

藏药药理命名实体识别^{*}

何家欢 刘勇国 蒋 羽 张 云 张春梅 李东晓 张 艺

(电子科技大学信息与软件工程学院
成都 610054)

(四川省中医药科学院
成都 610041)

(成都中医药大学
成都 611137)

〔摘要〕 通过中国知网获取藏药药理相关文献 155 篇, 构建中文藏医药药理实体识别语料库, 设计基于 BiLSTM - CRF 深度学习的藏药药理命名实体识别方法, 采用信息抽取技术从科技文献中提取并识别藏药药理, 为藏医药文献研究提供新途径。

〔关键词〕 藏医药; 药理; 命名实体识别; BiLSTM - CRF

〔中图分类号〕 R - 056 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673 - 6036. 2020. 04. 006

Pharmacological Named Entity Recognition of Tibetan Medicine HE Jiahuan, LIU Yongguo, JIANG Yu, ZHANG Yun, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China; ZHANG Chunmei, LI Dongxiao, Sichuan Academy of Chinese Medical Sciences, Chengdu 610041, China; ZHANG Yi, College of Ethnic Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

〔Abstract〕 The paper obtains a total of 155 literatures related to pharmacology of Tibetan medicine through CNKI, builds a Chinese pharmacological entity recognition corpus for Tibetan medicine, designs a pharmacological named entity recognition method for Tibetan medicine based on BiLSTM - CRF deep learning model, uses the information extraction technology to extract and identify the pharmacology of Tibetan medicine from scientific literatures, provides a new way for the study on Tibetan medicine literatures.

〔Keywords〕 Tibetan medicine; pharmacology; named entity recognition; BiLSTM - CRF

1 引言

1.1 藏医药研究情况

藏医学是一门古老科学, 已有数千年历史, 是 4 大传统医学体系之一, 是通过不断的实践和研究逐渐形成的集人体生理、病理、诊断、治疗、方剂

等为一体的民族医药学^[1]。近年来我国藏医药研究的科技文献逐年递增^[2]。藏医药科技文献的增长有利于发现藏医药领域研究热点和进展, 明确藏医药研究方向和思路, 促进藏医药的研究和发展。现有藏医药科技文献研究以手工方式或运用统计软件分析为主。扎西道知等^[3]通过查询藏医药文献并整理藏药不良反应描述, 为藏医药不良反应研究提供文

〔收稿日期〕 2019 - 09 - 18

〔作者简介〕 何家欢, 硕士研究生; 通讯作者: 刘勇国, 教授, 博士生导师。

〔基金项目〕 国家重点研发计划“民族医药发掘整理与学术传承研究”(项目编号: 2017YFC703900); 四川省科技基础条件平台项目“中医药研发创新信息支撑服务平台”(项目编号: 2018TJPT0039); 四川省应用基础研究计划“网络药理学/药效学并行的血小板活化抑制剂筛选及 SLE 通路干预研究”(项目编号: 2018JY0659)。

献依据。高路等^[2]使用软件 CiteSpace 对文献数据进行统计分析,发现藏医药研究进入稳定发展期。李琪等^[4]应用中医传承辅助平台^[5]对藏药治疗肝炎方剂进行药物频次统计和药物关联度分析,运用聚类分析提取核心组合。手工方式分析藏医药文献存在样本小、主观性强、周期长和人力成本高的问题,应用商业软件可分析藏医药概念(治法、方剂和药物等)出现频率和共现关系,但无法分析藏医药概念深层语义。药物药理用于阐明药物对机体的作用和原理,以及药物在体内吸收、分布、生物转化和排泄等过程,药物效应和血药浓度随时间消长的规律^[6]。药物药理研究能够了解药物治疗疾病原理,从而针对疾病研发新药物,针对药物寻找新的适应症,药理研究在临床中具有重要价值^[7]。

