doi:10.3969/j.issn.1001-4616.2021.01.018

基于多特征双向门控神经网络的领域专家实体抽取方法

张柯文,李　翔,严云洋,朱全银,马甲林

(淮阴工学院计算机与软件工程学院,江苏 淮安 223005)

[摘要]　命名实体识别是自然语言处理和信息提取的基本任务,传统专家命名实体识别方法存在过度依赖人工特征标注和分词效果、专家简介中大量专业新词无法识别等问题. 本文提出一种基于多特征双向门控神经网络结构并结合条件随机场模型进行领域专家实体抽取方法. 该方法首先通过构建领域专家语料库以训练实体抽取模型;接着,使用 Bert 方法进行字嵌入表示,对语料库专业领域词汇构造要素进行特征分析并提取边界特征;然后,利用双向门控神经网络和注意力机制有效获取特定词语长距离依赖关系;最后,结合条件随机场模型实现命名实体识别. 在同一数据集上进行 5 种方法实验比较分析,结果表明该模型较 BiLSTM-CRF 和 IDCNN-CRF 方法 $F1$ 值提高 9.98%以上.

[关键词]　命名实体识别,自然语言处理,信息提取,多特征,边界特征

[中图分类号]TP301.6　[文献标志码]A　[文章编号]1001-4616(2021)01-0128-08

Domain Expert Entity Extraction Method Based on Multi-Feature Bidirectional Gated Neural Network

Zhang Kewen,Li Xiang,Yan Yunyang,Zhu Quanyin,Ma Jialin

(Faculty of Computer and Software Engineering,Huaiyin Institute of Technology,Huai'an 223005,China)

Abstract:Named entity recognition is the basic task of natural language processing(NLP) and information extraction(IE). Traditional expert named entity recognition methods have problems,such as excessive reliance on artificial feature labeling and word segmentation effects,and the inability to recognize a large number of professional new words in the expert profile. This paper proposes a method based on multi-features bidirectional gated neural network structure combined with conditional random field model for the domain expert entity extraction. Firstly,train the entity extraction model by constructing a domain expert corpus. Secondly,use the Bert method to represent the word embedding,and perform feature analysis on the vocabulary structure elements of the professional field of the corpus and extract the boundary features. Thirdly,use the bidirectional gated neural network and attention mechanism to effectively obtain the long-distance dependence of specific words. Finally,combine the conditional random field model to achieve named entity recognition. The experimental comparison and analysis of five methods on the same data set show that the $F1$ value of the model is improved by more than 9.98% compared with BiLSTM-CRF and IDCNN-CRF.

Key words:named entity recognition,natural language processing,information extraction,multi-feature,boundary feature

信息抽取(information extraction,IE)的主要研究方法是基于自然语言处理和文本挖掘,从非结构化或半结构化的网络文本数据中挖掘出有价值的信息[1]. 命名实体识别(named-entity recognition,NER)是自然语言处理和信息抽取的基础任务,从文本中识别命名性指称项,为关系抽取、机器翻译和自动文摘等任务做铺垫[2].

收稿日期:2020-08-08.

基金项目:国家自然科学基金项目(71874067、61602202)、国家重点研发计划项目(2018YFB1004904)、江苏省产学研合作项目(BY2020067、BY2020309)、江苏省农业科技自主创新资金项目(CX203074)、淮阴工学院研究生科技创新计划项目(HGYK202024).

通讯作者:李翔,博士,副教授,研究方向:机器学习、数据挖掘、推荐系统. E-mail:hyitlixiang@ hotmail.com

专家信息是一种以网络文本形式存在的非结构化数据,是专家向社会展示个人基本信息和过去经历的重要载体[3]. 通过大数据技术对专家信息进行整理、分类和分析后,以不同的形式为政府、高校、企业提供精准的专家信息服务,可以构建高校科技人才与政府、企业的联通桥梁. 然而,随着互联网技术的普及,大量的电子文本信息在筛选过程中需要耗费大量的时间及精力[4]. 实体抽取的研究更好地满足人们信息检索的需求. 通过从非结构化文本中提取指定类型的关键性信息,自动转换为结构化信息以支持数据库的保存及数据的下一步处理[5]. 在实体抽取的研究中,Zhang 等[6]主要是关注人名、地名和组织机构名这三类名词的识别. 对于处理专家简历信息而言,除人名、机构名外,专家的其他信息(包括职称、研究领域名称、电子邮件地址及电话号码)的提取同样起着基础性作用,而特定领域的专有名词是非常重要的实体. 对研究领域名的识别研究还很薄弱,一方面,研究领域很大程度上与行业知识息息相关,另一方面领域的特殊性给实体的抽取带来了挑战. 因此,为更好地对复杂文本进行处理,将自然语言处理与行业知识深度融合受到了更多的关注.

