面向智能审图的规范条文命名实体识别

周育丞，郑哲，林佳瑞*，杨梓艺，陆新征

（清华大学土木工程系，北京 100084）

【摘　要】规范条文是建设工程各个阶段的知识依据，如何借助计算机自动构建规范的知识图谱，以及对规范进行自动规则检查，对工程建设智能化有重大意义。命名实体识别是指从一段文本中提取出具有特定意义的实体，是提取审图规范信息、构建知识图谱等的重要环节。本文针对建筑设计规范，定义了十种实体类别，并提出了基于深度学习的规范条文命名实体识别方法。实验结果表明该方法准确有效，且相较于传统基于统计的方法具备一定优势。

【关键词】命名实体识别；规范条文；深度学习；自然语言处理；智能审图

1　引言

命名实体，是指人为定义的一种标签，用于区分文本中的不同信息类别。命名实体识别是利用计算机技术从一段文本中识别出不同的实体信息，并打上相应的标签进行分类[1]。这个概念最早于 MUC-6 会议上提出，是“信息提取”的一项子任务，研究的主要目的是作为“关系抽取”的前置任务，提取出文本中的有效实体信息，并用实体标签来分类。对规范条文的这类研究属于特定领域的命名实体识别，旨在从规范文本信息中识别出具有特定意义的实体，从而为提取实体间的关系提供支撑，可以为土建行业规范知识库建设、智能规范审查、情报分析和数据挖掘等上层应用提供重要的支持[2]。

对于土建行业而言，知识图谱的构建非常依赖规范条文的内容，所以需要将规范条文中的数据结构化地提取出来，即进行知识抽取，而命名实体识别作为知识抽取的关键步骤之一，虽然可以采用人工为主、计算机算法为辅的方式完成，但是当需要处理的数据过多时，人工处理数据就会缺乏效率。特别是土建行业有着数量庞大、种类繁杂的规范条文，仅仅依靠人力完成更为困难。基于上述现状，本研究旨在采用计算机与人工智能技术，研究基于深度学习技术对规范条文进行命名实体识别的方法，以此作为知识图谱构建、规则检查、智能审图等上层应用的基础。

【基金项目】国家自然科学基金资助项目(51908323，72091512)，清华大学-广联达 BIM 联合研究中心(RCBIM)

【作者简介】林佳瑞(1987-)，男，助理研究员。主要研究方向为智能建造、BIM/CIM 与数字防灾技术。E-mail：lin611@tsinghua.edu.cn

2 规范的命名实体类别定义

进行合理的实体类别定义，或者说标签定义，是命名实体识别任务最重要的步骤之一，通常需要从两个方面进行考虑，一个方面是专业术语覆盖面，即在土建行业的规范条文范围内，应该定义怎样的实体类别来尽量覆盖到文本中的所有信息；另一个方面就是识别结果语义丰富度，也就是命名实体识别应为上层应用的服务对象提供足够语义丰富度的识别结果。在实体类别定义的过程中，不仅要时刻考虑是否满足以上两个方面，也要结合具体的实例来进行，不断优化标签。

本文从体现实体间的内在关系出发，来进行命名实体的分类定义。对于土建行业的规范而言，每一条单一的规范条文一般可以理解为“对某一对象进行某种规则检查”，或者说描述了“某一对象应满足某种条件”，可以进行抽象化的描述为：“某一对象（obj）”的“某项属性（attr）”“等于（=）”“某属性值（attrV）”。这里的“等于（=）”仅仅是一种抽象化的表达，并不代表“等号”，可以将之理解为某种“行为（Behavior）”（或称约束行为），描述了某项属性和对应属性值之间的关系。另外，这里的“属性值描述（attrV）”也并不局限于数值，还可能是非数值类属性值。从这种角度出发进行命名实体的定义，能比较清晰地提取出规范条文的核心信息，将之转换为一种统一的规则检查公式，从而方便计算机进行自动的规则检查，也能提取出实体间的关系，服务于知识图谱的构建。

在上述基础可以将实体类别进行进一步细分。一般而言一条规范文本是围绕着某一主体进行规则检查，可以将该文本中描述的主体定义为“主要对象（mObj）”，得到第一个实体标签；而一条规范中可能会描述该主体的一个或多个“属性（Attr）”，可以定义为第二个标签；而对于“等于（=）”或者说“行为（Behavior）”来说，可以将其分为“比较行为（cBe）”和“非比较行为（oBe）”，从而得到第三和第四个标签；而“属性值描述（attrV）”正如前面所提到的，不仅仅代表数值，结合规范条文的具体案例，可以划分为四类标签，分别是“数值属性值（NAttrV）”、“单位（UAttrV）”、“非数值属性值（RAttrV）”和“动作补充（AAttrV）”，其结构如图 1 所示。