1.2 生物医学领域文献挖掘研究情况

近年来文献挖掘在生物医学研究领域逐步开展, Murugesan 等^[8]提出 BCC - NER 方法,使用双向条件随机场对 BioCreative II 语料库的基因和蛋白质进行识别, F1 值达 86.95%。何林娜等^[9]提出基于特征耦合泛化(Feature Coupling Generalization, FCG)的半监督学习方法,结合条件随机场(Conditional Random Fields, CRF)自动识别 1979 - 2009 年 PubMed 生物医学文献的药物名, F1 值达 76.73%。王振华等^[10]通过研究藏药镰形棘豆科技文献 117 篇,发现研究工作集中于化学成分、药理作用方面。仁青东主等^[11]研究藏药复方治疗慢性萎缩性胃炎方剂相关文献,统计药物频次并分析药物功效。通过文献挖掘方法分析研究主题及其演变趋势,总结研究内容并提取关键信息,能高效挖掘海量文献数据,显著降低手工分析强度,提高科技文献分析效率。文献挖掘技术可多方面辅助研究人员开展科研工作,例如文本分类可自动对文本进行筛选,缩小待研究的文本范围;命名实体识别可从生物医学文献中抽取所需特定事实信息,如疾病、药物、基因和蛋白等有特定意义的生物实体;自动摘要可从海量文献中自动进行文本总结并生成文档文摘。

1.3 研究背景

目前药理研究通常通过临床实验提取药物化学成分,研究药物活性成分以确定药物药理,研究过程周期长且成本高^[12]。从生物医学研究领域文献中提取药物药理,可为新药研发实验提供方向、辅助决策,减少实验周期时间并提高研究效率。当前生物医学研究领域文献的实体识别相关研究都针对英文文献,未开展针对中文文献实体识别研究。中文命名实体识别相较于英文命名实体识别更加困难,中文词与词之间没有类似英文单词之间明显的边界。本研究构建中文藏医药药理实体识别语料库,设计基于 BiLSTM - CRF 深度学习模型的藏药药理命名实体识别方法,采用文献挖掘的信息抽取技术从科技文献提取并识别藏药药理,协助藏医药研究人员开展藏药药理研究,为藏医药文献研究提供新途径。

2 语料库与方法

2.1 药理实体识别流程(图1)

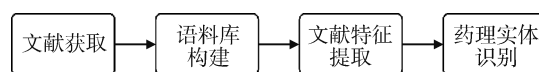


图1 药理实体识别流程

文献获取指获得藏药药理相关研究文献;语料库指去除文献中噪声信息(如页码、分类号、格式符等),保留文献正文文本并对文本句子中的药理实体进行标注;文献特征提取指将文献文本从高维空间向低维空间投影降维,以保存并表征文献内容的关键信息^[13];药理实体提取指使用提取的特征向量计算每个字最大概率标签,以实现药理实体提取。

2.2 文献获取

通过中国知网以“藏医药”和“藏药”为关键字检索 2014 - 2018 年《中草药》、《中国中药杂志》、《中医杂志》、《世界科学技术 - 中医药现代

化》、《中成药》、《中国民族医药杂志》、《中华中医药杂志》等期刊 479 篇文献，以文献中是否包含藏药药理为筛选标准，筛选获得藏医药药理相关文献 155 篇。

2.3 语料库构建

语料库指经科学取样和加工的电子文本库。将获取的文献中页码、分类号、格式符等不相关信息、图片和表格等数据剔除，保留正文文本，构建藏医药文献语料库并实现语料库标注。本研究使用标签集 {B, I, O} 进行语料库标记^[14]，其中 B 表示实体开始，I 表示实体中首词以外部分，O 表示非藏药药理实体部分，标记实例，见图 2，其中包含药理实体“抗氧化”。语料库中给定语句 $X_m = \{x_m^1, \cdots, x_m^t, \cdots, x_m^n\}$ ，其中 x_m^t 表示语料库第 m 句的第 t 个词， n 表示语句长度。以句子为单位，随机选取 90% 语料为训练数据以训练实体识别模型，其余 10% 语料为测试数据以评价模型识别性能，本研究藏医药文献语料库，见表 1。

