以得到部分结果如下。

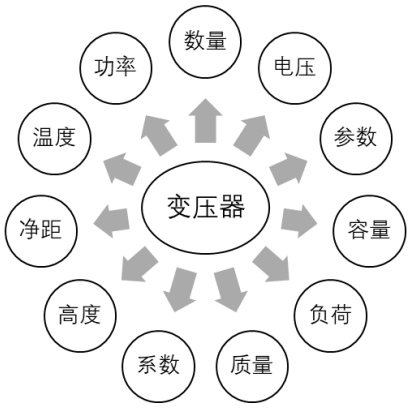

图 2　与变压器具有“设备_属性”关系的部分属性

从图 2 中可以看到，在我们构建的知识图谱中，变压器与许多属性是有关联的，如“高度”、“负荷”、“容量”等等，这些也确实是变压器所具有的属性。对于不了解变压器的人来说，通常很难知道它包含哪些属性，也就很难找到对应的约束条件。而根据这一查询结果，我们可以方便的了解规范中所描述的变压器属性。

进一步的，对于变压器的某一特定属性，我们可以给出这一条关系的来源语句，从而辅助工作者对 BIM 模型进行审查。例如，对于“变压器”的“净距”这一属性，通过查询得到其部分来源如下，当“变压器”和“净距”两个实体出现在同一句子中时，该句子就被认为是关系来源之一。

当露天或半露天变压器供给一级负荷用电时，相邻油浸变压器的净距不应小于5m；当小于5m时，应设置防火墙。

变压器的外廓与围栏的净距不宜小于0．6m，变压器之间的净距不应小于1．0m。

4．2．4本条规定的净距仅为巡视通道，不考虑变压器的就地检修条件。

规定的变压器外廓与围栏的净距和变压器之间的净距，是考虑安全运行和巡视的需要。

因为油浸变压器的火灾危险性比可燃液体贮罐大，它又是变电设备中的核心设备，其重要性远远大于可燃液体贮罐，所以变压器之间最小防火净距应大于0．75D计算数值。

根据变压器着火后其四周对人的影响情况来看，对地面最大辐射强度是在与地面大致成45°的夹角范围内，要避开最大辐射温度，变压器之间的水平净距必须大于变压器的高度。

4．5．9变压器外廓(防护外壳)与变压器室墙壁和门的净距不应小于表4．5．9的规定。

4．5．10多台干式变压器布置在同一房间内时，变压器防护外壳间的净距不应小于表4．5．10及图4．5．10-1和图4．5．10-2的规定。

图 3　规范中变压器与净距的关系来源语句

根据这些句子，我们可以很快得出变压器的净距所需满足的约束条件，如一般变压器之间的净距不应小于 1.0m、与围栏的净距不宜小于 0.6m 等，从而对 BIM 模型进行相应的检查。通过关系来源查找，我们能够直接得到反映实体对关系的所有句子，而不需要再到规范中去人工检索，从而节约了翻阅、查找约束语句的时间，提高了人工审查的效率。

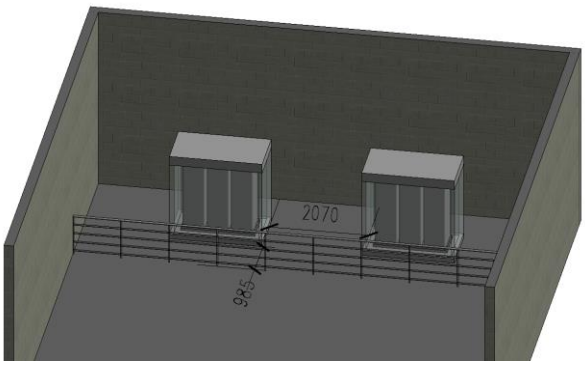

图 4　三维模型中变压器的净距

通过应用知识图谱技术，我们实现了从机电领域相关规范中自动提取设备和设备所具有的属性，并进一步对“设备_属性”关系进行来源查找。即使是不熟悉规范的工作者，在进行 BIM 模型的审查工作时，也能够方便的了解该设备所具有的属性，并快速查找到规范

对相关属性的约束语句。

在本研究的基础上，可以考虑进一步对相应属性的属性值进行提取，例如根据“变压器”、“净距”提取出对应的属性值“不小于 1.0m”，从而减少工作者阅读、理解规范中句子的时间，进而实现部分内容的自动化审查；另外，从图 3 中也可以看到，规范中的一些约束条件是以图片、表格的形式给出的，仅依靠文本数据来构建知识图谱并不全面，因此可以考虑进一步对规范中的非文本内容进行解析提取，从而更加充分的利用已有数据。

5 总结

本文研究了知识图谱的自动构建技术，选用了 BiLSTM-CRF 和 ResCNN 两个模型，并分别应用于机电领域相关规范中的实体提取和关系抽取工作，结果表明它们具有良好的性能，能够较好的应用在建筑领域知识图谱的自动构建工作之中。在此基础上，我们将得到的实体和关系应用到规范约束条件的自动提取之中，在一些机电设备及其属性上取得了理想的效果，说明这一方法对于 BIM 模型的审查能够起到有效的辅助作用。

参 考 文 献

[1] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(04):589-606.

[2] 房栋. 高校知识图谱的构建与数字资源分配新融合[J]. 中国信息技术教育, 2018(Z3):164-167.

[3] 任飞亮, 沈继坤, 孙宾宾, 等. 从文本中构建领域本体技术综述[J]. 计算机学报, 2019, 42(03): 654-676.

[4] 王元卓, 贾岩涛, 刘大伟, 等. 基于开放网络知识的信息检索与数据挖掘[J]. 计算机研究与发展, 2015, 52(02):456-474.

[5] Chinchor N , Marsch E . MUC-7 Information Extraction Task Definition[C]// A Seventh Message Understanding Conference. 1998.

[6] Rau L F . Extracting company names from text[C]// Artificial Intelligence Applications, 1991. Proceedings. Seventh IEEE Conference on. IEEE, 1991.

[7] Liu X , Zhang S , Wei F , et al. Recognizing Named Entities in Tweets[J]. Acl, 2011, 1:359-367.

[8] 刘克彬, 李芳, 刘磊, 等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007(08):136-141.

[9] Guo J Y, Li Z , Yu Z T , et al. Extraction and relation prediction of domain ontology concept instance, attribute and attribute value[J]. Journal of Nanjing University(Natural Sciences), 2012.

[10] Wu F , Weld D S . Autonomously semantifying wikipedia[C]// Sixteenth Acm Conference on Information & Knowledge Management. ACM, 2007.

[11] 吴浪韬. 建筑机电领域知识的自动获取[D]. 北京: 清华大学, 2019.

[12] 吴雪峰, 赵志凯, 王莉, 等. 煤矿巷道支护领域知识图谱构建[J]. 工矿自动化,2019,45(06):42-46.

[13] Huang Y Y, Wang W Y. Deep residual learning for weakly-supervised relation extraction[J]. arXiv preprint arXiv:1707.08866, 2017.