控制与决策
Control and Decision

领域专业知识富关联关系提取方法

李青， 钟将， 李立力， 张剑， 李琪

引用本文:
李青, 钟将, 李立力, 等. 领域专业知识富关联关系提取方法[J]. 控制与决策, 2021, 36(1): 52−60.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0727

领域专业知识富关联关系提取方法

李 青[1], 钟 将[1†], 李立力[2], 张 剑[3], 李 琪[4]

(1. 重庆大学 计算机学院，重庆 400044；2. 重庆大学 土木工程学院，重庆 400044；3. 重庆西信天元数据资讯有限公司，重庆 401121；4. 绍兴文理学院 计算机科学与工程系，浙江 绍兴 312000)

摘 要: 面向知识服务业中领域专业内容资源的多模态、智能化、精细化、知识化和重组化的碎片性管理需求,如何高效生成和应用专业知识,促进实体经济创新发展,成为共同的战略选择与难题. 对此,重点研究八大战略新兴产业内容资源的富关联体系和知识关系标引规范,制定面向服务专业内容资源的一致性富关联关系的描述体系. 构建内容资源表示实体(知识、信息、资源、服务、对象)间的富关联模式,满足实体间自动解构、聚合及智能抽取的需求,提出基于领域专业知识的富关联关系提取模型. 运用多层注意力机制来凸显重要表征性信息,通过知识图谱设计并优化异构环境中核心源对象与目标对象间元属性. 与以往基线模型不同,所提出的模型结构支持在特定领域下端到端的学习,不必显式依赖外部知识. 实验结果表明,领域专业知识富关联关系提取方法,可有效提升富关联关系识别精度及专业知识服务效率.

关键词: 富关联关系；领域专业知识；关系提取；自然语言处理；人工智能；深度学习

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0727 **开放科学(资源服务)标识码(OSID):**

引用格式: 李青,钟将,李立力,等. 领域专业知识富关联关系提取方法[J]. 控制与决策, 2021, 36(1): 52-60.

Extraction method of multiple semantic relations in domain knowledge

LI Qing[1], *ZHONG Jiang*[1†], *LI Li-li*[2], *ZHANG Jian*[3], *LI Qi*[4]

(1. College of Computer Science，Chongqing University，Chongqing 400044，China；2. School of Civil Engineering，Chongqing University，Chongqing 400044，China；3. Chongqing Xixintianyuan Data Information Co.，Ltd.，Chongqing 401121，China；4. Department of Computer Science and Engineering，Shaoxing University，Shaoxing 312000，China)

Abstract: For the fragmentary management needs of multi-modal, intelligent, refined, knowledgeable and reorganized professional content resources in the knowledge service industry, how to efficiently generate and apply professional knowledge to promote the innovation and development of the real economy has become a common strategic choice and a challenging problem. Therefore, this paper studies the multi-relation classification system and knowledge relationship labeling standard of content resources in eight strategic emerging industries, and develops a consistent multi-relation classification description system for service-oriented professional content resources. Then an extraction model of multiple semantic relations based on domain knowledge is proposed, which can distinguish various entities (e.g., knowledge, information, resources, services, objects), and meet the requirements of automatic deframe, aggregation and intelligent extraction among entities. A multi-level attention mechanism is used to highlight representational details. At the same time, it designs and optimizes meta-attributes between the core source and the target in heterogeneous contexts through the knowledge graph. Unlike previous baseline models, the proposed model structure supports end-to-end learning in the specific domain without explicit dependence on external knowledge. The experiments show that the proposed method can effectively improve the accuracy of the multi-relation classification and the efficiency of professional knowledge service.

Keywords: multiple semantic relation；domain knowledge；relation extraction；natural language processing；artificial intelligence；deep learning

0 引 言

自21世纪以来,知识已成为生产要素中最重要的组成部分,市场经济也逐渐转型为以知识经济为主体的知识再生产、再消费和再分配. 面对全球海量知识内容资源的创新服务发展契机,如何打通全球互联环节中用户生产内容(UGC, user-generated content)与

收稿日期: 2019-05-26；**修回日期:** 2019-07-30.

基金项目: 国家重点研发计划项目(2017YFB1402400)；中央高校研究生科研创新项目(2018CDYJSY0055)；重庆市研究生科研创新项目(CYB18058)；重庆市技术创新与应用示范项目(cstc2018jszx-cyzdX0086)；重庆市重点产业共性关键技术创新专项(cstc2017zdcy-zdyf0150)；陕西省教育厅科学技术研究项目(18JK1130).

†通讯作者. E-mail: zhongjiang@cqu.edu.cn.

专业生产内容(PGC, professionally-generated content)关联壁垒,促进实体经济创新发展,成为各国共同的战略选择与难题. 20世纪90年代起,现代知识服务业逐渐转型为基于提供高智力附加值的专业知识(或技能)的产品或服务. 而这类专业知识是一种高质量的显性知识,是领域专家思维成果的固化. 通过文字、数字、语音、图像、视频等载体形式进行表现,经过专业化机构的生产加工,为消费群体提供知识服务. 然而,如何高效精准地挖掘海量跨领域专业知识实体间的隐性关联关系,是专业化知识服务机构的生产难题[1].