句子	余甘子、诃子均具有明显的抗氧化作用。										
标签	O	O	O	O	O	O	O	O	O	B	I

图2 文本标记实例

表1 藏医药文献语料库		
项目	训练语料	测试语料
句子	5 541 句	719 句
药理数	2 018 个	262 个
字数	403 488 字	52 450 字

2.4 文献特征提取

2.4.1 概述 药理实体识别的文献特征指药理实体出现于文献句子的位置，构成药理实体的特殊字等多种特征。文献特征提取能自动提取文本特征，将特征表示为向量形式。本研究将藏药药理命名实体识别视为序列标签预测问题，预测语料库中字 {B, I, O} 标签以检测藏药药理实体。通过文献提取藏药药理，探索相同或相似药理药物，为新药研

发提供实验思路，缩短实验周期^[15]。传统序列标注预测方法通过统计方法如条件随机场进行等，需人工定义大量特征且泛化能力不足^[16]。深度学习方法模拟人脑进行非线性分析学习，无需人工干预可自动完成藏医药文本特征学习。本研究引入 BiLSTM - CRF 深度学习模型实现藏医药文献的药文献的药理实体识别，见图 3。模型包括：输入层、嵌入层、特征提取层、输出层和 CRF 层。

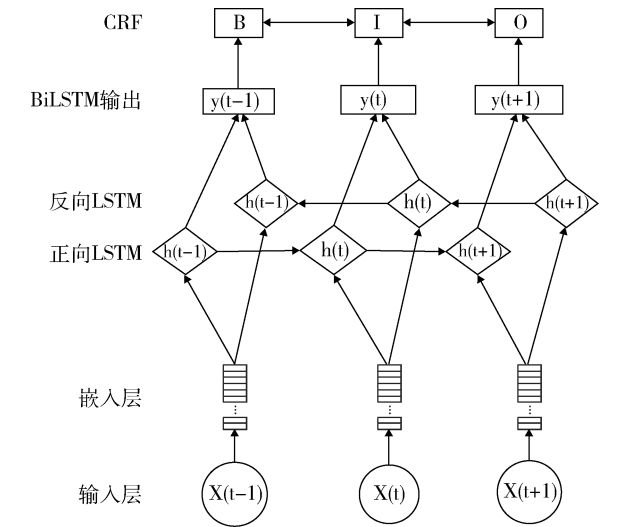


图3 BiLSTM - CRF 深度学习模型

2.4.2 输入层 BiLSTM - CRF 模型以藏医药文献语料库中词 x_m^t 为单元输入，以语句 X_m 为输入信息。例如 $X_m =$ “余甘子、诃子均具有明显的抗氧化作用。”，则 $x_m^1 =$ ‘余’， $x_m^2 =$ ‘甘’， $x_m^3 =$ ‘子’， $\cdots$ ， $x_m^{17} =$ ‘用’， $x_m^{18} =$ ‘。’。以语料库中每个语句输入到 BiLSTM - CRF 模型。

2.4.3 嵌入层 将语句 X_m 中字 x_m^t 数字化表示，方便模型学习藏药药理实体特征，字向量实例，见表 2。数字化表示是一种分布表示特征，即用连续实数向量表达藏医药文献语料库中字特征。

表2 字向量实例					
字	向量				
余	-1.196 4	-1.584 5	1.733 3	...	-1.067 4
甘	0.643 1	-1.265 5	0.331 5	...	-3.439 6
子	-0.666 0	-3.275 4	-0.234 4	...	1.051 4
用	-0.816 1	-1.635 6	-0.838 2	...	-1.244 5
。	0.318 8	-2.976 0	0.516 0	...	-1.130 6

2.4.4 特征提取层 长短期记忆网络 (Long Short Term Memory, LSTM) 结构^[17] 含有 1 个或多个具备可遗忘和记忆功能的单元组成, 可实现长距离文本信息利用, 其结构, 见图 4。