本文首先对领域专家实体抽取及相关问题进行介绍;然后阐述基于多特征双向门控神经网络的构建过程及命名实体识别抽取专家信息过程;最后,以化工专家网络文本作为实验数据,使用 HMM、IDCNN-CRF、BiLSTM-CRF 及多特征双向门控神经网络抽取方法进行化工专家实体抽取,根据实验结果分析本文提出的模型的优势及未来工作.

1 相关工作

实体抽取方法可分为传统实体抽取方法、基于机器学习的抽取方法和基于神经网络的抽取方法. 传统实体抽取方法都是基于词典和规则的,通过大规模语料库构建词典,在实体抽取的识别准确率和召回率上取得了很大的提升[7]. 面向专家领域的规则还需要领域专业人士去构建,此类方法在抽取专家实体过程中不仅受限于词典的规模和质量,还无法识别和抽取新的实体. 基于机器学习的抽取方法在预测性上可以预测新的实体,逐渐受到研究者们的广泛关注. Morwal[8]引入马尔科夫假设的隐马尔可夫模型(hidden Markov model,HMM)算法非常适合用于序列标注问题,但其局限于输出独立性假设,在实际文本中限制了上下文特征的选择. McCallum 等[9]提出的最大熵隐马模型(maximum entropy Markov model,MEMM)使用局部最优值解决了隐马的问题,同时也带来了标记偏见的问题. Lafferty 等[10]于 2001 年提出的条件随机场(conditional random field,CRF),结合了最大熵模型和隐马尔可夫模型的特点,通过监督学习可更加高效地进行实体识别任务,还可以准确地预测新的实体.

为减少特征工程的需求,深度学习方法给实体抽取方法提供了新的思路. 神经网络出色的非线性映射和自主学习的能力在很大程度上减少了特征工程的工作量. 2018 年 Google 发布的基于双向 Transformer 的大规模预训练语言模型(Bi-directional encoder representation from transformers,Bert)[11]在处理命名实体识别等序列标注任务中取得了很好的效果. Collobert 等[12]最早提出用 CNN 对序列标注任务来自动提取特征的模型. Strubell 等[13]提出使用 Iterated Dilated CNN+CRF 模型进行命名实体识别,取得了很好的效果. Huang 等[14]提出目前中文序列标注最常用的模型 BiLSTM-CRF,充分利用上下文特征,在实体抽取任务上取得了很高的成就. 深度学习模型在行业领域研究和应用中还处于起步阶段. 在实际研究中,基于神经网络的实体抽取任务多以英文语料为主,在中文文本的应用中效果差强人意.

本文针对中文专家信息的特点,以领域专业术语在文档中的特征进行分析,提出基于多特征双向门控神经网络的领域专家简介实体抽取的方法. 首先,挖掘网络文本并对其清洗及规范化,半自动标注构建领域专家简介语料库;接着,对语料库专业领域专业名词构造要素进行分析,使用 Bert 语言模型进行字嵌入表示;然后,将处理后的有监督文本向量输入双向门控神经网络,利用注意力机制有效获取特定词语长距离依赖关系;最后,结合边界特征构建条件随机场模型实现命名实体识别. 门控神经网络可以从上下文中自动找到更有用的单词以获得更好的 NER 性能,从而解决人工特征提取成本高和专业新词无法识别等问题.

2 问题描述

2.1 领域专家实体定义

领域专家实体抽取是进行专家信息抽取的首要工作,即从专家网络文本中识别并提取具有实际意义

的实体,从而表示专家信息. 专家信息中的领域术语能够快捷准确地了解专家的研究领域及研究方向,有效抽取并利用领域专家实体能够更好地检索或推荐专家信息. 因此,本文以化工领域的专家网络文本为例抽取实体,基于多特征和双向门控神经网络构建自动抽取模型.

Zhang 等[6]从新浪财经收集简历数据,将个人简历分为包括国家(country)、机构(educational institution)、所在地(location)、人名(personal name)、组织(organization)、行业(profession)、种族背景(ethnicity background)及职位(job title)8 种实体,使用门控循环单元使模型从句子中选择最相关的字符和词,以生成更好的 NER 结果,而与行业领域方向相关的实体没有涉及. 本文分析专家网络文本发现化工领域术语存在以下特点:(1)中文行业领域术语实体歧义多变,且随时间推移不断出现新词,在抽取过程中新词识别无法掌控;(2)化工领域术语组合模式复杂,其中包含字长及中英文混杂的特点,如 TAME 原料预处理、DNW 高温树脂合成异丙醚研究等;(3)领域术语多为嵌套或复合结构,如污染物防控及资源化利用、功能材料的合成及制备工艺等.

综上,本文将领域专家实体定义为 3 类,如表 1 所示. 第一类为普通名词性实体,包括人名、机构名及职称;第二类为数字性实体,包括联系方式和电子邮件;第三类为领域性实体,包括研究方向及领域关键词实体.