传统关联关系的提取方法是将关系表示为已经识别的两个实体(知识、信息、资源、服务、对象)间的语义关系. 通常此类方法,首先采用命名实体识别技术将文本中的相关实体区别标记,然后结合实体关系层次转换技术和领域知识库查询技术共同参与关联关系构建[2]. 例如,输入“基于贪心策略的自适应蚁群算法在TSP (tiaveling salesman problem)中的应用”,提到带注释的目标实体(entity)对: 实体1(X_1)“自适应蚁群算法”和实体2(X_2)“TSP”. 目标是输出实体对间的因果关系(C-1: 起因事物-结果行为关系). 与传统实体关系不同,面向知识服务中的关联关系形式纷繁复杂,语义隐含性更强. 基于传统关联关系的提取方法难以适应领域专业内容资源的多模态、智能化、精细化、知识化和重组化的碎片性管理需求. 与传统方法相比,新形势下的富关联关系提取技术将打破资源载体间隔阂,使得“内容”成为真正源头,推动实现“按主题”跨模态的专业知识“关联”出版.

富关联关系提取技术是语义理解、 语篇处理和更高层次自然语言处理任务的基础[3]. 然而,高效的富关联关系提取模型准确度常常受限于词法、句法的可变性,以及语用的隐含性. 随着机器学习技术的发展,富关联关系提取方法打破了传统目标实体在词法内容和手工特征上的浅层局限,并逐步向深层架构转变[4-5]. 近年来,绝大多数基线模型采用卷积神经网络(CNN)、递归神经网络(RNN)和其他神经架构,主要关注语义和句法结构特征[6-8]. 尽管如此,这些模型通常需要依赖外部解析器,常常忽略关键实体间的富关联关系. 基于以上分析,本文提出一种基于领域专业知识的富关联关系提取方法——DomulAttCNN (domain knowledge multi-relation classification via attention mechanism CNN)模型,通过多层注意力机制形成句内重要表征性语义信息的凸显图,采用知识图谱设计并刻画异构环境中领域知识链路,有效解决专业知识服务中跨领域分析的困惑性难题. 形成满足实体间自动解构、聚合及智能抽取需求的专业内容富关联关系模型.

本文主要贡献如下:

1) 分析专业知识服务中存在的各种关联类型,重新制定面向服务专业内容资源的一致性富关联关系的描述体系. 采用基于维基百科数据的DBpedia知识图谱[9]刻画异构环境中领域知识链路,设计并优化异构环境中核心源对象与目标对象间元属性.

2) 提出一种新的领域专业知识富关联关系提取方法. 该方法支持在特定领域下端到端的学习,不必显式依赖外部知识. 通过多层感知注意力机制,捕获特定领域实体与关系池间的权重.

3) 提出基于边缘距离的目标函数. 通过多个数据集的实验结果表明,基于边缘距离的目标函数收敛性能优于标准损失函数,具有更好的收敛性.

1　相关工作

富关联关系提取是识别自然语言中句子级的两个实体及其之间关系的任务. 现有的富关联语义关系提取模型大致可以分为以下5类: 人工特征模型[4]、依存树模型[8,10]、端到端模型[7,11]、远程监督模型[12-13]、注意力机制模型[14-15].

以支持向量机(support vector machine, SVM)为代表的人工特征模型,旨在求出n维空间的最优超平面将富关联实体间正负类区分开[4]. 由于其无法支持多分类、在大规模数据中训练困难等缺点,现已被神经网络模型所取代. 依存树模型依赖于显式自然语言预处理步骤中输出计算的一组特征,进而产生最短依存路径(shortest dependency path, SDP). 然而这些SDP通常是基于试误的选择,当自然语言中句子呈现离散结构时,SDP分析极易出错且严重影响模型性能. 同时,此时解析时间会随着句子长度增加而呈指数增长,难以收敛. 随后发展的端到端模型解决了这一问题,其根据未标记的实体间关联关系特征训练目标样本,最终进行概率区分. Zhang等[16]提出了一种基于卷积神经网络(convolutional neural networks, CNN)的端到端模型,这种模型可以自动捕获相关的词汇的句子级别特征. 为摆脱词法资源和自然语言处理工具包的依赖,发展出一系列以CR-CNN[7]、depLCNN+NS[11]、PCNN[13]为代表的模型. 但由于这类方法对实体的主体特征关注度不足,且受限于人工特征标记数据集的规模,而逐渐被取代. 为了解决这一局限性,文献[12]通过启发方式将文本与给定的知识库(knowledge base, KB)对齐来创建大

型数据集的远程监督(distant supervision, DS)假设方法. 由于这种假设并不总是成立,某些句子有可能被错误地标记. 为了解决这些不足,文献[17-19]提出了多实例多标签学习模型,该模型放宽了传统远程监督的多实例单标签学习模型的标签限制. 但因远程监督模型实体间易于重叠,且机械地引入知识库而产生大量错误标签,导致模型难以推广. 近年来发展的基于深度学习的注意力机制可有效填补这一不足,Lin等[14]提出了一种基于实例选择性注意力机制的卷积神经网络PCNN+ATT模型. 随后,文献[15]突破实例选择性局限,提出基于注意力语法实体感知机制的远程监督模型,以提升重要实体间的语法关注力度. 虽然上述模型成功引入了注意力机制,有效提升重要信息的关注力度,但由于单文本信息固化性强,导致模型泛化能力欠佳.

