面向化学资源文本的命名实体识别

马建红，　王立芹，　姚　爽

(河北工业大学 计算机科学与软件学院　天津 300401)

摘要：针对化学资源文本中的命名实体,提出一种适合于化学资源文本的命名实体识别方法,旨在将化学物质、属性、参数、量值 4 种命名实体进行识别. 该方法根据化学资源文本的语言规律及特点,建立 BLSTM-CRF 模型对命名实体进行初步识别,并使用基于词典与规则相结合的方法对识别结果进行校正. 实验结果表明,该方法在化学资源文本中能够较好地完成命名实体识别任务,在测试语料上的 $F1$ 值最高能达到 94.26%.

关键词：化学资源文本；命名实体识别；双向长短时记忆网络；条件随机场；规则

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1671-6841(2018)04-0014-07

DOI：10.13705/j.issn.1671-6841.2017299

0　引言

命名实体识别作为自然语言处理领域的一个重要研究方向[1],其在传统领域的研究较为成熟,$F1$ 值可达到 90% 以上[2]. 相比传统领域,专业领域命名实体的产生往往以该领域知识为依据,兼顾其语言规律特性,使得识别难度大大增加. 本文选择对化学资源文本的命名实体进行研究,提出一种面向化学资源文本的命名实体识别方法. 化学资源文本,是指包含化学资源的自由文本. 化学资源命名实体包括化学物质、属性、参数和量值. 其中属性是化学物质之间进行区分的必然性质. 参数是属性在某个时空状态下的显性度量,例如:可燃性是化学物质的属性,而自燃点是可燃性的一个参数,自燃点会随着外界条件(大气压等)的变化而变化,而可燃性却是客观存在的. 化学资源文本中命名实体识别主要存在以下难点:

1) 化学资源文本和化学资源命名实体数量庞大且表达方式复杂多样,命名实体结构没有严格的构词规律可循,例如化学物质由单个字的物质(溴、银等)到由字和相关符号组成的有机物(邻苯二甲酸二(2-乙基己)酯)等. 量值包含了更多的符号(g/cm^3、℃等),并且有的量值字数较多(1 g 该品可溶于约 5.5 mL 水(约 18%,25 ℃)、83 mL 醇).

2) 化学资源文本在描述化学物质时往往将属性与参数混为一谈,使得二者在句中所处的语法位置相同,区分造成困难.

很多学者对化学资源领域的命名实体识别均有研究. 文献[3]通过基于词典的方法来识别小分子物质和药品,这种方法依赖于词典和匹配算法的质量. 文献[4]分析化学物质的领域特征以及统计语言特征,制定规则识别有效词类序列,这种方法可移植性很弱,当数据稍有变化时,就要修改或增加相应的规则. 文献[5-6]使用条件随机场(conditional random field,CRF)等机器学习模型进行化学物质识别,然而基于机器学习的方法需要依据逻辑直觉手工定制大量特征,其识别性能很大程度上依赖于特征的质量. 近年来,深度学习被应用到命名实体识别任务中,其中最常见的神经网络模型有卷积神经网络(convolutional neural network,CNN)[7]、循环神经网络(recurrent neural network,RNN)[8-9]、长短时记忆网络(long short-term memory,LSTM)[10-11]以及双向长短时记忆网络(bi-directional LSTM,BLSTM)[12-13]. 使用深度学习模型不仅能减少机器学习中手工定制特征的工作量,还能够自动从词及句子中获得更加有效的特征.

针对化学资源文本命名实体识别难点,本文提出了一种面向化学资源文本的命名实体识别方法. 首先采

收稿日期:2017-09-27

基金项目:中国科学技术咨询服务中心计算机辅助创新设计公共服务平台建设服务采购项目(HSZT2015FD/254).

作者简介:马建红(1965—),女,河北保定人,教授,主要从事计算机辅助创新设计过程与方法、TRIZ、软件工程、CAI 软件技术研究,E-mail:m_zh2002@126.com;通信作者:王立芹(1992—),女,河北唐山人,硕士研究生,主要从事深度学习、自然语言处理研究,E-mail:wlq_hebut@126.com.

