基于法院判决文书的法律知识图谱构建和补全

王　宁，刘　玮，兰　剑

(武汉工程大学 计算机科学与工程系　湖北 武汉 430205)

摘要：由于法律领域知识图谱专业性强、结构复杂，而现有的关系抽取方法因各个领域的需求和术语不同，无法适用于法律领域知识图谱的构建和补全。首先，提出了基于 StanfordNLP 关系抽取机制的法律知识图谱构建方法；然后，构建基于设置谓语导向词的深度学习模型对法律知识图谱进行补全；最后，选用典型案例(伪卡盗刷判决书)作为文本对象验证模型的可行性。与其他知识图谱补全模型相比，本模型的准确率达到 95%以上。基于谓语导向词的深度学习模型综合了自动构建和人工参与，提高了关系抽取的准确率和补全的效率，能最大程度挖掘判决书文本中的深层隐式关系，更好地发挥判决书文本的应用技术。

关键词：关系抽取；领域术语；知识图谱构建；深度学习

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-6841(2021)03-0023-07

DOI：10.13705/j.issn.1671-6841.2020304

0　引言

谷歌提出的知识图谱概念作为语义网的升华，是结构化的语义知识库，以符号形式描述物理世界中的概念实体及其相互关系，是从“关系”的角度来分析问题并为搜索提供了新思路。所以关系抽取是知识图谱构建和补全的关键一步。通常实体关系抽取使用最多的是通用实体关系抽取和领域实体关系抽取[1-2]。完全自动化的通用实体关系抽取存在很多错误，难以构建出准确的法律知识图谱。如盛美伦提出一种句子卷曲法来抽取可接受解的字串[3]。这种非监督的外部方法需要大量的空间和时间开销，且不能保证一定能找到可接受解的字串。而领域实体关系的抽取是目前人们研究的重点和难点，通常采用的是基于特征法和规则法。如周琦等设计了一种语义方法 GeoRSG 来表现地理试题之间的位置关系，然后用规则法实现地理位置关系在文本中的语言表达方式[4]。国内对法律本体的研究较少，关于法律领域知识图谱的构建目前只有陈淑燕提出的一个简单法律知识框架[5]：使用法规知识库的方法来分析案件。这种方法的缺点是随着案例库的增大，效率明显下降，并且不能提供明确的语义信息，适用性不强。

知识图谱的补全就是向一个已有的知识图谱中增加新的三元组来不断完善和丰富知识图谱的内容。用于补全知识图谱的信息有：从一个知识图谱已有的三元组来推理新的三元组、从文本中抽取新的实体和三元组。传统方法有以 TransE[6]模型为代表的基于翻译转化的知识图谱补全算法，核心思想是从头实体到尾实体的翻译操作，之后在其基础上演化出了 TransH、TransR。另一种是基于关系路径的知识图谱补全算法，即在知识图谱中添加一条边(关系)来连接两个点(实体)。这种算法能够合理解释预测结果，但是无法在低连通图(关系稀疏的知识图谱)上进行有效计算。

已有的研究有王祯基于嵌入模型的知识图谱补全提出的一种多任务联合学习框架下的嵌入模型，该模型是对所有关系事实三元组进行处理[7]；唐慧琳提出的基于融合三角形子图的嵌入表示模型及实体间语义关联进行旅游知识图谱的补全[8]；罗琦提出的基于实体描述和关系路径的知识图谱补全[9]。王祯的方法仅使用了无标注的语料库，所谓的结合就是模型预测结果的直接合并，较为粗浅，且基于多源的文本；唐慧琳主要设计了一个基于垂直领域知识图谱的景点推荐问答系统，领域性过强，无法直接借鉴使用；而罗琦的算法是基于关系路径补全的典型代表，其最大的缺点就是需要有一个较为完善的现成的知识图谱来进行训练。

收稿日期：2020-09-23

基金项目：湖北省技术创新专项重大项目(2019AAA045)；湖北省自然科学基金项目(2019CFB172)。

作者简介：王宁(1997—)，女，硕士研究生，主要从事自然语言处理和深度学习研究，E-mail：1757674599@qq.com；通信作者：刘玮(1981—)，女，副教授，主要从事智能软件工程、服务计算、语义计算研究，E-mail：liuwei@wit.edu.cn。

