文章编号：1003-0077(2020)04-0069-08

融合词典特征的 Bi-LSTM-WCRF 中文人名识别

成于思[1]，施云涛[2]

（1. 东南大学 土木工程学院，江苏 南京 210096；
2. 苏宁科技集团云计算研发中心，江苏 南京 210042）

摘 要：受限于标注语料的领域和规模以及类别不均衡，中文人名识别性能偏低。相比人名识别训练语料，人名词典获取较为容易，利用词典提升人名识别性能有待进一步研究。该文提取人名词典特征，融入到双向长短期记忆(Bi-LSTM)网络模型中，在损失函数中提高人名标签权重，设计加权条件随机场(WCRF)。从人名词典中获取姓和名相关的特征信息，Bi-LSTM 网络捕获句子中上下文信息，WCRF 提高人名识别的召回率。在《人民日报》语料和工程法律领域语料上进行实验，结果表明：在领域测试语料上，与基于隐马尔可夫模型的方法相比，人名识别的 F_1 值提高 18.34%，与传统 Bi-LSTM-CRF 模型相比，召回率提高 15.53%，F_1 提高 8.83%。WCRF 还可以应用到其他类别不均衡的序列标注或分类问题中。

关键词：人名识别；双向长短期记忆网络；加权条件随机场；词典特征

中图分类号：TP391 **文献标识码**：A

Bi-LSTM-WCRF Incorporating Dictionary Feature for Chinese Person Name Recognition

CHENG Yusi[1], SHI Yuntao[2]

(1. School of Civil Engineering, Southeast University, Nanjing, Jiangsu 210096, China;
2. Cloud Computing Research Center, Suning Technology Corporation, Nanjing, Jiangsu 210042, China)

Abstract: Chinese person name recognition is restricted by the domain and size of the existing annotated corpus and the issue of class imbalance. Person name dictionaries and domain dictionaries are more easily achieved than humanly annotated training corpus. This article incorporates dictionaries into bi-directional long short-term memory (Bi-LSTM) networks with weighted conditional random field layer (WCRF). The model extracts the possibility of family name and given name from personal name dictionaries. The domain dictionaries provide information on human names. Bi-LSTM captured context information and weighted conditional random field improved recall of personal name recognition. Experiments on *People's Daily* corpus and construction law corpus show that, compared with the existing method based on hidden Markov model, the F_1 value of personal name recognition is improved by 18.34%; compared with traditional Bi-LSTM-CRF model, Recall value increases by 15.53% and F_1 value increases by 8.83%.

Keywords: person name recognition; bi-directional long short-term memory netwoork; weighted conditional random field; dictionary features

0 引言

中文人名识别是词法分析中未登录词识别的重点和关键[1]。一方面，错误的人名识别影响了相邻词的切分，降低了词法分析的正确率，影响了句法分析以及其他信息处理的质量。另一方面，人名识别属于命名实体识别，在信息抽取中起着重要作用，如安全事件分析[2]、商业信息提取[3]、不良信息过滤[4]等。

现有的人名识别方法大致分为两类：①统计方法，例如条件随机场[5](CRF)、隐马尔可夫模

收稿日期：2019-09-09 **定稿日期**：2019-12-24

基金项目：国家自然科学基金(71601047)；中国博士后科学基金(2015M581706)

型[1,6]、支持向量机[7]等；②基于规则的方法[8]，分析人名的内外部结构、用字等特征，建立构词规则。为了避免繁琐的特征工程，研究者将深度学习方法引入命名实体识别领域[9-11]，充分挖掘文本蕴含的特征。

中国人名识别的难点包括人名构成多样性、人名与上下文组合成词、人名内部相互成词[1]。同时，当测试领域与训练领域不一致时，识别的性能会大幅降低[5]。另外，在训练样本中，人名出现的比例偏低，导致样本类别分布不均衡[12-13]。为了解决以上问题，本文在双向长短期记忆(Bi-LSTM)网络模型中加入人名词典特征，提高人名识别正确率，设计加权 CRF 损失函数，解决样本类别分布不均衡问题，提高人名识别召回率。本文提出的人名识别方法特点体现在：①减少了特征设计和标注的工作量；②无须加入分词或词性等特征，因而不受分词或词性错误的影响；③人名词典特征和加权 CRF 提高了人名识别正确率和召回率。本文提出的加权 CRF 还可以应用到其他类别不均衡的序列标注或分类问题中。

1 相关研究

人名识别可以看作一种文本序列化标注问题，在标注语料上进行监督学习。而深度神经网络能够挖掘文本潜在的抽象信息，有助于提高人名识别的性能。Huang 等[10]提出双向长短期记忆模型(Bi-LSTM)和 CRF 模型相结合的序列标注模型，并应用于命名实体识别任务；Ma 等[11]在 Bi-LSTM-CRF 模型中加入卷积神经网络层(CNN)，提取字母层特征用于命名实体识别。张海楠等[14]针对中文命名实体识别，提出将字特征和词特征结合起来作为深度神经网络的输入。李明扬等[15]在 Bi-LSTM-CRF 模型中加入自注意力机制，提高中文社交媒体中命名实体识别的性能。王蕾等[16]选用 Bi-LSTM 和 semi-CRF 结合的片段神经网络结构学习人名识别的特征，在中文语料上验证算法的性能。