表 1　化工领域专家实体描述

Table 1　Entity description of experts in chemical industry

实体类别		实体描述	示例	类型
普通名词性实体	人名	姓和名组成,多为 2~3 个字,少数民族字数较多	周志华	Name
	机构名	以“学院”或“大学”为特征	南京大学/淮阴工学院	ORG
	职称	以等级区分专家职称	教授/副教授	PRO
数字性实体	联系方式	手机号多为 11 位数字	(13|14|15|17|18|19)[0-9]{9}	PHO
	电子邮件	英文数字,以@ 为标识	\w[-\w.+]*@([A-Za-z0-9][-A-Za-z0-9]+\.)+(cn|com|edu|net)	EMA
领域性实体	研究方向	以中文字为主,且多具有复合词特征	人工智能/一般化学工业	FIE
	关键词实体	以字长不等、中英文混杂和复合结构为特征	TAME 原料预处理/DNW 高温树脂合成异丙醚研究	KEY

2.2　实体抽取目标

本研究的最终目标是从领域专家网络文本中提取定义的专家实体类型,重点解决领域性实体抽取过程中存在的领域实体无法识别及现有方法对人工特征过度依赖的问题. 本文从多特征角度对 3 类实体进行分析,提取相关性特征. 使用 Bert 语言模型以字符为单位进行文本向量化表示,统计特定词汇上下文边界信息;使用双向门控神经网络获取长文本上下文信息;训练条件随机场模型处理有强依赖性数据的难题,从而对文本实现更好的标注. 输出结果为:c = {“content text”offset“content type”},其中,c 表示输出内容;content text 表示专家实体内容;offset 表示实体起始到结束的标识;content type 则表示定义的专家实体类型.

3　领域专家简介实体抽取模型

领域专家简介实体抽取过程以化工领域为例如图 1 所示,首先对化工专家网络文本进行预处理,包括分词、词性标注及特征抽取等;然后,将化工专家实体抽取转化为序列标注问题,将抽取的特征通过多特征双向门控神经网络提取隐藏层特征;最后将其输入到条件随机场模型对上下文标注进一步约束,得到序列标注结果,实现化工专家实体的识别和抽取.

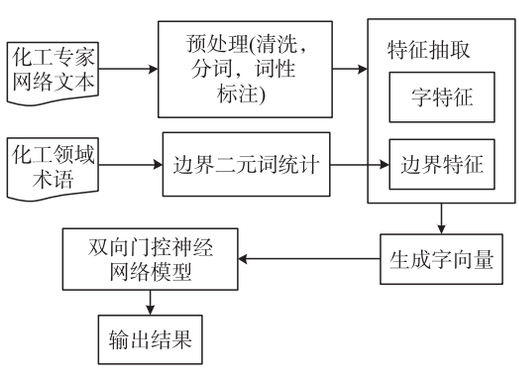

图 1　领域专家简介实体抽取过程

Fig. 1　Process of domain expert introduction entity extraction

3.1　语料库构建

3.1.1　数据清洗及规范化

通过数据源搜索的数据一定要经过清洗,才能让数据发挥价值,最终保证数据分析结果的准确性. 对于爬取的领域专家网络文本,通过预定义

的清理规则,将脏数据转化为满足数据质量要求的数据,使数据变得完整和精准,从而保证后续数据分析结果的准确性[15]. 数据清洗方法包括:(1)筛查文本数据的一致性,根据数据源内部及数据元之间的规范,将文本转换为统一结构的规范化数据;(2)检测并清除特殊字符,使用规则匹配去除 JavaScript 代码及编号等无效字符;(3)检测重复文本,基于时间节点保留最新数据,保证数据的唯一性.

3.1.2 半自动标注构建领域专家语料库

张华平等[16]认为中文分词是中文自然语言处理的基础. 在中文自然语言处理中,词是最小的能够活动独立的有意义的语言成分,因此进行中文自然语言处理通常是先将中文文本中的字符串切分成合理的词语序列,然后在此基础上进行其他分析处理.

对于中文的分词规范取决于不同的应用,在领域专家文本中,本文使用半自动标注构建领域专家语料库. 首先,在基本分词步骤中引入专家姓名、机构名称及职称为基础词汇表,以保证分词结果的准确性;接着,对于数字性实体,使用正则表达式对邮箱及电话进行规则匹配并标记;然后,通过专业领域关键词,对研究方向及领域关键词实体进行匹配标记,其中嵌套或复合结构的领域术语不做细粒度拆分;最后,对分词后的结果进行人工检验,对于未标注实体,使用 YEDDA 工具进行人工补充.

3.2 特征抽取

数据预处理是自然语言处理的基础任务,处理的质量决定了模型实现的质量. 中文文本不同于英文文本,无法以空格进行划分,通常以词为单位. 分词结果的好坏同样影响着模型对实体抽取的性能. 本文根据所抽取的实体类别,引入字嵌入特征和边界特征进行分析.