因此本文在以上问题的基础上,以“伪卡盗刷判决书”为研究对象,目标是为每一份判决书文本构建出的知识图谱进行自动补全。主要的贡献有:

1) 整合了基于 StanfordNLP(斯坦福自然语言处理包)的伪卡盗刷知识图谱构建流程,实验结果验证了该流程的可行性与有效性,为下一步的补全工作提供了数据基础;

2) 提出了一种基于谓语导向词的深度学习模型,用来对 1)中建立的伪卡盗刷知识图谱进行补全。相比于传统的词袋模型(bag of words)和 word2vector 能更好地表达语法信息。

1 相关研究综述

本节介绍关系抽取过程中面临的主要问题:构建法律领域本体、语义标注、构建三元组。

1.1 本体构建

本体的构建[10-11]复杂且重要,是构建知识图谱的基础。现在多为手工构建本体,费时费力,特定领域需要专家参与,对于构建较大的知识图谱并不现实。自动化构建的结果受训练集、数据集以及训练方法的影响大,正确率低。本文采用半自动构建来弥补上述两种方法的不足。首先使用 python 中的模块对判决书进行分词得到法律领域的关键词集,再使用自然语言处理工具进行诸如词性标注、命名实体识别、关键字抽取等预处理。最后在法律领域专家的指导下(人工干预)构建出高质量的领域本体。

1.2 语义角色标注

语义角色标注[12]是一种浅层语义分析技术,其任务就是以句子的谓词(通常是动词)为中心,研究句子中各成分与谓词之间的关系,并用语义角色来描述这种关系。序列标注是语义角色标注的基础工作,包括分词、词性标注、实体识别和依存分析。解决的方法有传统法和深度学习法。传统法是采用条件随机场[13](conditional random field,CRF)模型来针对序列数据进行分类;长短期记忆网络[14](long short-term memory,LSTM)是深度学习的主要方法。由于两种方法各有利弊,所以目前最好的方法就是结合其优点,先用 LSTM 自动抽取特征,再通过 CRF 进行序列数据标记,也就是在 LSTM 的输出层中再加一个 CRF 层。本文采用的是浅层语义分析技术来标注语义角色。

1.3 关系抽取和知识图谱

资源描述框架[15-16](resource description framework,RDF)是以元数据的概念提出的。其形式为三元组,可作为关系抽取结果的一种存储方式。三元组建立的主流方法有有监督的学习法、半监督的学习法和无监督的学习法。有监督的学习法将关系抽取任务作为分类问题。半监督的学习法采用 BootStrapping[17]。无监督的学习法是利用每个实体的上下文信息来代表该实体的语义关系并进行聚类。由于有监督学习法具有严重的依赖性,近年来又有学者提出一种基于深度学习的关系抽取,如 Socher 等提出了使用递归神经网络来解决关系抽取问题[18],以及邵明光用卷积神经网络进行关系抽取[19],还有基于端到端神经网络的关系抽取模型均取得了较大的提升。

三元组是关系抽取结果的一种存储方式,同时也是知识图谱的最小组成单元。从数据结构的角度考虑,知识图谱代表了一张巨大的关系图,而三元组文本形式的事实数据则对应关系图中的边[20]。目前主流方法倾向于人工建立规则和基于统计的方法来从标签信息中抽取关系[21-25]。

2 基于伪卡盗刷判决书文本的知识图谱构建和补全

具体介绍基于 StanfordNLP 的伪卡盗刷判决书知识图谱的构建,主要步骤如下。

Step1 规范化处理。将搜集的数据统一进行规范化处理,得到处理后的数据集。

Step2 本体构建。基于法院判决书及具体刑事判决资料,用统计的方法在法律领域专家以及知识图谱老师的指导建议下,构建出高质量的伪卡盗刷领域本体。

Step3 标注数据。为了完善 Step2 中的伪卡盗刷领域本体,利用浅层语义分析技术对法院判决书进行语义角色标注,得到标注数据。

Step4 关系抽取。通过使用 StanfordNLP 对 Step3 中的标注数据进行实体和关系抽取,得到结构化的实

体与关系集,建立三元组。

2.1　法律本体构建

对法律领域的本体构建需要明确专业术语、关系及其领域,使其形式化以实现一定程度的法律领域知识复用。构建的原则:明确性和客观性(用自然语言定义)、一致性、最小承诺(尽可能少约束)、最小编码偏差以及使用多样的概念层次结构实现多继承机制。本文构建本体的主要步骤如下。