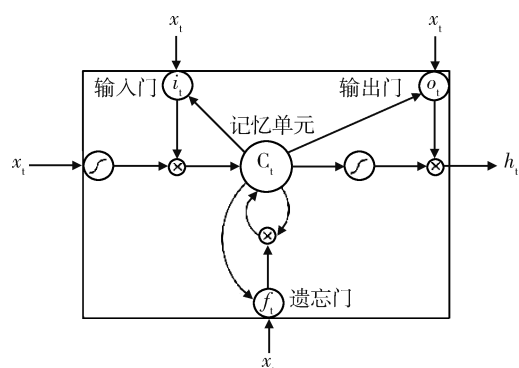


图 4 LSTM 单元结构

LSTM 结构核心是记忆单元, 通过输入门 i , 遗忘门 f , 输出门 o 控制记忆单元每时刻输出, 计算过程如式 (1) - (5) 所示:

$$i^t = \sigma(W_{ix} x_m^t + W_{ih} h_m^{t-1}) \quad (1)$$

$$f^t = \sigma(W_{fx} x_m^t + W_{fh} h_m^{t-1}) \quad (2)$$

$$c^t = f^t \cdot c_m^{t-1} + i^t \cdot \varphi(W_{cx} x_m^t + W_{ch} h_m^{t-1}) \quad (3)$$

$$o^t = \sigma(W_{ox} x_m^t + W_{oh} h_m^{t-1}) \quad (4)$$

$$h^t = o^t \cdot \varphi(c^t) \quad (5)$$

其中 W_{uv} 表示训练权重矩阵, $u \in (i, f, c, o)$, $v \in (h, x)$, σ 和 φ 分别表示 sigmoid 和 tanh 非线性激活函数^[18]。该门控机制能对记忆单元信息进行处理, 确定输入文本信息遗忘或保留, 以解决长距离的信息处理问题。如语句“含有诃子的方剂的主治疾病多与热毒、疼痛相关, 因此其高频次药物也多具有清热解毒、活血通络、理气止痛作用, 其中以——、理气药使用最多”, 需对横线词进行预测, LSTM 单元通过学习, 记忆单元保留“诃子”、“热毒”等强相关信息, 遗忘门遗忘“的”、“方剂”等弱相关信息, 输出门将最后所获信息输出, 预测横线词为“清热药”。

双向长短期记忆网络 (Bi-directional Long Short Term Memory, BiLSTM) 模型同时考虑藏医药文献的正向和反向文本信息, 例中通过反向传递信息预先明确“清热药”, 能预测上文对应的“热毒”

信息。BiLSTM 模型相比 LSTM 模型能获取更佳特征表示^[19]。本研究引入 BiLSTM 自动提取藏医药科技文献语料库中藏药药理命名实体特征, 如藏医药药理实体位置、前后特殊文字等。BiLSTM 单元输出 $h^t = (\vec{h}^t, \overleftarrow{h}^t)$, $\vec{h}^t$ 表示正向 LSTM 单元输出, $\overleftarrow{h}^t$ 表示反向 LSTM 单元输出, 将正向和反向输出拼接以获得 BiLSTM 输出。给定输入句 $X_m = \{x_m^1, \dots, x_m^t, \dots, x_m^n\}$, 通过特征提取层输出表示序列 $h_m = \{h_m^1, \dots, h_m^t, \dots, h_m^n\}$, h_m^t 表示特征提取层输出, 即表示当前字是否属于实体的语义特征。

2.5 药理实体识别

2.5.1 输出层 利用字特征表示向量 h_m 计算词归属标签 $\{B, I, O\}$ 概率 y_m^t , 如式 (6) 所示。获得每个字的标签后, 根据字标签对句中药理实体进行提取。其中 M 为参数矩阵, b 为偏移量。

$$y_m^t = M h_m^t + b \quad (6)$$

2.5.2 条件随机场层 为藏药药理实体标签序列添加约束规则, 模型训练过程中 CRF 层通过转移矩阵 A_{ij} 学习藏药实体标签间约束规则, 如标签 B 后应为标签 I, 因为药理实体通常由多词组成; 标签 O 后面不可能是标签 I, 因为实体开始标签应为 B 标签。通过上述约束可避免生成错误标签序列。给定语句 X_m 和藏药药理实体标签序列 z , 所预测标签序列得分如式 (7) 所示:

$$s(X_m, z, \theta) = \sum_{t=1}^n A_{z^{t-1} z^t} + y_{z^t}^t \quad (7)$$

其中 $y_{z^t}^t$ 表示 BiLSTM 模型获得的语句 X_m 中第 t 个词属于药理实体的概率, $A_{z^{t-1} z^t}$ 为输入句第 $t-1$ 个字的实体标签 z^{t-1} 转移到第 t 个字的实体标签 z^t 概率, $\theta \in \{W_{uv}, W, b \mid u \in (i, f, c, o), v \in (h, c)\}$ 表示可训练参数。