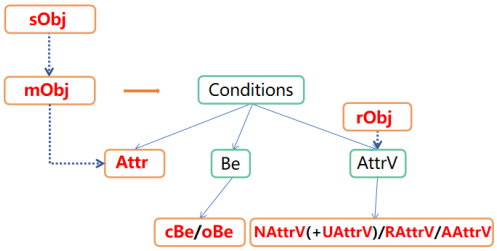

图 1 实体关系结构图

下面通过一个示例进一步说明。如图 2 所示，该条规范文本描述了主体“丙类建筑（mObj）”，当其属性“高度（Attr）”满足“不大于（CBe）”“24（nAV）m（uAV）”时，其属性“结构类型（Attr）”应该“不宜采用（OBe）”非数值属性值“单层框架结构（rAV）”。注意到第二个属性“结构类型”其实被省略了，这是规范文本中经常出现的现象，如果将这条规范识别的结果按前文所述的规则检查公式进行格式化，就可以得到：“丙类建筑的高

度大于 24m”以及“丙类建筑的结构类型不宜采用单跨框架结构”。

除了前文基于“obj.attr = attrV”这一基本公式细分定义的 8 个标签，再额外定义一个描述文本中不同对象的实体类别：“引用对象（rObj）”，它是指在描述“主要对象（mObj）”的“属性（Attr）”满足什么条件时，可能会在“属性值（attrV）”部分出现引用的“其他对象”。最后，在前述 9 个标签的基础上，再额外定义一个“上级对象（sObj）”，它是指该条规范文本描述的“主要对象（mObj）”的“父类对象”，即该“主要对象（mObj）”在其他规范文本中可能是该“上级对象（sObj）”的“属性（Attr）”。

图 2 规范条文标注示例

此外，在对某一规范条文进行规则检查时，有时不是直接判断某一属性满足什么条件，而是会出现“前置条件”，即在满足前一条件的情况下，再满足后置条件，即类似 IF-THEN 格式的条文说明。考虑到如果对标签区分是属于 IF 部分还是 THEN 部分会增加大量标签数量，本研究决定不在实体识别过程中区分前置条件（IF）与后置条件（THEN），而是考虑可以在实体识别的结果上通过关键词匹配或是其他算法，来区分出前置条件与后置条件，例如检查该条件是“当……时……”等句式或者是定语从句时，可以判断为前置条件。

3 规范条文命名实体识别数据集构建

在命名实体识别研究中，数据准备工作一般分为文本爬取和文本标注，从而得到作为神经网络模型输入的训练数据。考虑到土建行业的规范条文数量庞大，且内容繁杂，如何选择合适的规范条文，如何进行适当的数据处理，都是非常重要的前期工作。同时，在模型训练阶段，也可以根据训练结果针对性地对数据集进行一定的优化处理，从而提高识别效果。

实验前期的数据准备阶段，首先要考虑的是规范文本的选取。为保证能得到足够的数据集，本研究的训练数据均选自文本量较多的《建筑抗震设计规范》GB50011-2010（2016 年版）[3]，其中规范条文通过网络爬取等手段获取，并应用一些文本结构化等操作进行预处理[4,5]。考虑到规范条文中一般会带有大量的图片和表格内容，且其包含的信息一般为数学公式、结构图例等，不适合作为命名实体识别任务的研究对象，所以本研究的训练数据只包含人工处理后的纯文本数据。对《建筑抗震设计规范》进行文本提取后得到的 TXT 文件中，考虑选取的文本应该是有效文本，一条有效文本即一句包含完整信息的句子，以句号或分号作为分隔；而初步提取到的 TXT 文件中仍包含不少无效数据，需要删除其中的无效行，比如章节信息、符号注释等，同时，也需要对一些长难句进行人为划分，保证文本标注的合理性。

在得到规范条文的文本数据，并定义实体分类标签后，需要用这些标签对文本进行人工标注，本研究使用了文本标注工具 Doccano[6]来完成标签标注。之后将标注后的文本导出为 JSON 格式文件作为数据集。