Step1 确定本体的专业领域和范畴。即使是同一个法律领域,应用的本体不同,表示概念的侧重点也会不同(如婚姻案件和盗窃案件)。所以建立本体之前要明确本体建立的领域和应用目标。

Step2 列出本体涉及领域中的重要术语。为了保证准确率,我们采用统计的方法,参考了 1 236 份伪卡盗刷案件的法院判决书,列出其所涉及的重要术语,如在判决书中经常涉及的有刷卡人、刷卡时间、刷卡地点以及报警时间等。

Step3 领域概念分类。领域概念分类层次对应着一棵树,树中的节点体现了领域概念间的层次结构关系:根节点、枝节点、树枝和叶节点。建立领域概念的分类关系后,将分类概念的属性值添加到分类概念中,这样就把领域概念通过树形结构形象地描述出来,并且通过树结构清晰地体现了领域概念间的类属关系——每一个子树都对应着领域中独立的、模块化的知识模型。

Step4 定义概念之间的关系。概念的分类层次结构体现了分类概念间的一种继承关系。但是在领域本体中,概念和概念之间除了通过继承关系来交互,还根据需要定义其他关系。如在本文中,警察和刑警之间应该是相容关系。

根据上述本体的构建原则,本文构建了伪卡盗刷本体中的核心概念(部分)——人物:开户人、盗刷人、银行客服、警察等;报警:电话挂失、电话冻结、银行报警、电话报警等;刷卡:ATM 取现、柜台取现、POS 机刷卡、网上支出等。

2.2　伪卡盗刷判决书的语义角色标注和三元组的建立

本文采用的是基于浅层句法分析结果的语义角色标注,采用传统的三元组保存形式〈主语 谓语 宾语〉。根据浅层句法分析得到的结果来构建三元组。从标签信息中抽取关系,如“宋思宁”的一个标签信息为“2018 年 7 月 16 日在武汉市刷卡购买商品”,我们可以根据这个标签信息推出三种信息框:〈宋思宁　2018 年 7 月 16 日　武汉〉、〈宋思宁　刷卡　2018 年 7 月 16 日〉、〈宋思宁　刷卡　武汉〉。故三元组的保存形式有三种:某人在某地干了什么、某人什么时间干了什么、某人什么时间在某地。这种结构化的三元组为后续知识图谱的建立和补全工作奠定基础。图 1 为法律知识图谱的高层结构(实体和关系的数量可根据需要增减);图 2 为伪卡盗刷知识图谱示例。

2.3　伪卡盗刷判决书知识图谱补全

由于判决法案在不断更新,为了保证构建出的知识图谱具有一定的实用价值,要继续挖掘其潜在的关系对构建出的知识图谱进行补全,使其成为一个不断更新的动态知识图谱。补全工作包括向已有的知识图谱中增加新的实体和关系,以及将新的实体和关系添加到已有的知识图谱中。

我们提出了一种基于谓语导向词的深度学习模型对构建出的知识图谱进行补全。首先,深度学习是学习样本数据的内在规律,然后通过组合学习到的规律来发现数据的分布式特征表示。

我们使用无监督的训练向量来提升其泛化能力,因为有些端到端的方式可以克服传统模型“短距离压制”的缺点。深度学习模型的缺点就是受数据量的影响较大。不过据中国法律网的数据显示,伪卡盗刷的判决书每天都以数万份的量在增长,它们统一书写,统一格式,因此我们有大量的数据进行回标训练。具体的步骤如下。