用基于 BLSTM-CRF 模型的初始识别,然后利用基于词典与规则相结合的办法优化识别. 该方法一方面通过训练语料,减少了基于以往方法的复杂性和盲目性,同时可以很好地解决识别新的命名实体难的问题;另一方面利用规则和词典,降低了仅使用 BLSTM-CRF 模型对大规模语料的依赖性,有效地解决了识别问题.

1　基于 BLSTM-CRF 模型的初始识别

1.1　预处理

本文首先对化学资源文本进行分词处理,然后定义 9 种标签进行人工标注,即 L = (B-SUB, I-SUB, B-ATT, I-ATT, B-PAR, I-PAR, B-VAL, I-VAL, N),各标签依次分别代表化学物质首部、化学物质内部、属性首部、属性内部、参数首部、参数内部、量值首部、量值内部及其他. 此外,要使用 BLSTM-CRF 模型进行处理,需要将每个词转换为固定维度的 embedding 向量[14]. 设样本句子 $\boldsymbol{X}$ 由 n 个词组成,$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$,embedding 向量 $\boldsymbol{x}_t$ 计算公式为 $\boldsymbol{x}_t = \boldsymbol{W}^{\text{emb}} \boldsymbol{r}_t$,其中:$\boldsymbol{W}^{\text{emb}} \in \mathbf{R}^{d \times |\boldsymbol{V}|}$ 为 embedding 向量查询表,需要训练得到;$\boldsymbol{r}_t \in \mathbf{R}^{|\boldsymbol{V}|}$ 为词的 one-hot 表示;$\boldsymbol{x}_t \in \mathbf{R}^d$,$d$ 为 embedding 向量维度;$|\boldsymbol{V}|$ 为 one-hot 表示下词典 $\boldsymbol{V}$ 的大小.

1.2　基于 BLSTM-CRF 模型的命名实体识别

BLSTM-CRF 模型结构如图 1 所示. 其中,BLSTM 由前馈层、反馈层和输出层组成,词向量分别作为前馈层与反馈层的输入,输出层为前馈层与反馈层的输出向量连接. BLSTM 能够同时考虑文本的上下文信息,然而在处理输出标签有强烈依赖关系的数据时,效果却是有限的. 所以加入 CRF 层,CRF 能够有效获得句子级别的标注信息,进而获得最优的标注序列.

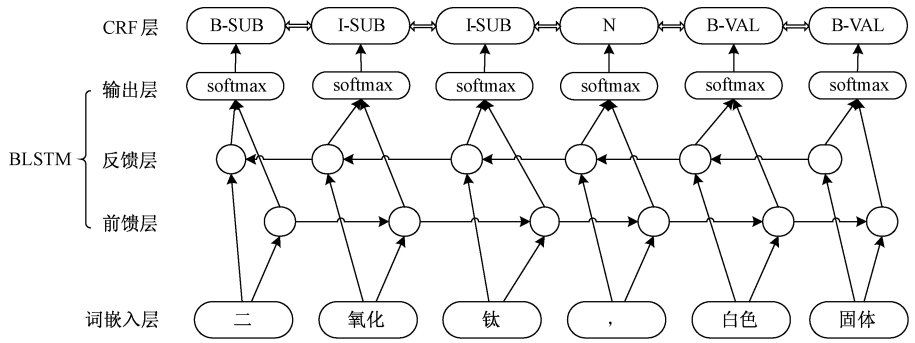

图 1　BLSTM-CRF 模型结构

Fig. 1　Architecture of BLSTM-CRF model

在 BLSTM 层,当输入为 $\boldsymbol{x}_t$,输出为 $\boldsymbol{y}_t$ 时的概率计算公式为

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{h}_{t-1} + \boldsymbol{U}_f \boldsymbol{x}_t + \boldsymbol{b}_f), \boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{h}_{t-1} + \boldsymbol{U}_i \boldsymbol{x}_t + \boldsymbol{b}_i), \tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \boldsymbol{h}_{t-1} + \boldsymbol{U}_c \boldsymbol{x}_t + \boldsymbol{b}_c),$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{h}_{t-1} + \boldsymbol{U}_o \boldsymbol{x}_t + \boldsymbol{b}_o), \boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t), \boldsymbol{P}(y_t \mid \boldsymbol{x}_t) = g(\boldsymbol{W}[\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}] + \boldsymbol{b}_y),$$