经过人工标注，最终得到了 810 条完成标注的规范条文。由于规范文本的语言特性，部分定义的标签在数据集中占比较小，如 sObj 和 aAV，这些数据量较小的标签通常在模型中的训练效果较差，识别结果也显著低于其他数量较多的标签，所以需要采取一定方式来增加这些标签的数量。比如可以人为地多选取包含这些实体的规范文本，但是因为这类文本本身就在规范中占比较少，所以效率不高。考虑到规范条文其实体的可替性，可以采用文本替换的方式来进行数据集的增广。比如，某条规范文本内的 sObj 为“部分框支抗震墙结构”，可以在保持文本其他部分不变的条件下，将之替换为“部分框支剪力墙结构”等相似内容，从而得到一句新的带有 sObj 标签的文本。通过这种不改变文本结构，只替换部分内容的方式，可以快速地实现低频实体标签的扩充，从而达成数据集增广和标签平衡的目的，最后，用于实验的数据集总数为 1000 条规范文本。经数据增广前后标签数量的分布如图 3 所示。

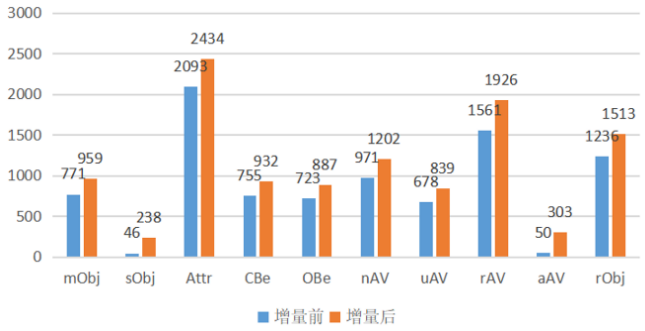

图 3 数据集增广前后的标签数量分布

4 基于深度学习的命名实体识别

随着深度学习技术的发展，越来越多的深度学习模型被提出，如循环神经网络（RNN）、卷积神经网络（CNN）等，它们在不同应用场景（如自然语言处理）有各自的长处。其中，基于 RNN 的各种变体在命名实体识别任务中得到了广泛应用，如长短时记忆神经网络（LSTM），其通过设置四种门结构对信息进行筛选，选择部分遗忘或记忆，这使其相较 RNN 能记住一些更久远的信息。若将两个 LSTM 模型融合在一起，分别接受正序和逆序的输入，则可以合成一个双向 LSTM（BiLSTM）模型，并往往可以提高性能。隐马尔可夫模型（HMM）[7]是一种概率图模型，它通常包含若干个随机变量，又被称为状态集，变量之间通过链式结构相连，所以又被称作隐马尔可夫链。条件随机场（CRF）[8]属于马尔可夫随机场的特例，是一种判别式概率模型；可以加在 LSTM 模型的输出层上，对结果进行进一步修正以提高性能。为了对比，本文实现了四种命名实体识别模型，分别为：基于 HMM、基于 CRF、基于 BiLSTM、基于 BiLSTM-CRF[9]。其中最后一个模型 BiLSTM-CRF 的架构如图 4 所示，主要分为输入层、字向量映射层、BiLSTM 层、CRF 层以及输出层五个部分。

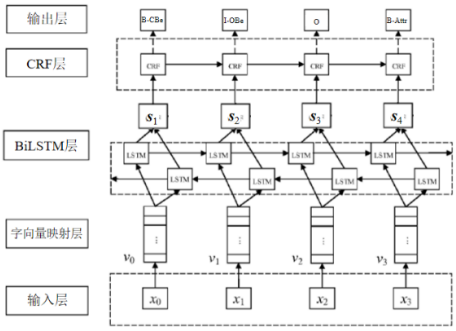

图 4　BiLSTM-CRF 模型架构

根据构建的 1000 条文本组成的数据集，按 6:2:2 的比例划分为训练、验证和测试集。本实验的运行环境如下：编程语言为 Python 3.8，深度学习框架为 Pytorch 1.7，GPU 并行运算平台为 CUDA 11.0，CPU 为 Intel i7-6700HQ。实验中性能评价指标采用精确率（Precision，P），召回率（Recall，R）以及 F1 值，三项指标的计算公式如下。其中，N_c 为被正确识别的命名实体个数，N_r 为被识别出的命名实体个数，N_t 为数据集中命名实体的总数。