Step1 构建伪卡盗刷知识图谱。使用 StanfordNLP 进行伪卡盗刷领域的知识图谱构建。

Step2 构建谓语导向词库。抽取伪卡盗刷知识图谱中三元组的谓语动词来构建谓语导向词库。

Step3 设置判决书文本的起止位置。起始位置:法院审理查明;结束位置:本院认为。

Step4 将起止位置中的内容以句子为单位进行编号($a1,\cdots,an$)。

Step5 利用谓语导向词逐个匹配新增判决书文本中的句子。

Step6 将 Step5 中匹配到的句子进行实体和关系抽取,构建三元组,添加到已有的知识图谱中。

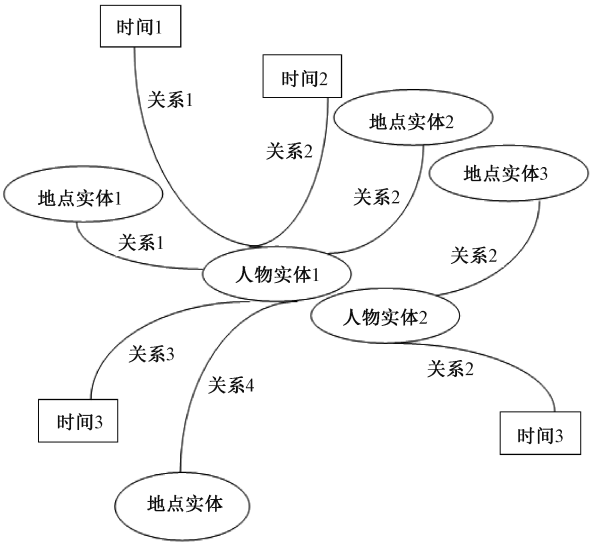

图 1 法律知识图谱高层结构图

Figure 1 The high-level structure diagram of knowledge graph

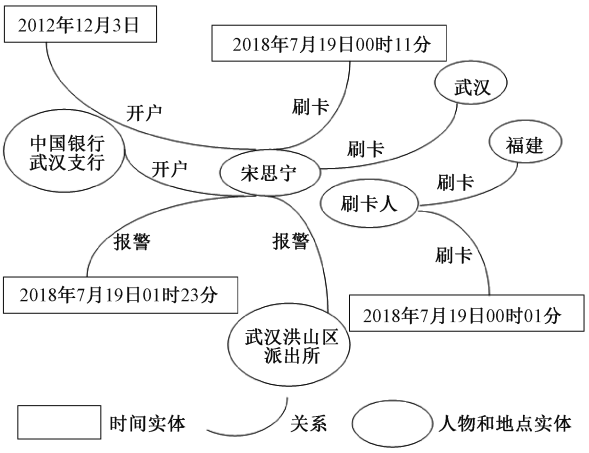

图 2 伪卡盗刷知识图谱示例

Figure 2 Examples of knowledge graph for counterfeiting card theft

Step7 将 Step5 中未匹配到谓语导向词的句子编号 $b1,\cdots,bn$。对 $b1,\cdots,bn$ 筛查,总结出新的谓语动词并判断是否将其加入到谓语导向词库中,将入库的谓语导向词编号 $c1,\cdots,cn$, 并选择构建知识图谱的判决书文本继续进行回标训练,重复操作 Step3~Step6。若没有新的谓语导向词入库,则结束训练。

3 实验

实验数据规模为伪卡盗刷判决书 1 263 份,土地分配判决书 500 份,实体 69 567 个,关系 125 个,三元组的数量为 96 845 个。首先我们采用设置谓语导向词抽取和直接抽取两种方法进行对比,如表 1 所示。随后我们对实验数据添加了噪音,实验结果如表 2 所示。其中我们使用准确率(P)、召回率(R)作为系统性能的评测指标,使用的公式有

准确率(P) = 提取出的正确信息条数 / 提取出的总信息条数;

召回率(R) = 提取出的正确信息条数 / 样本中的总信息条数。

表 1 设置谓语导向词的实验结果对比

Table 1 Comparison of experimental results of setting predicate-oriented words

抽取方法	总数/个	正确数/个	P/%	R/%
设定谓语导向词	95 157	94 867	95.04	97.96
直接抽取	96 152	91 036	94.68	94.00