本文利用远程监督模型在特征提取中的优势,着重解决基线模型泛化能力欠佳的问题. 通过领域知识图谱刻画各领域知识链路, 采用多层感知注意力机制捕获特定领域实体与关系池间的权重,弥补了上述模型的不足,并进一步提高了模型泛化能力和新领域模型应用的稳健性.

2 富关联关系类型

为结合中国“十三五”规划中,新一代信息技术、高端装备、新材料等8大战略新兴产业(新能源、新材料、生命生物工程、信息技术和移动互联网、节能环保、新能源汽车、人工智能、高端装备制造)所构建的领域专业内容资源库,特重构富关联关系的形式化描述体系,支持对关联的继承关系、约束特征、依赖关系等进行刻画. 将各领域专业出版社的数字内容资源整合,重点利用领域专业内中外文科研论文、技术成果报告、专业期刊、学术著作、企业标准等现有资源,重新进行专业知识加工、组织和管理. 推进“产、学、研、用”相结合的知识服务新业态,为富关联关系提取打下基础.

结合领域专业内容文献知识实体关联特征,按照相关度划分专业内容的富关联关系类型,可划分为: 1) 同一性关联,是指对实体间所具有的某种程度的相同(或相反)之处所形成的同指关联关系,如同指关系(SR); 2) 隶属性关联,是指构成某实体隶属于某一概念、范畴和类别的逻辑关系,且由实体本身的性质决定,如因果关系(CR)、从属关系(AR); 3) 相关性关联,是指在同一性、隶属性关联之外,实体间具有相互依存、相互渗透、相互制约、相互作用的关系,一般表现为时序与空间域的各种关系. 其关联关系可以是不严格固定,且数量关系也可以不完全确定,如时序相关关系(RT)、空间相关关系(RC).

各领域专业内容存在差异性,此体系留有可扩充关系组,以实现领域知识服务中富关联构建的弹性与规范性(如表1所示).

表 1 专业内容的富关联关系类型描述体系

一级关系	二级关系
同指关系 SR	同义关系 S-1
	反义等同关系 SR-2
	同类并列关系 SR-3
	同类对比关系 SR-4
因果关系 CR	起因事物-结果行为关系 CR-1
	起因行为-结果事物关系 CR-2
	起因行为-结果行为关系 CR-3
	起源事物形式-发展事物形式关系 CR-4
从属关系 AR	上下位从属关系 AR-1
	类型实例从属关系 AR-2
	整体部分同源从属关系 AR-3
	整体部分异源从属关系 AR-4
时序相关关系 RT	时序同现关系 RT-1
	时序先于关系 RT-2
	时序后于关系 RT-3
	时序时间长短关系 RT-4
空间相关关系 RC	空间毗连方位关系 RC-1
	空间上下方位关系 RC-2
	空间左右方位关系 RC-3
	空间包围相关关系 RC-4
	空间横贯相关关系 RC-5
	空间存现相关关系 RC-6
⋮	⋮

3 DomulAttCNN模型

一直以来,实体是关联关系提取赖以存在的基础,但因各领域实体间存在语义差异性难以获取,例如实体(X)“根”,在生命生物工程领域知识服务产业表现为“铝诱导植物的根分泌有机酸阴离子的机理”,在高端制造领域知识服务产业表现为“四元数系数多项式$Q(t)$的根的集合”,在新能源汽车领域知识服务产业表现为“利用根的轨迹法对汽车操纵稳定性的影响因素进行了研究”.

本文特提出一种多领域具有多层感知注意力机制的卷积神经网络模型(domain multi-layer attention convolutional neural network, DomulAttCNN),整体结构如图1所示. 首先,DomulAttCNN模型自动分类解构并聚合各领域专业知识,智能抽取结构化专业知识资源中的实体与关联关系部分; 其次, 针对以自然语