$$P = N_c/N_r \tag{1}$$

$$R = N_c/N_t \tag{2}$$

$$F_1 = 2PR/(P+R) \tag{3}$$

图 5 展示了实验中 BiLSTM-CRF 深度学习模型训练的轮数（Epoch）和损失函数（Loss）之间的变化关系。从图中可以看出，蓝色线代表的训练集 Loss 随 Epoch 的增加逐渐降低，其曲线在第 15 个 Epoch 后已趋于平缓，而橙色线代表的验证集 Loss 在第 6 个 Epoch 处达到最低，之后缓慢上升，说明模型已过拟合，在多次重复试验后，可以得出结论，对模型训练 8 个 Epoch 比较合适。同理，可以得到模型训练的各项参数指标的最佳值。经实验测试，最终 BiLSTM 模型训练取的最佳参数值为：Epoch=8，Batch size=8，Learning rate=0.005，Hidden size=240，Dropout rate=0.35。

最终，各个模型的性能实验测试结果如图 6 所示。可以看出，BiLSTM-CRF 模型有最佳的性能，命名实体识别可以达到 86%精确度，86%召回率，和 85%的 F1 值。该结果符合预期，也证明了基于深度学习的命名实体识别相较于传统方法（如 HMM）具备一定优势。

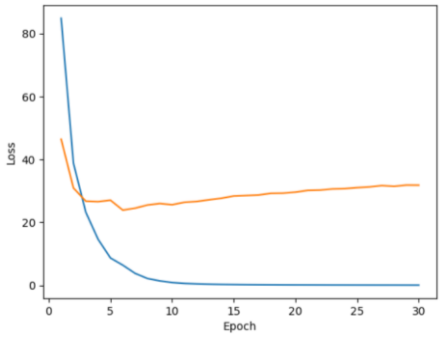

图 5　模型训练的 Epoch-Loss 变化曲线

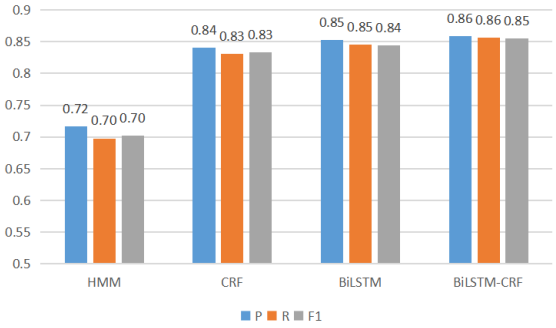

图 6　模型性能对比结果

5 结论

本文围绕土建行业的设计规范，研究了基于深度学习技术的命名实体识别方法。在调研文献的基础上，定义了十种命名实体类别，通过人工标注和数据集增广来构建数据集，并搭建了几种不同的模型来进行命名实体识别及性能优化和分析。实验结果表明，本文所定义的标签分类具备合理性，所构建的基于深度学习的命名实体识别方法准确有效，且相比传统方法具备优势。其中，基于 BiLSTM-CRF 的模型为四种模型中效果最好，在构建的测试数据集上达到了 85%的 F1 值。

本研究的工作尚有许多可改进之处值得在今后进行研究完善，如提高数据的标注质量或构建更大数据集来提升识别效果，改善实体分类标签以制定更细致的实体分类标准，以及引入无监督、弱监督等深度学习训练技术以进一步提升模型性能。

参 考 文 献

[1] 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010, 26(6).

[2] MOON S, LEE G, CHI S, 等. Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing[J]. Journal of Construction Engineering and Management, 2021, 147(1)

[3] GB50011-2010. 建筑抗震设计规范[S].

[4] 林佳瑞, 廖盤宇. 面向法规智能的消防规范图谱构建及应用初探[C]//第六届全国 BIM 学术会议论文集. 2020.

[5] ZHOU Y, LIN J, SHE Z. Automatic Construction of Building Code Graph for Regulation Intelligence[C]//Proceedings of the International Conference on Construction and Real Estate Management 2021.

[6] NAKAYAMA H, KUBO T, KAMURA J, 等. Doccano: Text Annotation Tool for Human[EB/OL](2018). https://github.com/doccano/doccano.

[7] 祝继锋. 基于 SVM 和 HMM 算法的中文机构名称识别[D]. 吉林大学, 2017.

[8] 单赫源, 张海粟, 吴照林. 小粒度策略下基于 CRFs 的军事命名实体识别方法[J]. 装甲兵工程学院学报, 2017, 31(01): 84-89.

[9] 张子睿, 刘云清. 基于 BI-LSTM-CRF 模型的中文分词法[J]. 长春理工大学学报(自然科学版), 2017, 40(04): 87-92.