其中:σ 为 sigmoid 函数;g 为 softmax 函数;$\odot$ 为点对乘积;$\boldsymbol{i}$,$\boldsymbol{f}$ 和 $\boldsymbol{o}$ 分别为输入门、忘记门和输出门;$\boldsymbol{c}$ 表示 BLSTM 中每个记忆单元的状态;$\boldsymbol{b}$ 是偏置项;$\boldsymbol{W}$、$\boldsymbol{U}$ 为权重矩阵;$[\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}]$ 代表前馈层与反馈层在 t 时刻的输出向量连接.

接着,利用 CRF 层对 BLSTM 层的输出进一步处理. 定义 BLSTM 层的输出概率矩阵为 $\boldsymbol{P}_{n*k}$,k 是标签 L 的个数. $\boldsymbol{P}_{i,j}$是指第 i 个词被标记为第 j 个标签的概率. 对于模型得到的标签序列 $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$,最优标签序列的计算公式为

$$s(\boldsymbol{X}, \boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{A}_{y_{i-1}, y_i} + \sum_{i=1}^{n} \boldsymbol{P}_{i, y_i}, \log p(\boldsymbol{y} \mid \boldsymbol{X}) = s(\boldsymbol{X}, \boldsymbol{y}) - (\log \sum_{\boldsymbol{y}' \in \boldsymbol{Y}} \mathrm{e}^{s(\boldsymbol{X}, \boldsymbol{y}')}),$$

其中:$s(\boldsymbol{X}, \boldsymbol{y})$为标签序列 $\boldsymbol{y}$ 的得分;$\boldsymbol{A}_{y_{i-1}, y_i}$ 代表从第 $i-1$ 个标签转移到第 i 个标签的概率;$\boldsymbol{Y}$ 代表所有可能的标签序列;$p(\boldsymbol{y} \mid \boldsymbol{X})$为标注为标签序列 $\boldsymbol{y}$ 的概率. 为使其达到最大,采用对数最大似然估计得到代价函数,最后使用维特比算法[15],即可求得得分最高的标签序列,即为最优标签序列.

2 基于词典与规则的识别校正

经过大量人工分析发现,化学资源文本中的命名实体构成存在某些特定规律.因此,本文使用基于词典与规则相结合的方法,对 BLSTM-CRF 模型的识别结果进行校正,进一步提高识别性能.

2.1 化学物质命名实体校正

对化学物质进行命名实体识别和校正的对象是化学物质的中文名称和别名.本文受文献[4]的启发,借助化学领域的专业命名规范[16]、化学物质名录[17]和化学资源语料库进行统计分析,建立化学物质特征库如表1所示.并制定如下规则.

以标号 C-E 结尾的化学物质在记录中的比例不足 1%.若模型求出的化学物质尾部标号为 C-E 中的内容,则将尾部移到下一个字,继续判断是否包含,包含继续移动,直到不包含为止.

表1 化学物质特征库

Tab.1 Feature library of chemical substances

标号	特征词类	示例
A	化学元素	氢、氦、锂、银、铂、金…
B	化学专用字	酸、胺、脂、酮…
C	化学介词	化、合、代、聚…
D	特定词头	亚、过、偏、原…
E	其他	阿拉伯数字、罗马数字、天干、英文字母、希腊字母、汉文数字

2.2 属性与参数命名实体校正

人工收集建立属性命名实体库和参数命名实体库.其中属性命名实体库词条总计 158 个,如“致密性”、“导电性”、“可燃性”等,参数命名实体库词条总计 287 个,如“密度”、“电导率”、“自燃点”等,其中 1 个词条表示 1 个命名实体.在以上两个命名实体库的基础上,使用 1 种文本实体匹配算法,该算法思想如下:首先求得库中最长的词条的字数 n,取出待校正样本的前 m 个字,令 $m=n$,判断是否能够和库中的某个词条完全匹配,若完全匹配,无论模型求得的结果如何,都将该字段进行标注,然后删除待校正样本的前 m 个字,继续取出前 $m=n$ 个字进行判断和匹配;若不匹配,$m=m-1$,若仍不匹配就继续 $m=m-1$,直到 m 等于 1,即 m 是 1 个字为止,这时删掉这个字,继续取出前 $m=n$ 个字进行判断和匹配.