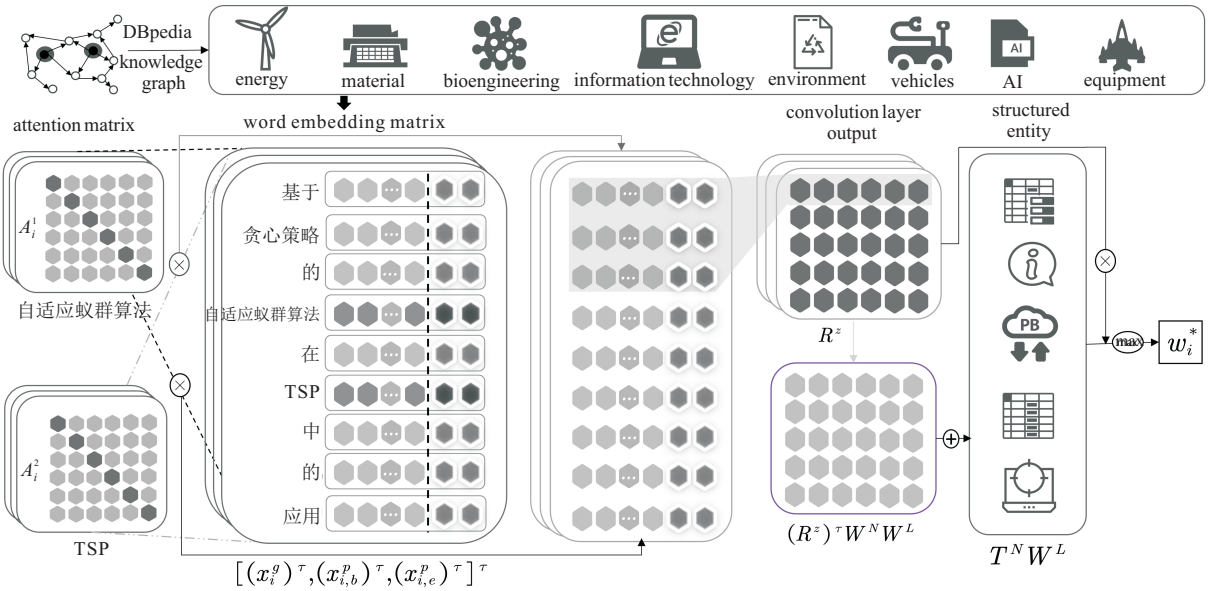

图 1　DomulAttCNN模型原理图

言形式存在的非结构化数据使用单词矢量表示对输入句子进行重编码,实体对采用句内上下文响应机制和位置编码机制高效地捕获实体词序,并利用基于对角矩阵的注意力机制来捕获句内单词相对于目标实体的相关性影响;然后,采用卷积层捕获句间关联信息,再经过有句间注意力机制的最大池层得到输出矩阵. 表 2 主要针对 DomulAttCNN 模型中的符号进行概述.

表 2　模型主要符号

符号	定义
x_i^M	词嵌入矩阵
c_i	上下文的嵌入矩阵
W_w	卷积核权重矩阵
b_c	卷积核偏置项
$\boldsymbol{G}$	相关矩阵
w_i^*	预测网络富关联关系输出

3.1　自然语言的输入预处理

首先结合领域专业内容文献知识实体关联特征,整理出资源的发表者(人物),发表机构(机构),关键词(知识点),发表载体(刊物)等类型的实体及各自之间的关系. 图 2 是一张典型的知识服务业知识图谱,展示了实体与实体间的关系及其部分属性. 其次链接融合了维基百科数据的 DBpedia[9] 知识图谱(knowledge graph, KG),将实体结合上下位关系自动分类成 8 大战略新兴产业领域. 通过知识图谱的领域分类辅助分析,解决基础专业知识的领域划分问题.

采用分词与命名实体识别技术[20]将自然语言文本子句中词转换为实值向量以提供词汇语义特征,表

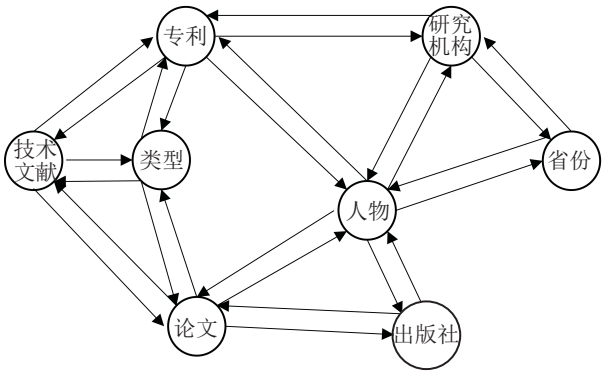

图 2　知识服务业知识图谱

示为 $T_n = (x_1, x_2, \ldots, x_n)$,并将全文档词嵌入矩阵设置为 x_v,其维数为 $g_x \times |v|$,v 为文档输入词,g_x 为词的矢量维度. 同时将每个 x_i 映射到列向量 $x_i^g \in \boldsymbol{R}^{g_x}$. 子句中两个标记的实体 $X_1(x_b)$ 和 $X_2(x_e)$ 中,$b, e \in [1, n], b \neq e$,如图2所示.

为融合上下文语义对实体对的关联关系影响,采用句内上下文响应机制和位置编码机制来高效地捕获实体词序. DomulAttCNN 模型将句内各词与头实体 x_b、尾实体 x_e 的相对距离嵌入词嵌入矩阵,第 i 个词的整体词嵌入矩阵为

$$x_i^M = [(x_i^g)^\tau, (x_{i,b}^p)^\tau, (x_{i,e}^p)^\tau]^\tau. \tag{1}$$

其中: x_i^g 表示每个词 x_i 映射后的列向量,$x_{i,b}^p$ 与 $x_{i,e}^p$ 分别表示第 i 个词相对于头实体 x_b、尾实体 x_e 的相对距离. 在图3所示子句“基于贪心策略的自适应蚁群算法在TSP中的应用”中,词“在”与实体对 X_1 =“自适应蚁群算法”和 X_2 =“TSP”的相对距离分别表示为 -1 和 1.