3.2.1 字嵌入特征

唐明等[17]利用词嵌入方法生成文档向量,通过单词在连续的低维空间中表示,捕获单词间的语义联系,在处理文档分类上取得了很好的效果. Mikolov 等[18]提出的 Word2Vec 和 Pennington 等[19]提出的 GloVe 在词嵌入上取得了很大的成功. 然而,对于中文语言没有明显词边界的特征,分词结果的好坏对语言处理的结果有很大的影响. 在专家网络文本中,除中文字符外,还包括标点符号、数字和英文字母,在处理词嵌入过程中给分词结果带来挑战. 因此,本文以字嵌入的方法对文本进行向量化表示,即每个汉字训练一个字嵌入. 根据训练集提取,在语言模型训练后生成一个大小为 $|\boldsymbol{C}|$ 的字典 $\boldsymbol{C}$,而未知字符也可以作为一个特殊的符号添加到字典中. 对于每个字 $\boldsymbol{c}$ 都可以映射为一个字向量 $\boldsymbol{v}_c \in R^d$,d 为向量维数,生成的字向量加入到字嵌入矩阵 $\boldsymbol{M} \in R^{d\times|\boldsymbol{c}|}$ 中. 本文通过对文本预处理,将文本以字表示,引入 Bert 语言模型生成字向量,作为实体抽取模型的输入.

Bert 模型采用 Transfomer 的编码器作为主体模型结构,舍弃了 RNN 循环式网络结构,引入了双向的语言模型任务如图 2 所示,完全基于注意力机制对文本进行建模. 通过注意力机制计算文本中每个词和所有词之间的相互关系,根据相互关系反应不同词之间的关联性及重要程度. 以词与词之间的权重获得每个词新的表征,通过自身及与其他词之间的关系得到全局性的表示. Transfomer 则对输入的文本不断进行注意力机制层和非线性网络层的交叠得到最终文本的表达. 将 Bert 模型引入实体抽取任务,不仅考虑到上下文信息,还充分利用了全局信息,在进行实体消歧上有很大的优势,在处理相似的未登录字符上更容易被识别,提高了实体抽取模型的召回率.

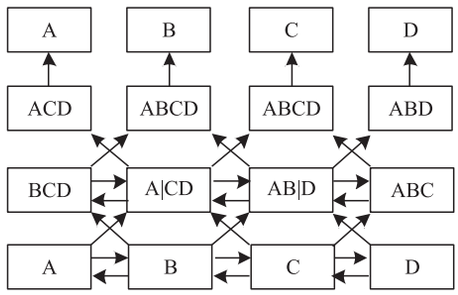

图 2 双向语言模型任务

Fig. 2 Task of bidirectional language model

3.2.2 边界特征

中文名词的表述上一般具有边界模糊的问题,即与名词相邻的词语具有很强的边界性. 传统基于词典和规则的方法可通过定义边界规则来区分名词信息,如联系方式与电子邮箱等具有明显的边界表示. 而在定义行业领域专业词汇上进行序列标注任务时,其组合模式多变、字长不固定及中英文混杂等特点使其在边界定义模糊. 本文以化工技术行业中英文关键词为标准,分词过程中基于关键词对嵌套或复合结构的领域术语实体不做细分,以减少对此类实体提取产生的影响. 在语料库中进行边界提取,提取结果如表 2 所示.

严云洋等[20]提出一种基于离群点检测的分类结果置信度的度量方法提高分类准确率. 本文将可信度作为边界的衡量标准,其定义如下所示:

$$P(\boldsymbol{c}_i)=\frac{\log(f_{\boldsymbol{c}_i}+1)}{\log(w_{\boldsymbol{c}_i}+1)}, \tag{1}$$

式中,$\boldsymbol{c}_i$ 代表语料库中的第 i 个字符,$f_{\boldsymbol{c}_i}$ 表示其作为边界的二元概率,$w_{\boldsymbol{c}_i}$ 表示 $\boldsymbol{c}_i$ 在未标注语料库中的共现频次. 通过可信度进行标准化得到离散化的特征数,作为边界特征输入到模型中.