词典特征常被应用于基于统计的人名识别方法中，指示当前字符或者一段字符序列是否出现在词典中[1,6-7,17]。深度神经网络模型也可以利用词典特征提高序列标注性能，如构造词典特征向量[18]，运用词典得到特征词向量[19-21]，以及训练额外的词典特征网络[22-23]。本文借鉴了中文分词中用词典提取成词的特征的方法[18]，从人名词典中抽取特征向量，提高人名识别性能。

He 等[12]发现命名实体识别的深度神经网络模型的召回率(Recall)明显低于准确率(Precision)，为了解决这个问题，在损失函数中加入 F 值。然而，召回率偏低的原因之一是训练样本中非命名实体个数远多于命名实体个数，即样本类别不均衡。Lannoy 等[13]提出加权 CRF 模型，用于解决 CRF 分类器的样本不均衡问题。本文将加权 CRF 模型引入深度神经网络的 CRF 层，并对加权 CRF 模型的损失函数进行改进。

本文的组织结构如下：第 2 节搭建了融合人名词典特征的 Bi-LSTM-WCRF 深度神经网络模型；第 3 节进行实验，说明本文模型在人名识别中的有效性；最后总结并提出下一步工作方向。

2 Bi-LSTM-WCRF 人名识别模型

2.1 基本架构

人名识别被看作是序列化标注问题。输入句子表示为 $\boldsymbol{s}=c_1,c_2,\cdots,c_m$，其中 c_i 表示第 i 个字符(包括汉字、数字、字母或标点符号等)。输出人名标注序列为 $\boldsymbol{y}=y_1,y_2,\cdots,y_m$，其中 $y_i\in\{\text{B},\text{M},\text{E},\text{S},\text{O}\}$是 c_i 的标签，B、M、E、S 和 O 分别代表人名首字，人名中间字，人名结尾字，单字表示的人名和其他。人名识别就是对每个字符进行 B、M、E、S、O 的分类标注。

人名识别的深度神经网络模型基本结构分为三层：输入层、Bi-LSTM 层和 CRF 层，如图 1 所示。

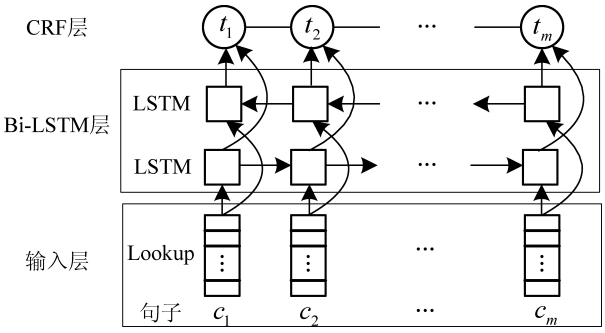

图 1 人名识别深度神经网络基本结构

2.1.1 输入层

输入层将句子中的字符 c_i 转换成向量形式 $\boldsymbol{e}_i$，输入 Bi-LSTM 层。本文采取随机方式，生成 d 维字符向量，形成 $d\times N$ 的字符矩阵 $\boldsymbol{W}_e$，其中 N 表示

训练语料库中有效的字符个数。$\boldsymbol{e}_i$ 如式(1)所示。

$$\boldsymbol{e}_i = \boldsymbol{W}_e(c_i) \tag{1}$$

2.1.2　Bi-LSTM 层

一个 LSTM 单元由输入门、遗忘门、输出门和细胞状态构成。$\boldsymbol{e}^t \in R^d$ 是时刻 t 的输入向量，$\boldsymbol{h}^{t-1} \in R^K$ 是 LSTM 单元的时刻 t-1 的输出，$\boldsymbol{c}^{t-1} \in R^K$ 是时刻 $t-1$ 的细胞状态。时刻 t 的 LSTM 的工作流程可以表示为式(2)～式(7)：

$$\boldsymbol{i}^t = \sigma(\boldsymbol{W}_i\boldsymbol{h}^{t-1} + \boldsymbol{U}_i\boldsymbol{e}^t + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}^t = \sigma(\boldsymbol{W}_f\boldsymbol{h}^{t-1} + \boldsymbol{U}_f\boldsymbol{e}^t + \boldsymbol{b}_f) \tag{3}$$

$$\boldsymbol{o}^t = \sigma(\boldsymbol{W}_o\boldsymbol{h}^{t-1} + \boldsymbol{U}_o\boldsymbol{e}^t + \boldsymbol{b}_o) \tag{4}$$

$$\tilde{\boldsymbol{c}}^t = \tanh(\boldsymbol{W}_{\tilde{c}}\boldsymbol{h}^{t-1} + \boldsymbol{U}_{\tilde{c}}\boldsymbol{e}^t + \boldsymbol{b}_{\tilde{c}}) \tag{5}$$

$$\boldsymbol{c}^t = \boldsymbol{f}^t \odot \boldsymbol{c}^{t-1} + \boldsymbol{i}^t \odot \tilde{\boldsymbol{c}}^t \tag{6}$$