2.3 量值命名实体校正

同样建立量值命名实体库,词条总计 344 个,如“无毒”、“白色”、“固体”等.在此基础上,建立量值特征库,如表2所示.对量值命名实体校正制定如下规则.

1)量值命名实体库校正规则. 与文本实体匹配算法相结合进行标注.

2)单位相关量值命名实体校正规则. 以标号 A 的内容结尾,向前标注到不包含 B 的内容为止.

3)溶解性相关量值命名实体校正规则. 以标号 C 的内容开头,向后查询到标号 E 的内容结束,进行标注,以标号 D 的内容结尾,向前查询到标号 E 的内容结束,进行标注.

表2 量值特征库

Tab.2 Feature library of values

标号	特征词类	示例
A	单位	g/cm^3、克/立方厘米、℃、摄氏度、万、%…
B	单位包含词	阿拉伯数字. - ~
C	溶解性词头	不溶、溶于、易溶、微溶
D	溶解性词尾	互溶、混溶
E	点号(除“、”外)	,:;. ? !

3 实验设计与结果分析

3.1 实验语料及语料评价

由于目前没有比较权威统一的语料库,因此语料库从百度百科科学分类下的化工科技词条库的词条中

爬取所得,其中段落按照句号进行切分,整理出 4 328 个句子进行标注. 考虑到化学资源文本的领域性较强,采取人员标注为主,规范制定人员从旁指导的模式,遇到疑难问题经过讨论后达成一致,以此来不断完善规范. 其中标注人员为实验室两名研究生,主要研究方向均为自然语言处理. 规范制定人员主要包括本文作者和一名化学工程专业硕士生. 整个标注分为 4 轮,前 3 轮为预标注,每人标注随机选取的 1 000 个句子相同的语料,旨在完善标注规范及修正问题. 经过 3 轮预标注后,规范趋于稳定,进行正式标注. 语料标注质量采用 F 值[18]的方式进行 IAA 评价[19],F 值计算公式为

$$P = \frac{\text{两个人标注的一致数}}{\text{第一个人标注总数}}, R = \frac{\text{两个人标注的一致数}}{\text{第二个人标注总数}}, F = \frac{2 \times P \times R}{P + R},$$

其中 F 值是通过将一个标注者的标注视为标准,通过计算另外一个标注者的正确率 P 和召回率 R 所得. 对于 4 轮标注,均采用 IAA 评价,随着每轮标注的深入,IAA 结果逐渐上升并趋于稳定. 正式标注 IAA 结果为 94.89,标注结果视为可靠的[20],标注样例如图 2 所示.

聚	甲醛	的	拉	伸	强度	达	70 MPa	,	吸水性
B-SUB	I-SUB	N	B-PAR	I-PAR	I-PAR	N	B-VAL	N	B-ATT
小	,	尺寸	稳定	,	有	光泽	,	这些	性能
B-VAL	N	B-PAR	B-VAL	N	N	B-VAL	N	N	N
都	比	尼龙	好	,	聚	甲醛	为	高度	结晶
N	N	N	N	N	B-SUB	I-SUB	N	B-VAL	I-VAL
的	树脂	,	在	热塑性	树脂	中	是	最	坚韧
N	N	N	N	N	N	N	N	B-VAL	I-VAL
的	.								
N	N								

图 2　标注样例

Fig. 2　Tagging example

3.2　实验评价标准

实验将语料划分为训练集、验证集和测试集,各集合句子数量及 4 种命名实体数量如表 3 所示. 并使用 $F1$($F1$-Measure)值对结果进行评测,其计算公式为

$$P = \frac{N_r}{N_a},\ R = \frac{N_r}{M},\ F1 = \frac{2 \times P \times R}{P + R},$$

其中:N_r 为标注正确的命名实体总数;N_a 为识别出的命名实体总数;M 为测试集合中的命名实体总数.