表 2 添加噪音后的实验结果对比

Table 2 Comparison of experimental results after adding noise

抽取方法	总数/个	正确数/个	P/%	R/%
设定谓语导向词	68 954	65 350	94.77	52.85
直接抽取	113 500	67 511	59.48	54.60

从表 1 的结果来看,效果并不是很明显,体现不出设定谓语导向的优点。这是因为我们的实验数据是比较规范的,领域性太强(每份法律文书都是伪卡盗刷的判决书),所以为了体现这一方面的性能,我们在数据中加了一些噪音(土地分配的法院判决书)进行了新的对比,新的数据规模为:1 000 份伪卡盗刷判决书,500 份土地分配判决书。实验结果如表 2 所示。

表 2 可以很明确地看出添加噪音后的系统抽取性能大幅度提升。这是由于设定了伪卡盗刷的谓语导向,抽取的结果全是可用的伪卡盗刷三元组。而直接抽取因未设置谓语导向,抽取出的三元组中包含很多土地分配三元组,必须通过筛选得到需要的内容。通过设定谓语导向来获取特定的内容是一个很大的改进。实验通过变换噪音的种类不断改进,最后确定出伪卡盗刷的谓语导向词。

我们通过从方法、数量两个方面来分别讨论数据变换对补全率的影响。根据方法的不同,将数据分为训练集、验证集和测试集三个部分。表 3 展示了不同方法的补全效果,表 4 展示了数量的变化对补全率的影

响。本文通过提取的实体、关系以及有效三元组的数量比来衡量补全率。具体的计算公式为

$$补全率(C) = (m/n + a/b + c/d)/3,$$

其中:m 和 n 分别表示抽取出正确实体数量和全部实体数量;a 和 b 表示抽取出正确关系数量和全部关系数量;c 和 d 表示抽取出有效的三元组数量和实际的三元组数量。

表 3　变换方法的补全效果

Table 3　Complementary effect of transformation method

类别	实体/个	关系/个	三元组/个	C/%
TransE	49 534	86	88 512	77.1
TransH	51 655	98	85 632	80.4
TransR	51 859	103	96 741	85.6
DPKR(2-step)	61 732	108	89 987	89.4
FS-M	66 854	121	94 867	97.0

表 4　变换数据规模的补全率

Table 4　Complementary rate of changing data scale　单位:%

数据规模/份	补全率				
	TransE	TransH	TransR	DPKR(2-step)	FS-M
100	43.5	45.5	51.2	31.3	95.2
1 000	67.7	63.4	59.4	53.4	96.4
10 000	72.8	71.7	73.5	81.7	96.5
20 000	75.9	77.4	79.5	88.3	96.5
50 000	76.4	79.5	85.3	88.5	96.3
80 000	76.8	82.2	87.9	88.2	96.9

表 3 列出了各种补全方法的补全率,我们选用当前主流的 Trans 系列和关系路径补全算法进行对比,基于实体描述和关系路径建模的知识图谱补全算法(description and path for knowledge representation,DPKR)是典型的根据关系路径的补全算法。根据文献[9]的实验结果,结合本实验的特性采用 2-step 为实验对比。由表 3 可以看出我们提出的基于谓语导向词的深度学习方法最为有效,由于在关系提取方面规定了谓语为固定动词,故关系的提取率大幅度提升,提取三元组的正确率也有较好的效果。

表 4 可以很清晰地看到数据规模的变换对补全效率有一定的影响。Trans 系列由于需要一定规模的数据来训练关系进行抽取,故三元组的数量对其影响很大。当数据达到一定规模后才能达到理想效果。同理,以 DPKR 为代表的关系路径补全算法在关系稀疏的知识图谱上也无法有效进行。但是我们的方法在固定谓语动词之后,相当于规定了关系抽取的框架,只需要抽取实体进行填充匹配即可。我们的方法在伪卡盗刷这个固定领域进行实验,由于领域粒度较小,故数据规模的变换对补全效果的影响不大,从而提高了效率和准确率。