$$\boldsymbol{h}^t = \boldsymbol{o}^t \odot \tanh(\boldsymbol{c}^t) \tag{7}$$

其中，σ 函数和 tanh 函数是对向量中每个元素做相应运算，具体表示为 $\sigma(x) = (1+e^{-x})^{-1}$，$\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$，$\odot$ 表示每个元素对应相乘。$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$、$\boldsymbol{W}_{\tilde{c}}$ 是 $\boldsymbol{h}^{t-1}$ 的权重矩阵；$\boldsymbol{U}_i$、$\boldsymbol{U}_f$、$\boldsymbol{U}_o$、$\boldsymbol{U}_{\tilde{c}}$ 是 $\boldsymbol{e}^t$ 的权重矩阵；$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_{\tilde{c}}$ 是偏置向量；$\boldsymbol{i}^t$、$\boldsymbol{f}^t$、$\boldsymbol{o}^t$、$\boldsymbol{c}^t$ 分别代表输入门、遗忘门、输出门和细胞状态。

Bi-LSTM 中的输出 $\boldsymbol{h}^t$ 可以表示为：$\boldsymbol{h}^t = \overrightarrow{\boldsymbol{h}}^t \oplus \overleftarrow{\boldsymbol{h}}^t$，$\overrightarrow{\boldsymbol{h}}^t$ 和 $\overleftarrow{\boldsymbol{h}}^t$ 分别是时刻 t 的前向输出和后向输出，$\oplus$表示向量拼接操作。

2.1.3　CRF 层

在人名识别的训练数据中，句子 $\boldsymbol{s} = c_1, c_2, \cdots c_m$ 对应的正确标注序列 $\boldsymbol{y} = y_1, y_2, \cdots, y_m$，$y_i \in \{\mathrm{B}, \mathrm{M}, \mathrm{E}, \mathrm{S}, \mathrm{O}\}$，经过 CRF 层后得到预测结果 $\hat{\boldsymbol{y}} = \hat{y}_1, \hat{y}_2, \cdots, \hat{y}_m$，$\hat{y}_i \in \{\mathrm{B}, \mathrm{M}, \mathrm{E}, \mathrm{S}, \mathrm{O}\}$。定义 $\boldsymbol{P}_t \in R^5$ 为 c_t 在{B，M，E，S，O}上的得分，有 $\boldsymbol{P}_t = \boldsymbol{W}_s\boldsymbol{h}^t + \boldsymbol{b}_s$，$\boldsymbol{W}_s \in R^{5\times 2K}$ 和 $\boldsymbol{b}_s \in R^5$ 分别是可训练的参数。$\boldsymbol{P}_{t,\hat{y}_t}$ 表示 c_t 在预测标注 $\hat{y}_t$ 上的得分。定义转移得分矩阵 $\boldsymbol{A} \in R^{5\times 5}$，$A_{i,j}$ ($i, j \in \{\mathrm{B}, \mathrm{M}, \mathrm{E}, \mathrm{S}, \mathrm{O}\}$) 表示从前一时刻预测的标注结果 i 转移到下一时刻标注结果 j 的得分。给定 $\boldsymbol{s}$ 和 $\hat{\boldsymbol{y}}$，定义预测得分如式(8)所示。

$$\mathrm{score}(\boldsymbol{s}, \hat{\boldsymbol{y}}) = \sum\nolimits_t (\boldsymbol{A}_{\hat{y}_{t-1},\hat{y}_t} + \boldsymbol{P}_{t,\hat{y}_t}) \tag{8}$$

已知 $\boldsymbol{s}$，$\hat{\boldsymbol{y}}$ 的条件概率 $P(\hat{\boldsymbol{y}} \mid \boldsymbol{s})$ 如式(9)所示。

$$P(\hat{\boldsymbol{y}} \mid \boldsymbol{s}) = \frac{e^{\mathrm{score}(\boldsymbol{s}, \hat{\boldsymbol{y}})}}{\sum_{\tilde{y} \in \boldsymbol{Y}_s} e^{\mathrm{score}(\boldsymbol{s}, \tilde{\boldsymbol{y}})}} \tag{9}$$

其中，$\boldsymbol{Y}_s$ 表示句子 $\boldsymbol{s}$ 的所有可能的标注序列集合。

训练时，最大化正确标注序列的 $\ln P(\boldsymbol{y} \mid \boldsymbol{s})$，预测时，预测结果为得分最高的序列，如式(10)所示。

$$\hat{\boldsymbol{y}}^* = \underset{\hat{y} \in Y_s}{\mathrm{argmax}}\ \mathrm{score}(\boldsymbol{s}, \hat{\boldsymbol{y}}) \tag{10}$$

2.2　融入人名词典特征

2.2.1　人名词典特征向量构建

中文人名的用字具有一定的规律性，有助于提高人名识别的性能[7]。参考中文分词模型中加入词典特征的做法[18]，在中文人名识别的深度神经网络模型中加入人名词典特征。由于中文人名形式多变，用字灵活，因此，并未采取直接匹配人名的方式，而是拆分为姓和名分别考虑。