3 训练

3.1 目标函数

给定语句 X_m 获得可能的药理实体标签序列 $Z(X_m)$, 则药理实体标签序列 z 概率为:

$$p\left(\frac{z}{x}\right)=\frac{e^{s\left(X_m, z\right)}}{\sum_{z^{\prime} \in Z\left(X_m\right)} e^{s\left(X_m, z^{\prime}\right)}} \quad(8)$$

最终获得预测输出序列 z^{n*} 是最大概率藏药药理实体标签序列，如式（9）所示：

$$z^{n*}=\operatorname{argmaxs}\left(X_m, z, \theta\right), z \in Z\left(X_m\right) \quad(9)$$

利用 Adam^[20] 算法训练 BiLSTM – CRF 深度学习模型，更新参数并最大化正确标签序列概率。将测试文本输入模型，获取最大概率藏药药理标签序列，以实现藏医药文献语料标注并提取藏药药理实体。

3.2 Dropout 策略

过拟合现象是神经网络模型的常见问题，如藏医药文献语料多处出现藏药药理实体“清热”，其对应标签分别为“B”和“I”，过拟合导致后续藏医药文献语料出现“清”字时被标记为实体开始，错误识别药理实体。Dropout 策略使神经网络节点随机失活，避免训练模型过度依赖特定特征，具备正则化作用^[21]。本研究为避免模型过拟合，在嵌入层与特征提取层间、特征提取层和 CRF 层间引入 Dropout 策略。

3.3 参数设置

嵌入层中不同维度的词向量表达藏医药文献中不同语义信息，故序列标签预测问题中词向量维度影响学习模型的预测性能^[22]。本研究采用 Word2vec 模型^[23] 和 Cw2vec 模型^[24] 训练词向量。以 475 篇藏医药文献为训练数据，训练 50 维、100 维、200 维和 300 维藏医药文献词向量。Word2vec 模型采用词上下文训练，Cw2vec 模型采用词笔画训练。同时，在 $[-0.25, 0.25]$ 区间随机采样，对藏医药文献每个词生成随机向量对比。Word2vec 模型训练参数参照文献^[23] 设置，Cw2vec 模型训练参数参照文献^[24] 设置。BiLSTM – CRF 模型其他参数，见表 3。

表 3 BiLSTM – CRF 模型参数

参数名	参数值
隐藏层单元数目	$H=300$
学习率	$\alpha=0.001$
Dropout 比率	$p=0.5$
迭代次数	$epoch=40$

4 实验结果

4.1 评价标准

药理实体是否正确识别主要依据实体边界是否正确。本研究采用精确率（ P ）、召回率（ R ）、 $F1$ 测度值（ $F1$ ）评价指标描述模型性能，如式（10）–（12）所示：

$$P=\text{识别正确的药理数} / \text{识别出的药理数} \quad(10)$$

$$R=\text{识别正确的药理数} / \text{样本的药理数} \quad(11)$$

$$F1=(2 \times P \times R) / (P+R) \quad(12)$$

精确率（ P ）表示所识别藏药药理实体中正确识别的藏药药理实体占比；召回率（ R ）表示语料库中藏药药理实体正确识别占比； $F1$ 测度（ $F1$ ）表示精确率和召回率的综合结果，判断识别模型能否准确、全面识别藏药药理实体。