表 3　数据集描述

Tab. 3　Description of data sets　　个

	句子数	化学物质	属性	参数	量值
训练集	3 000	2 156	578	4 871	6 239
验证集	500	322	51	766	954
测试集	828	639	318	940	2 281

3.3　不同参数下的实验分析

本文使用中科院研制的汉语词法分析系统对语料进行分词. 为了获得高质量的 embedding 向量查询表,本文首先爬取化工科技词条库中所有词条中的自由文本,然后利用 word2vec 工具中的 Skip-gram 模型[21]进行训练得到. BLSTM-CRF 模型使用 theano 实现,并使用反向传播算法进行训练. 针对模型参数复杂的问题,本文做了大量的对比实验,以分析各个参数对模型的标注性能的影响,其中涉及到的参数有:embedding 向量维度、学习率、隐藏层单元数量以及 dropout[22]值. 对每种参数进行实验,其他参数设置为固定值,分别是:embedding 向量维度为 200,学习率为 0.01,隐藏层单元数量为 200,dropout 值为 0.1,实验结果如图 3 所示,其中均值 F 为 4 种命名实体 $F1$ 值的均值.

通过实验可知,各参数均存在局部最优值,即 embedding 向量维度为 200,学习率为 0.005,隐藏层单元数量为 200 以及 Dropout 值为 0.1. 除此之外,各参数对模型的标注性能影响也不尽相同,在此实验中,

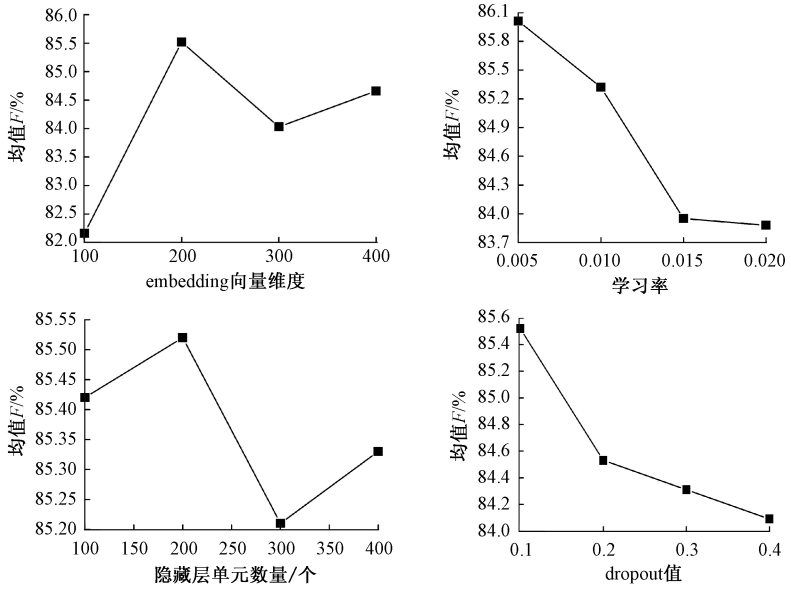

图3 不同参数对结果的影响

Fig. 3 The effect of different parameters on the results

embedding向量维数对模型的标注性能影响最大,隐藏层单元数量对模型的标注性能影响微乎其微.

3.4 不同模型结果的对比分析

本文使用CRF、RNN、LSTM、BLSTM、RNN-CRF、LSTM-CRF、BLSTM-CRF、BLSTM-CRF+校正8种模型进行化学资源文本的命名实体识别.其中CRF使用CRF++0.58版本工具实现,特征选择如表4所示.其他模型参数采用3.3节实验中参数的局部最优值,实验结果如表5所示.