选用“中文人名语料库①”中“中文常见人名”作为人名词典，包含人名 120 万。对人名词典进行分类统计，分为复姓、单姓、单字名和双字名。考虑句子 $\boldsymbol{s} = c_1, c_2, \cdots, c_m$ 中的第 i 个字符，定义词典向量 $\boldsymbol{f}_i = [f_{i1}, f_{i2}, \cdots, f_{i6}]$，$f_{ij} \in \{0, 1\}$，$j \in \{1, 2, \cdots, 6\}$，每个元素分别代表当前字符是否是复姓第一个字、是否是复姓第二个字、是否是单姓、是否是单字名、是否是双字名首字、是否是双字名第二个字。设置长度为 $l = 5$（l 的值可调）的窗口，提取 c_i 的上下文字符序列 $(c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2})$，将字符序列中字符的词典特征向量拼接起来，作为最终 c_i 的词典特征向量 $\boldsymbol{F}_i$。文献[18]中的词典特征与本文词典特征的性能比较见 3.3 节。

2.2.2　带特征向量的深度神经网络模型

给定句子 $\boldsymbol{s}$，字符 c_i 有字符向量 $\boldsymbol{e}_i$ 和词典特征向量 $\boldsymbol{F}_i$。基本的 Bi-LSTM-CRF 模型以字符向量作为输入，为了加入词典特征，一种常用方式是将词典特征向量作为另一个 Bi-LSTM-CRF 模型的输入，最后将这两个并行的 Bi-LSTM-CRF 模型的输出拼接起来[18]。然而，词典特征向量中存有潜在的人名边界信息，可能影响当前字符的 LSTM 模型的输出。由于在并行拼接方式中，字符的 LSTM 模型存在权重共享的约束，因此，为了在字符 LSTM 中更好地加入词典特征信息，采用 Zhang 等[18]提出的超网络（Hyper-network）结构深度神经网络。超网络结构包括主 LSTM（MainLSTM）层和超 LSTM（HyperLSTM）层，如图 2 所示，HyperLSTM 的隐藏层状态影响了 MainLSTM 的权重。并行拼接和超网络结构对人名识别性能的影响见 3.3 节。

HyperLSTM 层由 Bi-LSTM 网络构成，输入词典特征向量 $\boldsymbol{F}_i$，输出 $\boldsymbol{h}_i^{(t)}$ 被送入 MainLSTM 层。MainLSTM 层的门 $g \in \{i, \tilde{c}, f, o\}$ 的输入表示为：

① https://github.com/wainshine/Chinese-Names-Corpus

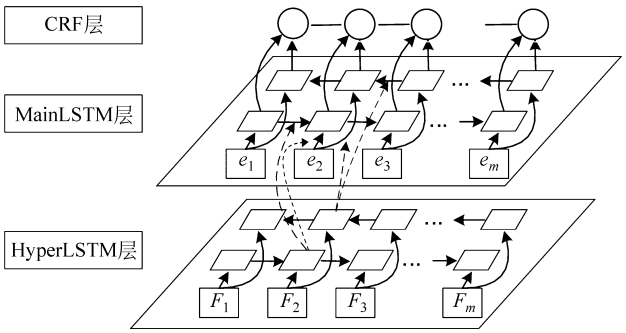

图2 超网络结构人名识别神经网络模型图

$$\boldsymbol{g}_i^{(M)} = \boldsymbol{d}_{gh} \odot \boldsymbol{W}_g^{(M)} \boldsymbol{h}_{i-1}^{(M)} + \boldsymbol{d}_{gx} \odot \boldsymbol{U}_g^{(M)} \boldsymbol{e}_i^{(M)} + \boldsymbol{b}_g \quad (11)$$

$$\boldsymbol{d}_{gh} = \boldsymbol{W}_{gh}^{(M)} \boldsymbol{z}_{gh}^{(t)} = \boldsymbol{W}_{gh}^{(M)} (\boldsymbol{W}_{gh}^{(t)} h_i^{(t)} + \boldsymbol{b}_{gh}^{(t)}) \quad (12)$$

$$\boldsymbol{d}_{gx} = \boldsymbol{W}_{gx}^{(M)} \boldsymbol{z}_{gx}^{(t)} = \boldsymbol{W}_{gx}^{(M)} (\boldsymbol{W}_{gx}^{(t)} \boldsymbol{h}_i^{(t)} + \boldsymbol{b}_{gx}^{(t)}) \quad (13)$$

$$\boldsymbol{b}_g = \boldsymbol{W}_{gb}^{(M)} \boldsymbol{z}_{gb}^{(t)} = \boldsymbol{W}_{gb}^{(M)} (\boldsymbol{W}_{gb}^{(t)} \boldsymbol{h}_i^{(t)} + \boldsymbol{b}_{gb}^{(t)}) \quad (14)$$

式(12)表示HyperLSTM的隐藏层输出 $\boldsymbol{h}_i^{(t)}$ 经过线性变换 $\boldsymbol{W}_{gh}^{(t)} h_i^{(t)} + \boldsymbol{b}_{gh}^{(t)}$，再乘以权重矩阵 $\boldsymbol{W}_{gh}^{(M)}$，得到 $\boldsymbol{d}_{gh}$，与MainLSTM的 $\boldsymbol{h}_{i-1}^{(M)}$ 一起，影响MainLSTM的门的状态计算，见式(11)。式(13)、(14)与(12)计算类似。具体过程参见文献[18]。