4.2 结果分析

随机向量、Word2vec 和 Cw2vec 向量模型在不同维度下的藏药药理识别结果，见表 4。可知相同向量模型条件下 50 维向量时藏药药理实体识别效果最好。相同维度条件下 Word2vec 和 Cw2vec 模型识别效果优于随机向量，说明 Word2vec 和 Cw2vec 模型能够有效获取藏医药文献的文本语义信息。Cw2vec 模型 50 维、200 维和 300 维向量条件下藏药药理实体识别效果优于 Word2vec 模型。Word2vec 模型仅使用藏医药文献的上下文信息，没有考虑单词内部特征；而 Cw2vec 模型使用藏医药文献单词的笔画特征训练模型，考虑上下文信息同时考虑单

词内部语义信息，更适合训练中文词向量，其在语义相似性方面也被证实比 Word2vec 具有更好效果^[24]。不同模型 $F1$ 测度值，见图 5。实验模型随着维度变化的召回率和精确率变化趋势，见图 6。

可见向量维度降低导致模型 $F1$ 测度值增加，模型召回率随着维度降低而增加， $F1$ 测度值和召回率成正比。

表 4 不同维度词向量性能对比

向量	50			100			200			300		
	P	R	$F1$	P	R	$F1$	P	R	$F1$	P	R	$F1$
随机向量	74.42	61.07	67.09	84.18	50.76	63.33	82.89	48.09	60.87	68.23	50.00	57.71
Word2vec	72.53	64.50	68.28	75.90	61.36	67.86	80.50	52.27	63.39	72.55	48.05	57.81
Cw2vec	89.02	55.73	68.54	82.22	56.49	66.97	78.57	54.58	64.41	76.25	46.56	57.82

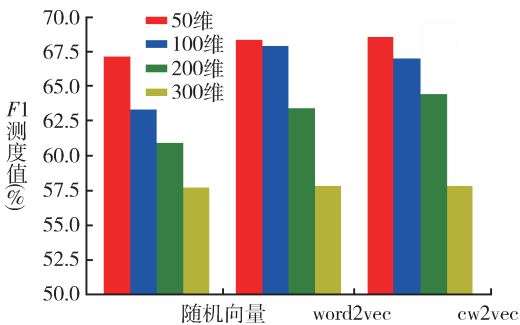


图 5 实验模型 $F1$ 测度值

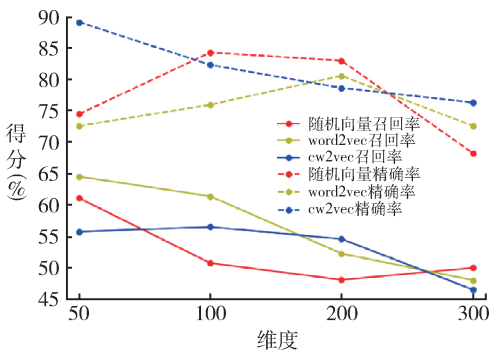


图 6 模型精确度、召回率

4.3 实例分析

模型对文本序列进行标注的实例，见图 7。模型正确识别句子 1 中药理实体“开窍通络”、“镇静安神”、“解痉止痛”。识别句子 2 中药理实体“清血热”、“清浊”，但真实药理实体为“清血热”、“分清血液清浊”。实体“分清血液清浊”错误识别

原因在于训练语料中出现大量动词 – 宾语结构的词标记为药理实体，如“清热”、“抗氧化”等，导致模型以动词 – 宾语结构短语为药理实体，造成含有动宾结构的长药理实体分割，产生识别错误。此错误因实体分布不平衡引发，即语料库中动宾短语药理实体占比过大，后续研究中尝试构建长短药理实体均匀分布语料库训练模型。

句子	麝香具有开窍通络，镇静安神，解痉止痛功效。
标签	O O O O B I I I O B I I I O B I I I O O O
句子	余甘子具有清血热和分清血液清浊之功效。
标签	O O O O O B I I O O O O O B I O O O O

图 7 药理识别实例

5 结语

针对现有藏医药科技文献研究过程中手工方式

和统计软件应用方面存在的不足，本研究提出基于 BiLSTM – CRF 深度学习模型的藏药药理命名实体识别方法，通过科技文献自动提取并识别藏药药理。模型采用少量标注数据训练并使用未标记的藏医药