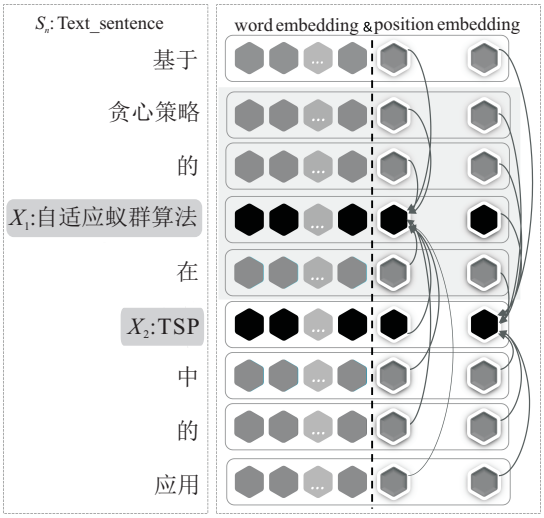

图 3 自然语言的输入预处理

当每次在子句中以第i个词为中心进行卷积时,为了充分利用上下文信息,使用大小为l的滑动窗口. 将l个词语的信息都融合到c_i中,最终表现形式为

$$c_i = [(x^M_{i-(\frac{l-1}{2})})^\tau, \dots, (x^M_{i+(\frac{l-1}{2})})^\tau]^\tau, \tag{2}$$

其中c_i表示上下文的嵌入矩阵,且$c_i \in \boldsymbol{R}^{(g_x+2g_p)l}$.

3.2 多层注意机制

因各领域专业知识存在多源异构实体对象,DomulAttCNN模型运用多层注意力机制可凸显各部分的重要性. 多层注意机制主要可以分为两类,既上下文注意力机制和卷积池化级注意力机制. 尤其在拥有多个子句的长句中,只有个别动词或名词与目标实体产生强影响. 此时采用注意力机制的词嵌入矩阵将形成句内重要表征性语义信息的凸显图,体现各词语元素与给定实体间联系的强弱程度,如图1所示.

3.2.1 上下文注意力机制

在图3所示的实例中,非实体词“应用”在决定关联关系时具有重要的因果关系意义. 因此,DomulAttCNN模型在共存语料库中引入两个实体对上下文相关的对角矩阵

$$A_i^{(1,2)} = \frac{\exp(f(X_{(1,2)}, x_i))}{\sum\limits_{i^*=1}^{n} \exp(f(X_{(1,2)^{i^*}}, x_{ij}))}. \tag{3}$$

具有注意力机制的对角矩阵$A_i^{(1,2)}$,分别表示实体与上下文相关词的联系强弱关系. $X_{(1,2)}$为句中两个实体,评分函数f则是$X_{(1,2)}$与x_i的内积.

在处理多个源的各个相关对角矩阵时,DomulAttCNN模型主要训练上下文嵌入矩阵c_i,并根据矩阵规模,分别采用两种处理方式进行联合矩阵融合,如下所示:

当矩阵规模较小($n \leqslant 15$)时,采用

$$r_i = [(c_i A_i^1)^\tau, (c_i A_i^2)^\tau]^\tau; \tag{4}$$

当矩阵规模较大($n > 15$)时,采用

$$r_i = c_i \frac{A_i^1 - A_i^2}{2}. \tag{5}$$

最终输出具有注意力的矩阵$R = [r_1, r_2, \dots, r_n]$,其中$n$是句子长度.

3.2.2 卷积池化级注意力机制

在浅层引入了上下文注意力机制输出矩阵R后,模型为提取更多抽象的高级特征,在更高层引入卷积池化级注意力机制代替传统的池化层. 此方式可以更好地确定在滑动窗口内各部分的重要性.

首先,将输出矩阵R送入卷积核大小为g_c的卷积层,引入权重矩阵W_w和偏置项b_c,即

$$R^z = \tanh(W_w R + b_c), \tag{6}$$

其中$W_w \in \boldsymbol{R}^{g_c \times l(g_x+2g_p)}$.

其次,引入注意力机制,构建一个关联矩阵S. 已忽略了句间无关成分,捕获上下文窗口之间相关连接的强关注部分,有

$$G = (R^z)^\tau W^N W^L + T^N W^L. \tag{7}$$

其中:W^N为窗口间的权重矩阵,T^N为结构化实体矩阵,W^L为关联关系权重矩阵.

最后,通过softmax函数得到富关联关系预测权重,再经过最大化输出关联关系,即

$$w_i^* = \max_j \Big(R^z_{i,j} \frac{\exp(G_{i,j})}{\sum\limits_{i^*=1}^{n} \exp(G_{i^*,j})} \Big). \tag{8}$$

3.3 边缘距离目标函数

DomulAttCNN模型采用基于富关联距离的损失函数,通过使该目标函数最小化来训练网络中的各参数. 在整个DomulAttCNN网络模型中,针对输入句T_n映射在实体关系空间之中,并由网络输出w_i^*. 同时,将W_φ^L映射为每一类实体关系的嵌入φ向量,且$\varphi \in \Upsilon$. 富关联距离采用欧氏距离函数为该向量差的L2范数进行描述,即

$$\rho_\theta(T_n, \varphi) = \Big\| \frac{w_i^*}{|w_i^*|} - W_\varphi^L \Big\|, \tag{9}$$

则损失函数为

$$L_{\text{loss}} = [\rho_\theta(T_n, \varphi) + (1 - \rho_\theta(T_n, \varphi^-))] + \xi\|\theta\|^2. \tag{10}$$

其中:$\rho_\theta(T_n, \varphi)$表示预测的输出与标注关系(ground truth)之间的距离,$\rho_\theta(T_n, \varphi^-)$为预测的输出与所有错误类别中得分最高的那个类别之间的距离,ξ为超参数,$\xi\|\theta\|^2$为正则项.