表 2　边界提取统计表

Table 2　Statistics table of boundary extraction

前一个词	后一个词	共现频次	二元概率
的	影响	1 762	inf
进行	了	1 405	0.432 707
研究	了	1 039	0.547 996
中	的	857	inf
一	种	631	0.504 396
条件	下	614	0.368 326
性能	的	486	0.222 834
过程	中	436	0.337 461

3.3　BIGRU-CRF

循环神经网络(RNN)是一种能够有效解决序列标注问题及处理文本序列上下文依赖的神经网络模型. 而 RNN 无法很好地处理长距离依赖问题,在训练过程中存在梯度消失和梯度爆炸的问题. 基于这个问题,非线性激活函数长短期记忆(long short-term memory,LSTM)和门控循环单元(gated recurrent unit,GRU)被提出. LSTM 在神经元中加入输入门(input gate)、输出门(output gate)、忘记门(forget gate)及记忆单元(cell state)改善梯度消失的问题. GRU 作为 LSTM 的变体,将忘记门和输出门合并为一个更新门,结构更简单,训练时间更短,在训练结果上与 LSTM 取得相当的结果. 本文采用 GRU 学习文本的结构信息,其内部结构如图 3 所示,公式定义如下:

$$\boldsymbol{z}=\delta(W_z[\boldsymbol{x}^{\langle t\rangle},\boldsymbol{a}^{\langle t-1\rangle}]+b_z), \tag{2}$$

$$\boldsymbol{r}=\delta(W_r[\boldsymbol{x}^{\langle t\rangle},\boldsymbol{a}^{\langle t-1\rangle}]+b_r), \tag{3}$$

$$\hat{\boldsymbol{h}}^{\langle t\rangle}=\tanh(W_c[\boldsymbol{x}^{\langle t\rangle},\boldsymbol{r}\cdot\boldsymbol{a}^{\langle t-1\rangle}]+b_c), \tag{4}$$

$$\boldsymbol{h}^{\langle t\rangle}=\boldsymbol{z}\cdot\hat{\boldsymbol{h}}^{\langle t\rangle}+(1-\boldsymbol{z})\cdot\boldsymbol{h}^{\langle t-1\rangle}, \tag{5}$$

式中,$\boldsymbol{z}$,$\boldsymbol{r}$ 和 $\hat{\boldsymbol{h}}^{\langle t\rangle}$ 分别代表更新门、重置门和 t 时刻候选隐藏状态. W_z、W_r、W_c 和 b_z、b_r、b_c 分别为更新门、重置门和候选隐藏状态的权重和偏置参数. $\boldsymbol{x}^{\langle t\rangle}$、$\boldsymbol{a}^{\langle t-1\rangle}$ 分别为当前神经网络的输入和前一隐藏节点输出的激活值. $\cdot$ 表示 Hadamard 乘积. 重置门决定了如何将新的输入信息与前面的记忆相结合,更新门定义了前面记忆保存到当前时间步的量,候选隐藏状态为当前时刻的输入信息. 通过两个门控机制能够保存长期序列中的信息,且不会随时间而清除或因为与预测不相关而移除.

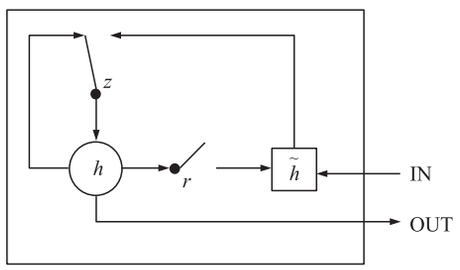

图 3　GRU 内部结构

Fig. 3　The internal structure of GRU

GRU 只能从一个方向获取序列信息,而对于基于上下文的文档表示序列,不能获取后文对语义之间的影响. 因此,本文使用 GRU 的扩展双向门控神经单元,分别从前向及后向建模获取文本的依赖关系,获得 t 时刻的隐藏状态 $\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t]$,通过前后向隐藏状态拼接同时获取上下文特征.

CRF 可以关注句子级别利用邻居标签信息,产生更高的标记精度. 给定一组随机变量 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,对应随机变量 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$ 满足 $P(\boldsymbol{Y}_v|\boldsymbol{X},\boldsymbol{Y}_w,w\neq v)=P(\boldsymbol{Y}_v|\boldsymbol{X},\boldsymbol{Y}_w,w\sim v)$ 的马尔科夫随机场为条件随机场,$(\boldsymbol{X},\boldsymbol{Y})$ 是条件随机字段,其中,$\boldsymbol{X}$ 表示观察到的序列,$w\sim v$ 表示与节点 v 相连的 w 的所有相邻节点. $\boldsymbol{Y}$ 的候选标签的联合概率分布可以在因子分解下表示为:

$$P(\boldsymbol{Y}|\boldsymbol{X})=\frac{1}{Z(\boldsymbol{x})}\exp\Big(\sum_{i,k}v_it_k(\boldsymbol{y}_{i-1},\boldsymbol{y}_i,\boldsymbol{x},i)+\sum_{i,l}u_is_l(\boldsymbol{y}_i,\boldsymbol{x},i)\Big), \tag{6}$$

$Z(\boldsymbol{x})$ 为归一化因子,可表示为:

$$Z(\boldsymbol{x})=\sum_{y}\exp\Big(\sum_{i,k}v_it_k(\boldsymbol{y}_{i-1},\boldsymbol{y}_i,\boldsymbol{x},i)+\sum_{i,l}u_is_l(\boldsymbol{y}_i,\boldsymbol{x},i)\Big), \tag{7}$$

式中,t_k 是状态转移函数,s_l 是发射函数,v_k 和 u_i 分别为 t_k 和 s_l 对应的权值.