2.3 加权CRF层

在人名识别的训练样本中，人名与非人名相比，数目明显偏少，即样本类别不均衡。这导致训练时损失函数会偏向非人名的一方，人名识别召回率偏低。解决方法之一是对损失函数中较少类别的分类错误导致的损失分配较大的权重，较多类别的分类错误导致的损失分配较小的权重。

深度学习模型的优化目标函数可表示为式(15)：

$$\min \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i;\theta)) + \lambda \varphi(\theta) \quad (15)$$

其中，N 是训练样本个数，$L(y_i, f(x_i;\theta))$ 是第 i 个训练样本的损失函数值，反映正确结果 y_i 与预测结果 $f(x_i;\theta)$ 的差距。假设正确结果共有 K 类，则有 $y_i \in \{c_1, c_2, \cdots, c_K\}$，将式(15)中的损失函数分解到不同类别上，如式(16)所示。

$$\min \frac{1}{N} \sum_{k=1}^{K} \sum_{\{i|y_i=c_k\}} L(y_i, f(x_i;\theta)) + \lambda \varphi(\theta) \quad (16)$$

给不同类别的损失函数值分配不同的权重 w_k，有式(17)：

$$\min \frac{1}{N} \sum_{k=1}^{K} w_k \sum_{\{i|y_i=c_k\}} L(y_i, f(x_i;\theta)) + \lambda \varphi(\theta) \quad (17)$$

然而，从式(9)可以看出，CRF层的损失函数的分母是对所有可能标注序列求和，因此，无法直接写成式(15)的形式，也无法如式(17)计算加权损失函数。本文通过对CRF层损失函数的近似计算，将近似值分解为式(15)形式，进而分配类别权重，减少样本不均衡的影响。

命题1 深度神经网络训练时，CRF层损失函数满足以下不等式：

$$-\ln P(\boldsymbol{y} \mid \boldsymbol{s}) < -\ln \frac{\exp\left(\sum_t (\boldsymbol{A}_{y_{t-1}, y_t} + \boldsymbol{P}_{t, y_t})\right)}{\left(\sum_{y_1} \exp(\boldsymbol{P}_{1, y_1})\right) \prod_{t=2} \left(\sum_{y_{t-1}, y_t} \exp(\boldsymbol{A}_{y_{t-1}, y_t} + \boldsymbol{P}_{t, y_t})\right)} \quad (18)$$

证明：训练时，CRF层目标函数为：

$$\max \ln P(\boldsymbol{y} \mid \boldsymbol{s})$$

等价于损失函数 $-\ln P(\boldsymbol{y}|\boldsymbol{s})$，带入式(8)和式(9)，有：

$$-\ln P(\boldsymbol{y} \mid \boldsymbol{s}) = -\ln \frac{\exp\left(\sum_t (\boldsymbol{A}_{y_{t-1}, y_t} + \boldsymbol{P}_{t, y_t})\right)}{\sum_{\tilde{y} \in \boldsymbol{Y}_x} \exp\left(\sum_t (\boldsymbol{A}_{\tilde{y}_{t-1}, \tilde{y}_t} + \boldsymbol{P}_{t, \tilde{y}_t})\right)} \quad (19)$$

令B=0，M=1，E=2，S=3，O=4，设行向量 $\boldsymbol{\alpha}_1$ 和矩阵 $\boldsymbol{M}_t$，$t = 1, \cdots, m$，有：$[\boldsymbol{\alpha}_1]_i = e^{\boldsymbol{P}_{1, y_1 = i}}$ 和 $[\boldsymbol{M}_t]_{i,j} = e^{\boldsymbol{P}_{t, y_t = j} + \boldsymbol{A}_{y_{t-1} = i, y_t = j}}$，其中，$i, j \in \{0, 1, 2, 3, 4\}$，$[\cdot]_i$ 代表向量的第 i 个元素，$[\cdot]_{i,j}$ 代表矩阵第 i 行第 j 列元素。

式(19)中的分母记作 Z，根据CRF前向算法[24]，Z 可表示为：

$$Z = \sum_{i=0}^{4} [\boldsymbol{\alpha}_1 \boldsymbol{M}_2 \boldsymbol{M}_3 \cdots \boldsymbol{M}_m]_i \quad (20)$$

将 $[\boldsymbol{\alpha}_1]_i$ 和 $[\boldsymbol{M}_t]_{i,j}$ 带入式(20)，可得：

$$Z = \sum_{y_1, y_2, \cdots, y_m} \exp(\boldsymbol{P}_{1, y_1} + \boldsymbol{A}_{y_1, y_2} + \boldsymbol{P}_{2, y_2} + \cdots + \boldsymbol{A}_{y_{m-1}, y_m} + \boldsymbol{P}_{m, y_m})$$

式(18)中的分母记作 Z'，进一步计算可得：

$$\begin{aligned} Z' &= \left(\sum_{y_1} \exp(\boldsymbol{P}_{1, y_1})\right) \left(\sum_{y_1, y_2} \exp(\boldsymbol{A}_{y_1, y_2} + \boldsymbol{P}_{2, y_2})\right) \cdots \left(\sum_{y_{m-1}, y_m} \exp(\boldsymbol{A}_{y_{m-1}, y_m} + \boldsymbol{P}_{m, y_m})\right) \\ &> \sum_{y_1, y_2, \cdots, y_m} \exp(\boldsymbol{P}_{1, y_1} + \boldsymbol{A}_{y_1, y_2} + \boldsymbol{P}_{2, y_2} + \cdots + \boldsymbol{A}_{y_{m-1}, y_m} + \boldsymbol{P}_{m, y_m}) \\ &= Z \end{aligned} \quad (21)$$