4　结束语

本文首先总结了知识图谱构建过程中的一般方法,分析每种方法的利弊。通过比较每种方法的优劣,综合考虑各个方法对法律关系抽取中每个环节的影响,加以改进后提出了“谓语导向”的概念用来提高伪卡盗刷领域中三元组的抽取效率以及知识图谱的补全工作。基于谓语导向词的深度学习模型在自然语言处理中能更深层次地挖掘伪卡盗刷判决书文本中存在的丰富语义关系。与传统的词袋模型相比,能够更好地表达语法信息,并取得较好的性能。

我们将继续研究 StandfordNLP 的相关算法来提高构建过程中关系抽取所消耗的时间,提高整体性能。另外,当前实验的范围仅是伪卡盗刷案件且数据来源于一个法院,粒度较小、领域单一,后期将扩大覆盖范围,横向搜集多个法院判决书以扩充数据规模,纵向考虑在其他法律领域中推行此方法。最后,我们考虑将补全后的知识图谱应用到精准推荐、相似案件判决书的自动生成等领域中。

参考文献:

[1]　彭乾慧. 领域知识图谱的自动化构建[D]. 重庆: 重庆大学, 2017.
PENG Q H. Domain knowledge graph auto-construction[D]. Chongqing: Chongqing University, 2017.

[2]　杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法[J]. 软件学报, 2018, 29(10): 2931-2947.
YANG Y J, XU B, HU J W, et al. Accurate and efficient method for constructing domain knowledge graph[J]. Journal of software, 2018, 29(10): 2931-2947.

[3]　盛美伦. 开放领域下复杂文本的关系抽取[D]. 上海: 上海交通大学, 2014.

SHENG M L. Relation extraction from complex text in open domain[D]. Shanghai: Shanghai Jiaotong University, 2014.

[4] 周琦, 陆叶, 李婷玉, 等. 基于语义文法的地理实体位置关系的获取[J]. 计算机科学, 2016, 43(7): 208-216.
ZHOU Q, LU Y, LI T Y, et al. Acquiring relationships between geographical entities based on semantic grammar[J]. Computer science, 2016, 43(7): 208-216.

[5] 陈淑燕, 瞿高峰. 通用法规知识库系统的设计[J]. 计算机工程, 2001, 27(11): 90-91,181.
CHEN S Y, QU G F. The design of a general code knowledge base system[J]. Computer engineering, 2001, 27(11): 90-91, 181.

[6] 方阳, 赵翔, 谭真, 等. 一种改进的基于翻译的知识图谱表示方法[J]. 计算机研究与发展, 2018, 55(1): 139-150.
FANG Y, ZHAO X, TAN Z, et al. A revised translation-based method for knowledge graph representation[J]. Journal of computer research and development, 2018, 55(1): 139-150.

[7] 王桢. 基于嵌入模型的知识图谱补全[D]. 广州: 中山大学, 2017.
WANG Z. Embedding model based knowledge graph completion[D]. Guangzhou: Sun Yat-sen University, 2017.

[8] 唐慧琳. 融合结构和语义信息的知识图谱补全算法研究[D]. 北京: 北京邮电大学, 2017.
TANG H L. Research on knowledge graph completion by combining structural and semantic information[D]. Beijing: Beijing University of Posts and Telecom, 2017.

[9] 罗琦. 基于实体描述和关系路径的知识图谱补全研究[D]. 济南: 山东大学, 2018.
LUO Q. Entity description with relation path modeling for knowledge graph completion[D]. Jinan: Shandong University, 2018.

[10] 翟畅. 面向非结构化中文文本的本体构建[D]. 武汉: 武汉工程大学, 2017.
ZHAI C. Constructing ontology for unstructured Chinese text[D]. Wuhan: Wuhan Institute of Technology, 2017.

[11] 李涛, 王次臣, 李华康. 知识图谱的发展与构建[J]. 南京理工大学学报, 2017, 41(1): 22-34.
LI T, WANG C C, LI H K. Development and construction of knowledge graph[J]. Journal of Nanjing university of science and technology, 2017, 41(1): 22-34.

[12] PALMER M, GILDEA D, XUE N W. Semantic role labeling[M]//HIRST G. Synthesis lectures on human language technologies. Willston: Morgan & Claypool Publisher, 2010:1-103.