3.4　Attention 机制

双向 GRU 获取的上下文信息无法完全融入当前字符信息. Attention 机制在不同时刻计算输出特征向量的权重,突出字符的重要特征.

$$\boldsymbol{score}=\boldsymbol{v}^{\mathrm{T}}\tanh(W_1\boldsymbol{h}_i+b_1), \tag{8}$$

$$\boldsymbol{\alpha}_i = \frac{\exp(\boldsymbol{score})}{\sum_{t=1}^{M}\exp(\boldsymbol{score})}, \tag{9}$$

$$\boldsymbol{c}_i = \sum_{j=1}^{T}\alpha_{ij}\boldsymbol{h}_j, \tag{10}$$

式中,$\boldsymbol{score}$ 为包含语义信息的 $\boldsymbol{h}_i$ 输入到单层感知机中获得单篇文档隐藏层的输出,计算出当前字符权重矩阵 $\boldsymbol{\alpha}_i$ 与文本特征向量 $\boldsymbol{h}_i$ 进行加权和,得到包含文档各字符重要性信息的向量 $\boldsymbol{c}_i$. 通过 Attention 机制控制当前字符权重,从而增加文档表示之间的语义联系,使整个模型获得更好的效果.

3.5 多特征双向门控神经网络的领域专家实体抽取

本文设计了一种在多特征选择的基础上,扩展基本字符单元,使用双向门控神经网络并添加注意力机制,CRF 对获取的信息再利用进行序列标注,抽取领域专家实体信息,抽取结构如图 4 所示.

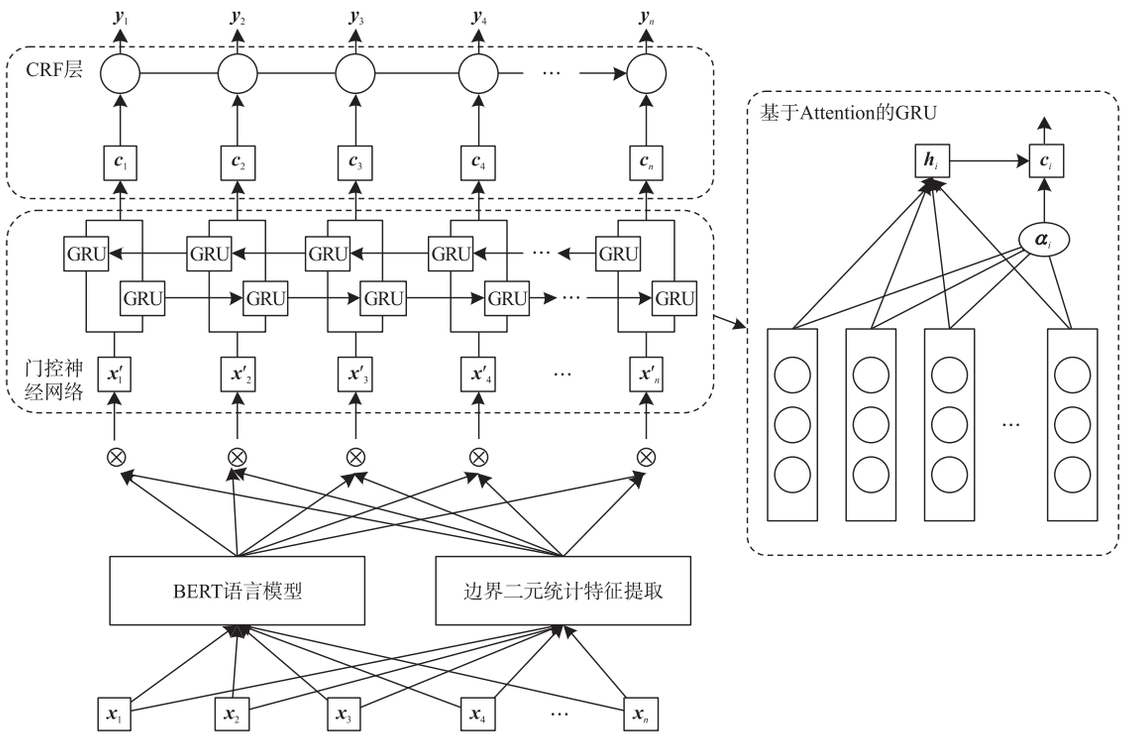

图 4 多特征双向门控神经网络实体抽取结构

Fig. 4 Multi-feature bidirectional gated neural network entity extraction structure

多特征双向门控神经网络实体抽取结构本质是通过 Bert 模型构建字向量特征 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 和边界向量表计算边界特征 $\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_n\}$,将其拼接后作为 BIGRU 网络层的输入,经过双向门控神经网络获取隐藏层输出 $\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$ 并输入注意力机制层,使用 CRF 进行序列标注建模,获取全局最优标签序列 $\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$,实现专家网络文本的实体抽取. 本模型充分利用文本字嵌入特征及领域术语实体的边界特征,为字符的向量化表示提供特征支持. BiGRU+Attention 更加有效地利用了字符的上下文表示,通过 Attention 机制分配字符在网络中的结构,有效利用特定词语在文本中的长距离依赖关系. CRF 进一步增强了前后标注的约束,避免了不合法的标注情况出现.