表4 CRF特征集合

Tab. 4 Feature sets of CRF

特征	含义	特征词	含义
$W(i)$	当前词	$POS(i)$	当前词的词性
$W(i-1)$	向前第1个词	$POS(i-1)$	向前第1个词的词性
$W(i-2)$	向前第2个词	$POS(i-2)$	向前第2个词的词性
$W(i+1)$	向后第1个词	$POS(i+1)$	向后第1个词的词性
$W(i+2)$	向后第2个词	$POS(i+2)$	向后第2个词的词性

表5 8种模型对比实验结果

Tab. 5 Comparison of experimental results of eight models %

模型名称	化学物质	属性	参数	量值
CRF	84.56	83.33	84.17	82.19
RNN	83.14	81.95	80.61	78.13
LSTM	83.15	80.56	79.24	78.33
BLSTM	84.46	80.75	78.46	80.96
RNN-CRF	86.29	83.78	83.45	81.46
LSTM-CRF	87.88	82.13	83.10	84.24
BLSTM-CRF	88.23	84.26	83.24	86.33
BLSTM-CRF+校正	89.32	94.26	90.33	88.78

从表5可以看出,本文所用的BLSTM-CRF+规则的识别效果好于其他模型的识别效果.通过BLSTM-CRF和BLSTM-CRF+校正的对比实验可知,引入校正能够明显提升$F1$值,尤其两个模型在识别属性和参数的结果相差很大.由于属性和参数两种命名实体组成结构类似,在句中所处的语法位置相同,造成仅使用BLSTM-CRF模型识别困难,难以区分,校正能够较好地区分这两种命名实体.对于属性,由于属性数量较少,

属性命名实体库已经比较完整和规范,所以加入校正后效果得到非常明显的提升. 相比于属性,参数的数量较多且表达方式多样,所以参数命名实体库无法尽可能地包括所有参数,通过校正得到的结果稍微差一些. 由于标签之间有相对较强的依赖性,例如被标注为“I-SUB”的词,其上一个词的标签一定是“B-SUB”或“I-SUB”,所以在深度模型的基础上引入 CRF 模型能够进一步提升 $F1$ 值. 从 4 种命名实体的识别结果可以看出,对量值的识别结果基本上是最低的,其次是化学物质,这也说明了复杂多样的命名实体结构会给识别造成困难. LSTM 解决了 RNN 梯度消失或爆炸的问题,BLSTM 又在 LSTM 仅考虑上文信息的基础上同时考虑了下文的信息,所以识别结果均优于后一个模型. 由于 CRF 作为传统机器学习的代表模型,在命名实体识别领域已经比较成熟[23-24],并且特征也是人工设定,在文本数量未达到一定规模时能够取得比一般深度学习模型更好的实验效果.

4　结束语

综上所述,本文提出一种面向化学资源文本的命名实体识别方法,该方法在 BLSTM-CRF 模型初步识别的基础上,根据每种命名实体的规律特征,进一步制定不同的词典和规则进行校正,经过对比实验表明,该方法具有较好的准确性和鲁棒性. 接下来将在继续增加语料的基础上,对 4 种命名实体的识别研究扩充到对分子式、特定外界条件的识别研究上来,使识别结果更加严谨而有效. 并制定不同命名实体之间的关系,进行关系抽取,进而扩充化学资源库.

参考文献:

[1] Al-AHMARI S S, Al-JOHAR B A. Cross domains arabic named entity recognition system[C]// International Workshop on Pattern Recognition. Tokyo, 2016: 100111I.

[2] CHOU C L, CHANG C H, HUANG Y Y. Boosted web named entity recognition via tri-training[J]. ACM transactions on Asian and low-resource language information processing, 2016, 16(2): 10.

[3] HETTNE K M, STIERUM R H, SCHUEMIE M J, et al. A dictionary to identify small molecules and drugs in free text[J]. Bioinformatics, 2009, 25(22): 2983 – 2991.

[4] 李楠,郑荣廷,吉久明,等. 基于启发式规则的中文化学物质命名识别研究[J]. 数据分析与知识发现, 2010, 26(5): 13 – 17.

[5] ROCKTÄSCHEL T, WEIDLICH M, LESER U. ChemSpot: a hybrid system for chemical named entity recognition[J]. Bioinformatics, 2012, 28(12):1633 – 1640.