[13] 何晓艺. 面向领域文本知识实体识别及关系抽取的关键技术研究[D]. 石家庄: 河北科技大学, 2018.
HE X Y. Key technology research on knowledge entity recognition and its relation extraction for specific domains text[D]. Shijiazhuang: Hebei University of Science and Technology, 2018.

[14] 胡新辰. 基于 LSTM 的语义关系分类研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.
HU X C. Research on semantic relation classification based on LSTM[D]. Harbin: Harbin Institute of Technology, 2015.

[15] 郭喜跃. 面向开放领域文本的实体关系抽取[D]. 武汉: 华中师范大学, 2016.
GUO X Y. Entity relation extraction for open domain text[D]. Wuhan: Central China Normal University, 2016.

[16] 邢美凤. DBpedia 本体知识库关键技术及应用展望[J]. 图书馆理论与实践, 2013(1): 43-46.
XING M F. Key technologies and application prospects of DBpedia ontology knowledge[J]. Library theory and practice, 2013(1): 43-46.

[17] 陈思佳. 实体关系抽取技术研究[D]. 北京: 北京邮电大学, 2014.
CHEN S J. Research on entity relationship extraction[D]. Beijing: Beijing University of Posts and Telecom, 2014.

[18] SOCHER R, PENNINGTON J, HUANG E H. Semi-supervised recursive autoencoders for predicting sentiment distributions [C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Strasbourg: Computational Linguistics Association, 2011:151-161.

[19] 邵明光. 基于深度卷积网络的知识图谱补全模型[D]. 哈尔滨: 哈尔滨工业大学, 2018.
SHAO M G. Deep convolutional neural network for knowledge graph completion[D]. Harbin: Harbin Institute of Technology, 2018.

[20] 鄂世嘉, 林培裕, 向阳. 自动化构建的中文知识图谱系统[J]. 计算机应用, 2016, 36(4): 992-996,1001.
E S J, LIN P Y, XIANG Y. Automatical construction of Chinese knowledge graph system[J]. Journal of computer applications, 2016, 36(4): 992-996,1001.

[21] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2008: 1247-1250.

[22] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia-a crystallization point for the web of data[J]. Journal of web

semantics, 2009, 7(3): 154-165.

[23] XU B, XU Y, LIANG J Q, et al. CN-DBpedia: a never-ending Chinese knowledge extraction system[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Cham: Springer International Publishing, 2017: 428-438.

[24] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge unifying wordnet and wikipedia[C]//Proceedings of the 16th International Conference on World Wide Web. New York: ACM Press, 2007: 697-706.

[25] 鲍晓. 领域本体概念及概念间关系学习算法研究[D]. 武汉: 华中科技大学, 2013.
BAO X. Research on domain ontology concepts and relations learning algorithm[D]. Wuhan: Huazhong University of Science and Technology, 2013.

Construction and Completion of Legal Knowledge Graph Based on Court Judgment Documents

WANG Ning, LIU Wei, LAN Jian

(*School of Computer Science and Engineering*, *Wuhan Institute of Technology*, *Wuhan* 430205, *China*)

Abstract: Due to the specialty and complex structure of the legal-domain-knowledge-graph, the existing relationship extraction methods were not suitable for the construction and completion of the legal-domain-knowledge-graph because of the various needs of different domains. A legal-knowledge-graph-construction method based on StanfordNLP relation extraction mechanism was proposed. Then, a deep learning model based on predicate-oriented words was proposed to complete the legal knowledge graph. At last, a typical case (the judgment of counterfeiting card theft) was selected as the text object to verify the feasibility of the model. Compared with other knowledge graph completion models, the accuracy of this model exceeded 95%. The deep learning model based on predicate oriented words integrated automatic construction and human participation, which improved the accuracy of relation extraction and the efficiency of completion. This method could excavate the deep implicit relationship in the text of the judgment to the greatest extent and achieved better application of judgment text technology.

Key words: relation extraction; domain term; knowledge graph construction; deep learning

(责任编辑:王浩毅　方惠敏)