训练目标是让$\rho_\theta(T_n,\varphi)$尽可能的小,$\rho_\theta(T_n,\varphi^-)$尽可能的大. 因此,DomulAttCNN模型采用随机梯度下降(SGD)来更新损失函数的参数,即

$$\begin{cases} G=\sum_{i=1}^{n}[\rho_\theta(T_i,\varphi)+(1-\rho_\theta(T_i,\varphi^-))], \\ \theta^*=\theta+\lambda\dfrac{\mathrm{d}G}{\mathrm{d}\theta}+\varphi\dfrac{\mathrm{d}(\xi\|\theta\|^2)}{\mathrm{d}\theta}, \end{cases} \tag{11}$$

其中λ和φ是学习率.

4 实验结果与分析

4.1 数据集

为了更好地验证DomulAttCNN模型的效果,本文分别在Semeval-2010 (Task 8)数据集和领域专业内容文献(DPC 30)数据集上对模型进行了评估. 其中,Semeval-2010 (Task 8)数据集重点验证在9种基线关联关系类型领域的识别效果,而专业内容文献(PC 30)数据集侧重于验证领域专业内容富关联关系的识别效果.

4.1.1 Semeval-2010 (Task 8)数据集

Semeval-2010 (Task 8)数据集[3]内含10 717个注释句,分为8 000个训练句和2 717个测试句. 注释句主要标记9种基础关联关系类型和一个附加“other”类型. 9种类型分别是: cause-effect, component-whole, content-container, entity-destination, instrument-agency, entity-origin, member-collection, product-producer, message-topic. 由于数据集刻画了关系方向,当两个实体以相反的顺序出现时,其需要被视为表达不同的关系,例如cause-effect(X_1,X_2)和cause-effect(X_2,X_1). 同时,DomulAttCNN模型采用官方评估指标准确率、召回率、F_1分数来测试模型性能.

4.1.2 领域专业内容文献(DPC 30)数据集

领域专业内容文献(DPC 30)数据集跨越8大战略新兴行业(包含新能源、新材料、生命生物工程、信息技术和移动互联网、节能环保、新能源汽车、人工智能、高端装备制造)的领域专业内容资源库. 内涵30个学科的专业内容文献,120万条专业内容知识,700万条资源端口,数据总量已达2亿条. 领域专业内容文献的关联关系注释句由120位签约专家进行标记,其中共包含87 429条注释句,分为65 000个训练句和22 429个测试句,如表3所示.

表 3　专业内容文献(DPC 30)数据集

8大战略新兴领域	注释句/条
新能源	9 420
新材料	9 235
生命生物工程	17 007
信息技术和移动互联网	18 043
节能环保	6 437
新能源汽车	7 930
人工智能	11 036
高端装备制造	8 321
共计	87 429

4.2 实验环境与超参数设定

本文模型的实现基于Python 3.6,采用深度学习框架PyTorch. 在每次实验中,模型的所有网络最多运行300个epochs, dropout设置为0.2. 其中200维中英文词嵌入矩阵,选自经过百度百科与维基百科预先训练后的word2vec. 模型首先将全部超参数随机初始化,再通过交叉验证进行调整. 最优超参数如下: g_p(词相对位置维度)= 25; g_c(卷积核大小)= 1 000; l(窗口大小)= 3; λ(学习率)= 0.01; φ(学习率)= 0.000 1.

4.3 实验结果

本文采用传统的F_1分数方法(又称平衡F分数[21])提取富关联关系的评价指标. 通过这种方法可准确评价模型性能,如下所示:

$$F_1=2\times\frac{F_{\mathrm{precision}}\times F_{\mathrm{recall}}}{F_{\mathrm{precision}}+F_{\mathrm{recall}}}\times 100\,\%. \tag{12}$$

4.3.1 Semeval-2010数据集的实验结果

为评价DomulAttCNN模型基础关联的性能,首先在Semeval-2010 (Task 8)数据集上与5类9种基线模型进行传统关联关系提取效果对比实验,结果如表4所示.

由实验结果可以发现,DomulAttCNN模型性能更优. 同时,观察到传统基于人工特征的SVM模型[4]引入大量人工特征,获得了较高准确度,其F_1值为82.2. 由于该类特征人为干扰性强,极易产生不可逆的偏差导致推广性不强,现已被机选特征所代替. 基于依存树的MVRNN模型[10]和SDP-LSTM模型[8],主要通过寻找父实体节点与子实体节点间最短依赖路径来实现. 然而,此类方法极易忽略语义强相关实体间关联关系,以及偏差性链接语法无关联的两实体间关系,进而影响模型准确度. 同时,此类模型另一大缺点是计算复杂度会随着句子维度增加而剧增. 2015年,因深度学习方法的发展而产生的端到端模型,将传统基于支持向量机的模型准确度改善近1.9 %(对比CR-CNN模型)、3.4 %(对比depLCNN+NS模型). 由此可见,通过端到端的模型可有效改善网络结构,探查句内语义联系进而减少人为干预. 然而,由实验结果可以发现,深层网络结构存在表现力达到极限的情况,其性能上还需增强. 另一类远程监督模型,主要基于离散特征的Mintz逻辑回归分类模型与基