因此，

$$-\ln P(\boldsymbol{y} \mid \boldsymbol{s}) = -\ln \frac{\exp\left(\sum_t (\boldsymbol{A}_{y_{t-1}, y_t} + \boldsymbol{P}_{t, y_t})\right)}{Z} < -\ln \frac{\exp\left(\sum_t (\boldsymbol{A}_{y_{t-1}, y_t} + \boldsymbol{P}_{t, y_t})\right)}{Z'}$$

$$=-\ln\frac{\exp\left(\sum_t\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)\right)}{\left(\sum_{y_1}\exp\left(\boldsymbol{P}_{1,y_1}\right)\right)\prod_{t=2}\left(\sum_{y_{t-1},y_t}\exp\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)\right)}$$

证明完毕。

以式(18)中$-\log P(\boldsymbol{y}\mid\boldsymbol{s})$的上限为损失函数，进一步分解可得：

$$-\ln\frac{\exp\left(\sum_t\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)\right)}{\left(\sum_{y_1}\exp\left(\boldsymbol{P}_{1,y_1}\right)\right)\prod_{t=2}\left(\sum_{y_{t-1},y_t}\exp\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)\right)}$$

$$=-\left(\ln\frac{\exp\boldsymbol{P}_{1,y_1}}{\left(\sum_{y_1}\exp\left(\boldsymbol{P}_{1,y_1}\right)\right)}+\sum_{t=2}\ln\frac{\exp\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)}{\sum_{y_{t-1},y_t}\exp\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)}\right)$$

假设人名 B、M、E、S 的权重为 w_P，非人名的 O 权重为 w_O，加权 CRF 损失函数表示为：

$$\text{Loss\_}w=-w_1\ln\frac{\exp\boldsymbol{P}_{1,y_1}}{\left(\sum_{y_1}\exp\left(\boldsymbol{P}_{1,y_1}\right)\right)}-\sum_{t=2}w_t\ln\frac{\exp\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)}{\sum_{y_{t-1},y_t}\exp\left(\boldsymbol{A}_{y_{t-1},y_t}+\boldsymbol{P}_{t,y_t}\right)}\tag{22}$$

其中，$w_t\in\{w_O,w_P\}$，$t=1,2,\cdots,m$。

3　实验

3.1　实验环境与实验数据

本文的实验环境如下：处理器：Inter (R) Core (TM) i7-8700K CPU @ 3.70GHz；GPU：NVIDIA GeForce RTX 2080 Ti 图形加速卡：NVIDIA GeForce RTX 2080 Ti 11 GB；内存：16GB；操作系统：Windows 10 (64bit)；Google 开源深度学习框架 TensorFlow-GPU1.3。

训练语料由两部分组成，一部分是北京大学计算语言学研究所发布的 1998 年 1 月《人民日报》语料[25]，大小为 8MB，另一部分来自工程法律领域，取自中国裁判文书网①的建设工程施工合同纠纷文书，大小为 0.3MB。选取 90%作为训练集，剩余作为验证集。从训练语料之外选取了工程法律领域中 200 个含有人名的句子进行开放测试，这 200 个句子中的人名未在训练集包含的专业领域句子中出现过。

由于训练句子中非人名标签 O 数目过多，除了加权 CRF 方法以外，还采取了另外两个步骤：①将非人名标签 O 按照分词标注进一步细分为 O_B、O_M、O_E、O_S；②确保每个 batch 中有人名的句子和没有人名的句子个数相同。尽管对 O 进行了细分，然而，深度神经网络的输入与分词无关，因而不受分词错误的影响。

人名识别的性能评估采用准确率 P、召回率 R 和综合指标 F_1。P＝正确识别出人名数/所识别出人名总数×100%，R＝正确识别出人名数/文本中人名总数×100%，$F_1=2\times P\times R/(P+R)\times 100\%$。

3.2　参数设置

本节对参数的选择进行实验分析，对比不同学习率和不同$\{w_P,w_O\}$的情况下人名识别性能变化情况。其他的参数设置如表 1。

表 1　人名识别深度神经网络超参数

参数名	参数值
字符向量长度	100
batch size	128
Dropout(丢弃率)	0.2
L_2 正则项系数	0.000 1
Main LSTM 隐藏层维数	64
HyperLSTM 隐藏层维数	128
$z_{gh}^{(t)}$ 维数	16

学习率(lr)分别取 0.01、0.001 和 0.000 1，相应的验证集的召回率和准确率如图 3 所示。从图中可以看出，学习率为 0.000 1 时，学习率过小，收敛速度太慢；与学习率 0.001 相比，学习率为 0.01 时，神经网络训练收敛较快，但召回率上的性能低于前者，准确率上性能在 6 次迭代后两者比较相近。因此，学习率设置为 0.001。