[6] 潘国巍,吉久明,李楠,等. 基于两类统计机器学习模型的中文化学物质名称识别研究[J]. 现代情报, 2011, 31(11): 163 – 165.

[7] DONG X, QIAN L, GUAN Y, et al. A multiclass classification method based on deep learning for named entity recognition in electronic medical records[C]// Scientific Data Summit. New York, 2016: 1 – 10.

[8] LI L, JIN L, JIANG Z, et al. Biomedical named entity recognition based on extended recurrent neural networks[C]// IEEE International Conference on Bioinformatics and Biomedicine. Washington, DC, 2015:649 – 652.

[9] PANDEY C, IBRAHIM Z, WU H, et al. Improving RNN with attention and embedding for adverse drug reactions[C]// International Conference on Digital Health. London, 2017: 67 – 71.

[10] LEE C. LSTM-CRF models for named entity recognition[J]. IEICE transactions on information and systems, 2017, 100(4): 882 – 887.

[11] RONDEAU M A, SU Y. LSTM-based neuroCRFs for named entity recognition[C]// INTERSPEECH. San Francisco, 2016: 665 – 669.

[12] WANG W, BAO F, GAO G. Mongolian named entity recognition with bidirectional recurrent neural networks[C]// International Conference on TOOLS with Artificial Intelligence. San Jose, CA, 2016:495 – 500.

[13] ADAK C, CHAUDHURI B B, BLUMENSTEIN M. Named entity recognition from unstructured handwritten document images [C]// Document Analysis Systems. Santorini, 2016: 375 – 380.

[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. Lake Tahoe, 2013: 3111 – 3119.

[15] 陈宇.基于深度置信网络的中文信息抽取方法[D].哈尔滨:哈尔滨工业大学, 2014.
[16] 中国化学会.化学命名原则[M].北京:科学出版社, 1984.
[17] 国家环境保护总局化学品登记中心.中国现有化学物质名录[M].北京:中国环境科学出版社,2001.
[18] OGREN P V, SAVOVA G K, CHUTE C G. Constructing evaluation corpora for automated clinical named entity recognition [C]// International Conference on Language Resources and Evaluation. Marrakech, 2008:3143 - 3150.
[19] OGREN P V, M.S, SAVOVA G, et al. Building and evaluating annotated corpora for medical NLP systems[J]. Amia Annu Symp Proc, 2006:1050.
[20] PESTIAN J P, BREW C, HOVERMALE D J, et al. A shared task involving multi-label classification of clinical free text[C]// The Workshop on Bionlp 2007: Biological, Translational, and Clinical Language Processing. Prague, 2007:97 - 104.
[21] WERBOS P J. Backpropagation through time: what it does and how to do it[J]. Proceedings of the IEEE, 1990, 78(10): 1550 - 1560.
[22] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer science, 2012, 3(4): 212 - 223.
[23] 蔡肖红,刘培玉,王智昊.基于语境情感消岐的评论倾向性分析[J].郑州大学学报(理学版), 2017, 49(2): 48 - 53.
[24] CHOPRA D, JOSHI N, MATHUR I. Named entity recognition in hindi using conditional random fields[C]// International Conference on Information and Communication Technology for Competitive Strategies. Udaipur, 2016: 106.

Named Entity Recognition for Chemical Resource Text

MA Jianhong, WANG Liqin, YAO Shuang
(*School of Computer Science and Engineering*, *Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: A method was proposed for the recognition of four kinds of named entities, chemical substances, attributes, parameters, and values in the chemical resource text. The language rules and characteristics of the chemical resource text were used for reference. Firstly, BLSTM-CRF model was established to the recognition of named entity. Then the algorithm, which based on the combination of the dictionary and rule, was used to correct and improve the recognition results. The result of experiments showed that the algorithm was able to complete the named entity recognition task in the chemical resource text well, and the maximum *F*1-Measure on the test sets could increase to 94.26%.

Key words: the chemical resource texts; named entity recognition; BLSTM; CRF; rule

(责任编辑:方惠敏)