表 4　5类9种富关联关系提取模型的效果对比

模型细化类别	分类器	特征集合	F_1
人工特征模型	SVM[4]	+POS, WordNet, prefixes and other morphological features, dependency parse, Levin classes, PropBank, FanmeNet, NomLex-Plus, Google n-gram, paraphrases, TextRunner	82.2
依存树模型	MVRNN [10]	+POS, NER, WordNet	82.4
	SDP-LSTM [8]	+POS embeddings, WordNet embeddings, grammar relation embeddings	83.7
端到端模型	CR-CNN [7]	+Word embedding, POS embeddings	84.1
	depLCNN+NS [11]	+WordNet, words around nominals	85.6
远程监督模型	Mintz [12]	+KB, POS, dependency parse	67.6
	PCNN [13]	+KB, Word embedding, POS embeddings	75.0
注意力机制模型	PCNN+ATT [14]	+Attention, Word embedding, POS embeddings	80.7
	SEE-TRANS [15]	+Attention, Entity Embedding	85.0
本文模型	DomulAttCNN	+Attention, Position features, External lexical resources(KG)	**89.1**

于卷积神经网络的PCNN模型. 虽然此类模型在F_1值方面表现欠佳,但其成功通过知识库(KB)对齐文本,激增大规模语料而功不可没. 最后一类,注意力机制模型在富关联关系提取上获得了良好的结果,著名的有PCNN+ATT和SEE-TRANS. 然而,融合了各类基线模型优势的DomulAttCNN模型F_1值首次逼近90%大关,性能始终优于其他5类基线模型. 对比Mintz模型提高21.5%,同时验证了利用边缘距离的目标函数在富关联关系提取的效果方面表现更优.

为进一步分析模型边缘距离目标函数对模型的影响效应,与传统目标函数(Softmax交叉熵)的收敛性行为进行对比,结果如图4所示. 由图4可知,在多轮迭代的模型训练过程中,DomulAttCNN模型可以更为平滑地收敛到最终F_1值. 然而,传统模型易产生层间联合效应,会引起更强的反向传播效应,从而出现收敛震荡现象.

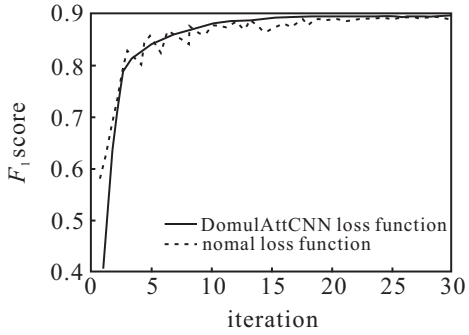

图 4　边缘距离目标函数与传统目标函数对比

4.3.2　DPC 30数据集的实验结果

为验证模型的泛化性,本文挑选在Semeval-2010 (Task 8)数据集上表现效果较优的6个模型进行领域专业文献富关联关系提取. 由于各领域存在差异性,传统的5种模型未经过领域训练,不同程度上存在领域混淆的情况. 而本文模型很大程度上为此类问题提供了解决方案,由于采用各领域知识图谱刻画异构环境中领域知识链路,使得本文的模型可以根据各领域特点挖掘其浅层语义关联关系,并采用分领域多层的注意力机制捕获特定领域实体与特定关系池的注意力. 尽管各领域的专业内容结构存在异构性,本模型可根据各领域特点刻画注意力响应机制. 由经过优化后成对的基于边缘距离的目标函数中可以看出,模型的收敛效率优于其他模型. 由表5的结果可以看出,DomulAttCNN模型在大多数领域下的富关联关系提取是非常有效的,在少数分词较困难领域会出现注意力机制效果不明显的情况.

通过两个不同数据集中的实验结果,本文分析得出:

1) 以传统特征为主的模型构建过程中,模型性能会随特征模块的丰富化而变优. 然而,此类特征模型的质量受限于人为标注特征的独创性或文本知识的先验性,而难以推广. 本文基于知识图谱预处理的DomulAttCNN模型很好地解决了这一难题. 专业知识服务中,基于新型远程监督的知识图谱技术成功刻画了关系链路,为知识、信息、资源、服务、实体对象间的富关联关系模式打下链接基础.

2) 具有多层感知的上下文注意力机制可以有效提取抽象特征. 同时,基于卷积池化级的注意力机制可以增加实体链接稠密性,有效融合各领域特征. 尽管输入的专业内容资源结构存在极强的异构性,本文模型仍然可以洞察更为细微的富关联线索.