非人名标签权重和人名标签权重(w_O,w_P)分别设置为(1,1)(1,5)(1,10)，相应的召回率和准确率如图 4 所示。从图中可见：$w_O=1,w_P=1$ 时，准确率的性能最好，召回率的最差；$w_O=1,w_P=10$ 时，召回率的性能最好，准确率的最差；$w_O=1,w_P=5$ 时，召回率上的性能接近 $w_O=1,w_P=10$，准确率上的性能接近 $w_O=1,w_P=1$。因此，设置 $w_O=1,w_P=5$。

① http://wenshu.court.gov.cn/

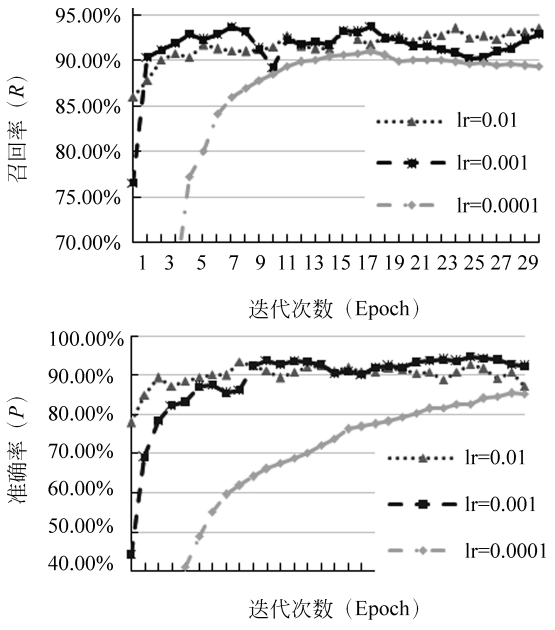

图3 学习率(lr)对性能的影响

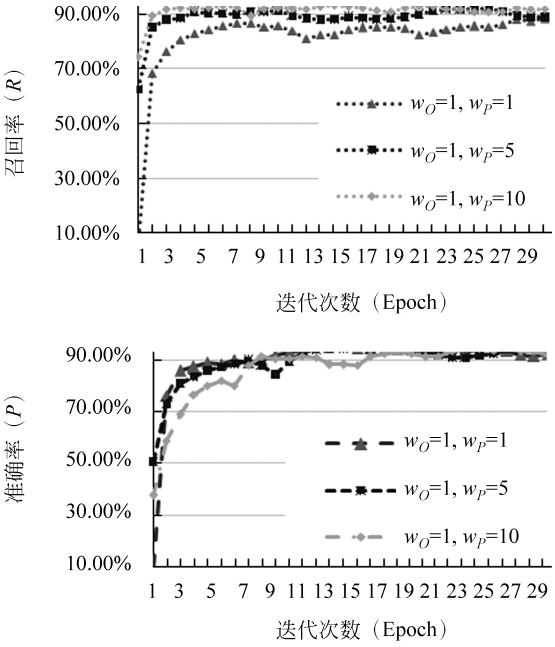

图4 非人名标签权重和人名标签权重(w_O,w_P)

3.3 不同方法性能对比

通过实验对比分析超参数对人名识别模型的影响,最终超参数的设置如下:学习率lr为0.001,标签权重$w_O=1$,$w_P=5$。在工程法律领域文本选取200个句子组成测试集,对比本文提出的模型与其他模型在测试集上的性能。表2列出了本文模型、NLPIR人名识别方法①(隐马尔可夫模型)、传统Bi-LSTM-CRF模型、本文模型(不带人名词典特征,加权CRF层)、本文模型(带人名词典特征,普通CRF层)、本文模型(带人名词典特征,文献[13]中的CRF层)、并行拼接人名词典特征和本文加权CRF、基于BERT的人名识别模型在测试集上的准确率P、召回率R和综合指标F_1。对比表2中的人名识别结果,可以得到如下结论。

表2 工程法律领域不同方法人名识别结果对比(%)

方法	召回率	准确率	F_1
模型1:本文模型	90.29	93.47	91.85
模型2:NLPIR人名识别方法	78.58	69.05	73.51
模型3:传统Bi-LSTM-CRF模型	74.76	93.33	83.02
模型4:本文模型(不带人名词典特征,加权CRF层)	88.35	87.92	88.13
模型5:本文模型(带人名词典特征,普通CRF层)	82.52	96.59	89.00
模型6:本文模型(带人名词典特征,文献[13]中的加权CRF)	89.32	85.98	87.62
模型7:并行拼接人名词典特征,本文加权CRF层	89.81	91.13	90.47
模型8:BERT+全连接层+softmax	92.72	95.98	94.32

(1) 本文模型(模型1)与NLPIR相比,召回率、准确率和F_1分别提高了11.71%、24.42%和18.34%,其中部分原因是本文的训练语料中加入了测试语料同领域的句子。然而,本文模型在特征提取方面更加简单,比如不需要像NLPIR中那样设计人名构成角色。并且,测试语料中的人名在训练语料中同领域句子里未出现,说明本文模型具有一定的自适应性。