4 结果与讨论

4.1 实验数据

本文使用高校官网收集的专家网络文本作为实验数据,其中包含 25 053 篇化工专家文档,共 5 162 个汉字. 使用 1 089 条化工技术行业中英文关键词对化工领域术语进行边界特征提取. 将 25 053 篇化工专家文档以 7:3 的比例分为训练集和测试集,数据描述如表 3 所示. 训练集

表 3 实验数据描述

Table 3 Description of experimental data

数据集	文档数量/篇	句子数量/句
训练集	17 537	737 199
测试集	7 516	263 943

共包含化工专家文档 17 537 篇、737 199 句文本,测试集共包含化工专家文档 7 516 篇、263 943 句文本. 通过使用本文提出的多特征双向门控神经网络算法进行命名实体识别与传统 HMM、BiLSTM-CRF 和 IDCNN-CRF 进行比较,验证算法的优越性.

4.2　评价指标

为减少外在人为因素的影响,本文通过精确率 P、召回率 R 和 $F1$ 值来评价模型效果[21]. 精确率为被识别为该分类的正确分类记录数与被识别为该分类的记录数之比;召回率为被识别为该分类的正确分类记录数与测试集中该分类的记录总数之比,召回率是覆盖面的度量,衡量了分类器对正例的识别能力;$F1$ 为精确率和召回率的调和均值. 其公式为:

$$P=n_p/n_t, \tag{11}$$

$$R=n_p/n_c, \tag{12}$$

$$F1=\frac{2\times P\times R}{P+R}, \tag{13}$$

式中,n_p 表示正确识别的实体数,n_t 表示抽取的实体数,n_c 表示语料库中的实体数.

4.3　实验结果与分析

根据化工专家信息将实体分为人名、机构名、职称、联系方式、电子邮件、研究方向及化工关键词 7 个类别. 通过 1 089 条化工技术行业中英文关键词对化工领域术语进行边界特征分析,统计出化工领域术语上下文边界二元词共 194 750 对,图 5 为边界二元词的可视化展示. 同时统计边界二元词频集及出现频率,得到边界特征向量矩阵. 尽管和弦图无法表达术语方向,但和弦图的节点宽度给出了重量的直观表示.

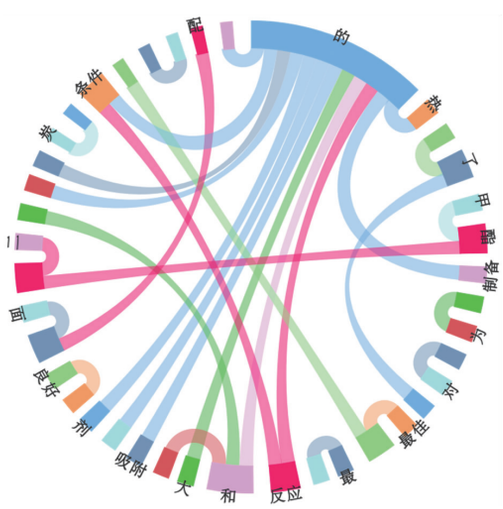

图 5　边界二元词的可视化展示

Fig. 5　Visual display of boundary binary words

实验以 BiLSTM-CRF 为基线对于加入特征提取算法进行测试,模型通过多特征与非字嵌入特征提取随机生成向量作为神经网络模型的输入,对比多特征神经网络模型在实体抽取任务上带来的增益. 实验结果如表 4 所示,相较于传统仅使用神经网络模型作为特征提取,加入字特征和边界特征对模型抽取的结果带来的效果最高. 模型的综合 $F1$ 值达到 94.69%,高于其他机器学习模型和深度学习模型.

本文重点以抽取领域专家实体为目标,以专业领域性实体为代表,通过抽取部分化工领域性实体对比本文提出的多特征双向门控神经网络和 BiLSTM-CRF 在化工专家语料上实体抽取的差异,如表 5 所示. 多特征双向门控神经网络模型能完整地识别专业术语“Pictet-Spengler 反应”,而 BiLSTM-CRF 只能识别“反应”两个字,可见本文模型在识别中英文混杂化工专业关键词实体上具有很好的效果.