3) 通过构建基于边缘距离的目标函数,可以增强

表5　6种富关联关系提取模型在各领域的F_1分值

8大战略新兴领域	F_1分值					
	SVM [4]	SDP-LSTM [8]	depLCNN+NS [11]	PCNN [13]	SEE-TRANS [15]	DomulAttCNN
新能源	68.97	52.73	68.48	61.35	71.46	**74.21**
新材料	64.69	58.26	**70.28**	64.95	62.22	68.88
生命生物工程	54.58	56.41	52.13	50.10	62.05	**66.30**
信息技术和移动互联网	64.36	72.23	71.39	65.10	57.63	**77.16**
节能环保	68.72	50.22	57.35	57.90	60.95	**72.74**
新能源汽车	65.02	58.42	**74.56**	65.63	65.96	74.05
人工智能	52.20	56.00	52.22	61.43	51.34	**69.74**
高端装备制造	60.75	68.30	72.93	63.23	66.64	**76.63**

模型的可解释性与有效性,且收敛效率明显优于标准损失函数. 通过缩小预测输出与标注关系间距离,可拉大与错误类之间的距离,进而逼近目标关系,有效提高模型训练效率.

5 结　论

本文提出了一种新型的领域专业知识富关联关系提取方法. 通过DomulAttCNN模型的构建过程,解决了传统领域知识富关联关系提取较为困难的现状. 本文制定了面向服务专业内容资源的一致性富关联关系的描述体系,运用多层注意力机制来凸显实体间重要表征性信息,并通过优化基于边距离的目标函数提升模型训练效率. 实验结果表明,这种简单而有效的端到端模型,可有效提升专业知识服务中富关联关系识别精度. 在未来的工作中,将进一步探索并优化网络结构,利用更准确的分词方法进行提升语义特征表现能力,从而提高在分词较困难的领域模型准确度.

参考文献(References)

[1] Qian L H, Zhou G D, Kong F, et al. Semi-supervised learning for semantic relation classification using stratified sampling strategy[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2009: 1437-1445.

[2] Chen J, Tandon N, de Melo G. Neural word representations from large-scale commonsense knowledge[C]. 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Singapore: IEEE, 2015, 1: 225-228.

[3] Hendrickx I, Kim S N, Kozareva Z, et al. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals[C]. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Morristown: Association for Computational Linguistics, 2009, 6: 94-99.

[4] Rink B, Harabagiu S. UTD: Classifying semantic relations by combining lexical and semantic resources[C]. Proceedings of the 5th International Workshop on Semantic Evaluation. Morristown: Association for Computational Linguistics, 2010, 4: 256-259.

[5] Tratz S, Hovy E. Isi: Automatic classification of relations between nominals using a maximum entropy classifier[C]. Proceedings of the 5th International Workshop on Semantic Evaluation. Morristown: Association for Computational Linguistics, 2010, 4: 222-225.

[6] 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. Dublin: Association for Computational Linguistics, 2014: 2335-2344.

[7] Santos C N D, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[J]. Computer Science, 2015, 86: 132-137.

[8] 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. Lisbon: Association for Computational Linguistics, 2015: 1785-1794.

[9] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data[M]. The Semantic Web, Berlin, Heidelberg: Springer, 2007: 722-735.

[10] 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. Jeju Island: Association for Computational Linguistics, 2012: 1201-1211.

[11] Xu K, Feng Y, Huang S, et al. Semantic relation

classification via convolutional neural networks with simple negative sampling[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 536-540.

[12] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec: Association for Computational Linguistics, 2009: 1003-1011.

[13] 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. Lisbon: Association for Computational Linguistics, 2015: 1753-1762.

[14] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 2124-2133.

[15] He Z, Chen W, Li Z, et al. SEE: Syntax-aware entity embedding for neural relation extraction[C]. The 30th Innovative Applications of Artificial Intelligence. New Orleans: AAAI, 2018: 5795-5802.

[16] Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification[C]. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai, 2015: 73-78.

[17] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2010: 148-163.

[18] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011: 541-550.

[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. Jeju Island: Association for Computational Linguistics, 2012: 455-465.

[20] Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling[C]. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018: 1638-1649.

[21] He X, Cai D, Niyogi P. Laplacian score for feature selection[C]. Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2005, 8: 507-514.

作者简介

李青 (1989–) 女, 博士, 从事自然语言处理、复杂事件检测的研究, E-mail: liqing@cqu.edu.cn;

钟将 (1974-), 男, 教授, 博士生导师, 从事自然语言处理、数据挖掘等研究, E-mail: zhongjiang@cqu.edu.cn;

李立力 (1989–), 男, 博士, 从事桥梁健康监测、数据挖掘的研究, E-mail: lilili@cqu.edu.cn;

张剑 (1977–), 男, 工程师, 硕士, 从事自然语言处理、数据挖掘的研究, E-mail: 13608341660@139.com;

李琪 (1987–), 男, 博士, 从事图计算、数据挖掘的研究, E-mail: liqi0713@foxmail.com.

(责任编辑：孙艺红)

著作权转让声明

论文作者须保证所投论文为原创作品且不存在涉密和一稿多投问题,若发生侵权或泄密问题,一切责任由论文作者承担. 论文作者保证所投论文的署名无争议,若发生署名争议,责任由论文作者承担.

本刊已许可中国知网以数字化方式复制、汇编、发行、信息网络传播本刊全文. 本刊支付的稿酬已包含中国知网著作权使用费. 所有署名作者向本刊提交文章发表之行为视为同意上述声明. 如有异议,请在投稿时说明.