文献训练字向量,能提取文献的藏药药理实体,为藏医药研究人员开展藏医药药理研究提供新途径。

参考文献

- 1 杨忠太,万德加.浅谈世界四大传统医学中的藏医药学地位[J].世界最新医学信息文摘,2016,16(75):200.
- 2 高路,高艳菊,杨青松,等.当代中国藏医药领域研究热点和前沿趋势的可视化分析[J].中央民族大学学报(自然科学版),2018,27(4):75-82.
- 3 扎西道知,白玛仁增,罗布顿珠.基于藏医药古籍文献对当前藏药说明书不良反应项窥见[J].中国民族医药杂志,2018,24(4):37-38.
- 4 李琪,李海娇,徐僮,等.基于数据挖掘的藏医药治疗肝炎用药规律分析[J].中国民族民间医药,2018,27(12):4-7.
- 5 卢朋,李健,唐仕欢,等.中医传承辅助系统软件开发与应用[J].中国实验方剂学杂志,2012,18(9):1-4.
- 6 李宏伟.实用医药基础[M].浙江:浙江大学出版社,2006:49.
- 7 李群.中药药性的理论与药理作用分析[J].世界最新医学信息文摘,2016,16(27):168.
- 8 Murugesan G, Abdulkadhar S, Bhasuran B, et al. BCC - NER: bidirectional, contextual clues named entity tagger for gene/protein mention recognition [J]. EURASIP J Bioinform Syst Biol, 2017, 2017(1): 7-22.
- 9 何林娜,杨志豪,林鸿飞,等.基于特征耦合泛化的药名实体识别[J].中文信息学报,2014,28(2):72-77.
- 10 王振华,孟祥云,杨丽霞,等.基于CNKI数据资料剖析藏药镰形棘豆的文献特征和研究现状[J].西部中医药,2017,30(7):65-67.
- 11 仁青东主,华青措,斗周才让,等.藏药复方治疗慢性萎缩性胃炎的配伍规律研究[J].中国中医基础医学杂志,2017,23(7):1009-1011.
- 12 李宏伟.实用医药基础[M].浙江:浙江大学出版社,2006:50.
- 13 徐冠华,赵景秀,杨红亚,等.文本特征提取方法研究综述[J].软件导刊,2018,17(5):13-18.
- 14 Luo L, Yang Z H, Yang P, et al. An Attention - based BiLSTM - CRF Approach to Document - Level Chemical Named Entity Recognition [J]. Bioinformatics, 2017, 34(8): 1381-1388.
- 15 向杨峰.基于数据挖掘的新药研发系统[D].北京:北京交通大学,2010.
- 16 Chen X C, Qiu X P, Zhu C X, et al. Long Short - term Memory Neural Networks for Chinese Word Segmentation [C]. Lisbon: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1197-1206.
- 17 Hochreiter S, Schmidhuber J. Long Short - term Memory [J]. Neural Comput, 1997, 9(8): 1735-1780.
- 18 田娟,李英祥,李彤岩.激活函数在卷积神经网络中的对比研究[J].计算机系统应用,2018,27(7):43-49.
- 19 Ma X, Hovy E. End - to - end Sequence Labeling via Bi - directional LSTM - CNNs - CRF [C]. Berlin: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1064-1074.
- 20 Kingma D P, Ba J. Adam: a method for stochastic optimization [C]. San Diego: The 3rd International Conference for Learning Representations, 2015.
- 21 Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. J Mach Learn Res, 2014, 15(1): 1929-1958.
- 22 Ronan C, Jason W, Léon B, et al. Natural Language Processing (Almost) from Scratch [J]. J Mach Learn Res, 2011(12): 2493-2537.
- 23 Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [C]. Scottsdale: Proceedings of Workshop at ICLR, 2013.
- 24 Cao S S, Liu W, Zhou J, et al. Cw2vec: learning Chinese word embeddings with stroke n - gram information [C]. New Orleans: Proceedings of the Thirty - second AAAI Conference on Artificial Intelligence, 2018: 5053-5061.