表 4　多特征神经网络模型实体抽取任务实验比较

Table 4　Experiment comparison of multi-feature neural network model entity extraction task

算法	精确率 P/%	召回率 R/%	$F1$/%
HMM	55.81	72.73	63.16
IDCNN-CRF	98.78	85.83	84.05
BiLSTM-CRF	98.81	85.26	84.71
W-BiLSTM-CRF	99.02	87.82	89.29
多特征神经网络模型	99.63	95.68	94.69

表 5　本文模型和 BiLSTM-CRF 的实体识别效果

Table 5　The entity recognition effect of the model in this paper and BiLSTM-CRF

关键词实体	BiLSTM-CRF	多特征门控神经网络模型
Pictet-Spengler 反应	反应	Pictet-Spengler 反应
TAME 原料预处理	TAME/原料	TAME 原料/处理
金属学及金属工艺	金属/工艺	金属学/金属工艺
ICP-OES	ICP/OES	ICP-OES

5　结论

本文提出的基于多特征双向门控神经网络的领域专家实体抽取方法,在抽取专家个人信息的同时,抽取代表其研究领域相关的关键词实体. 该方法通过分析使用 Bert 语言模型获取中文字符字特征,统计关键词上

下文边界词分析获取边界特征;采用双向门控神经网络结合 Attention 机制获取字符上下文依赖特征;使用 CRF 进行序列标注. 以化工专家数据集为例,实验结果表明,该方法能够有效识别化工领域关键词实体. 然而,在抽取的关键词实体中仍然存在相似性较高的词汇如“环氧化酶-2,环氧合酶-2”. 因此,在抽取领域专家信息实体之后,如何抽取并利用实体之间的关系进行歧义性分析是本文进一步研究的重点.

[参考文献]

[1] 邹博伟,钱忠,陈站成,等. 面向自然语言文本的否定性与不确定性信息抽取[J]. 软件学报,2016,27(2):309-328.

[2] LI J,SUN A,HAN J,et al. A survey on deep learning for named entity recognition[J]. IEEE transactions on knowledge and data engineering,2020,32(3):1558-2191.

[3] GE H,CAVERLEE J,LU H. Taper:a contextual tensor-based approach for personalized expert recommendation[C]//Proceedings of the 10th ACM Conference on Recommender Systems. Boston,2016:261-268.

[4] LI X,WANG Z,GAO S,et al. An intelligent context-aware management framework for cold chain logistics distribution[J]. IEEE transactions on intelligent transportation systems,2019,20(12):4553-4566.

[5] BARTOLI A,DE LORENZO A,MEDVET E,et al. Active learning of regular expressions for entity extraction[J]. IEEE transactions on cybernetics,2017,48(3):1067-1080.

[6] ZHANG Y,YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne,2018:1554-1564.

[7] 汪诚愚,何晓丰,宫学庆,等. 面向上下位关系预测的词嵌入投影模型[J]. 计算机学报,2019,43(5):868-883.

[8] MORWAL S,JAHAN N,CHOPRA D. Named entity recognition using hidden Markov model(HMM)[J]. International journal on natural language computing,2012,1(4):15-23.

[9] MCCALLUM A,FREITAG D,PEREIRA F C N. Maximum entropy Markov models for information extraction and segmentation[C]//Proceedings of International Conference on Machine Learning. Stanford,2000:591-598.

[10] LAFFERTY J,MCCALLUM A,PEREIRA F C N. Conditional random fields:probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning. Williams College,MA,2001:282-289.

[11] DEVLIN J,CHANG M W,LEE K,et al. Bert:pre-training of deep bidirectional transformers for language understanding[J]. Computation and language,2018,23(2):3-19.

[12] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural language processing(almost)from scratch[J]. Journal of machine learning research,2011,12(1):2493-2537.

[13] STRUBELL E,VERGA P,BELANGER D,et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen,2017:2670-2680.

[14] HUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer science,2015(8):1508-1518.

[15] WANG S,LI Y,LIU N,et al. Noisy-data-disposing algorithm of data clean on the attribute level[J]. Computer engineering,2005(9):86-87.

[16] 张华平,吴林芳,张芯铭,等. 领域知识图谱小样本构建与应用[J]. 人工智能,2020(1):113-124.

[17] 唐明,朱磊,邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学,2016,43(6):214-217,269.

[18] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems,2013,26:3111-3119.

[19] PENNINGTON J,SOCHER R,MANNING C D. Glove:global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha,2014:1532-1543.

[20] 严云洋,瞿学新,朱全银,等. 基于离群点检测的分类结果置信度的度量方法[J]. 南京大学学报(自然科学版),2019,55(1):102-109.

[21] GOUTTE C,GAUSSIER E. A probabilistic interpretation of precision,recall and F-score,with implication for evaluation[C]//Proceedings of European Conference on Information Retrieval. Springer,Berlin,Heidelberg,2005:345-359.

[责任编辑:丁 蓉]