(2) 本文模型(模型1)与传统Bi-LSTM-CRF模型相比性能有显著提高,召回率提高了15.53%,准确率提高0.14%,F_1值提高了8.83%。说明本文提出的人名词典特征和加权CRF对基于深度神经网络的人名识别性能有明显提升。

(3) 模型4比模型1少了人名词典和HyperLSTM层,召回率下降1.94%,准确率下降5.55%,说明人名词典特征对人名识别准确率有较明显帮助,同时也会影响召回率。同理可见,模型5比模型3多了人名词典和HyperLSTM层,准确率增加3.26%,召回率增加7.76%。

(4) 模型5比模型1少了加权CRF,召回率下

① http://ictclas.nlpir.org/

降 7.77%，准确率提高 3.12%，说明增加人名标签权重，牺牲了部分准确率，提高了召回率。同理可见，模型 4 比模型 3 多了人名标签权重，召回率增加 13.59%，准确率下降 5.41%，也是同样道理。

(5) 本文模型与模型 6 相比，性能有一定提升，召回率、准确率和 F_1 分别提高 0.97%、7.49% 和 4.23%。说明本文提出的加权 CRF 方法比文献[13]的方法在准确率上有改进。

(6) 模型 1 与模型 7 相比，超 LSTM 网络结构在准确率上，比并行拼接 LSTM 结构提高 2.34%。

(7) 对比本文模型与基于 BERT[26]的人名识别模型的性能，BERT 模型配置如下：模型包括 BERT 层、全连接层和 softmax 层，载入中文预训练模型参数，学习率设置为 5×10^{-5}，batch size 为 8，训练 10 个 epoch。由实验结果可知，BERT 模型性能优于本文模型。这主要因为 BERT 预训练模型很好地学习到语义信息特征以及句法短语信息，特征抽取能力较强。然而，随着 BERT 模型最大句子长度的增加，容易发生 GPU 或 CPU 内存不足。

在《人民日报》数据集上，将本文模型与加入词典特征的其他模型进行了对比，结果见表 3。本次实验中，不对训练集的句子按人名做样本平衡的预处理，因而调整标签权重。本文模型的参数设置如下：学习率 lr 为 0.001，标签权重 $w_O=1$，$w_P=10$，单层双向 HyperLSTM，初始字向量为 100 维预训练字向量①，其他参数见表 1。从表 3 中可见，本文模型提高了人名识别的召回率，准确率有所下降。

表 3　人民日报语料不同方法人名识别结果对比(%)

方法	召回率	准确率	F_1
本文模型	92.53	95.10	93.80
石等(2019)[20]	91.83	97.35	94.50
冯等(2018)[19]	89.49	98.23	93.66

4　结束语

本文构造了基于字符级别的深度神经网络模型(Bi-LSTM-WCRF)，用于识别中文人名。在传统 Bi-LSTM-CRF 模型基础上，增加 HyperLSTM 层，融合人名词典特征向量，改进 CRF 层，增加人名标签和非人名标签权重设置。在《人民日报》和工程法律领域的语料上分别进行实验，结果表明，Bi-LSTM-WCRF 在保证准确率的前提下，明显提高了召回率，改善了人名识别的类别不均衡问题。同时，Bi-LSTM-WCRF 无须输入分词或词性等特征，不受分词或词性错误的影响。

目前，领域训练语料被直接加入《人民日报》训练语料中，未考虑两种语料之间的差异，下一步研究将探索语料的不平衡对人名识别的影响，以及解决的方法；在实验中发现不同参数下，有的人名识别模型准确率很高，召回率很低，有的模型召回率很高，准确率很低，如何将这些模型结合起来也是下一步的研究方向；本文针对人名设计了词典特征，其他命名实体识别任务性能是否可以利用类似词典特征进一步提升也是后续研究方向。

参考文献

[1] 张华平，刘群.基于角色标注的中国人名自动识别研究[J]. 计算机学报，2004，27(01)：85-91.

[2] 徐飞，宋英华.海量食品安全事件下的命名实体识别研究[J].科研管理，2018，39(07)：131-138.

[3] 张悦，潘淑文，刘秀磊.人名识别技术在中国招中标领域的应用[J].北京信息科技大学学报(自然科学版)，2017，32(05)：72-76，83.

[4] Duan J，Zeng J. Web objectionable text content detection using topic modeling Technique[J]. Expert Systems with Applications，2013，40(15)：6094-6104.

[5] Chen W，Zhang Y，Isahara H. Chinese named entity recognition with conditional random fields [C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing，Sydney，Australia：Association for Computational Linguistics，2006：118-121.

[6] 俞鸿魁，张华平，刘群，等.基于层叠隐马尔可夫模型的中文命名实体识别[J].通信学报，2006，27(02)：87-94.

[7] 李丽双，黄德根，毛婷婷，等.基于支持向量机的中国人名的自动识别[J].计算机工程，2006，32(19)：188-190，201.

[8] 周昆，胡学钢.一种基于本体论和规则匹配的中文人名识别方法[J].微计算机信息，2010，26(31)：87-89.

[9] Peng N Y，Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon，Portugal：Association for Computational Linguistics，2015：548-554.

[10] Huang Z，Xu W，Yu K. Bidirectional LSTM-CRF models for sequence tagging [J]. arXiv preprint arXiv：1508.01991，2015.

① https://github.com/fudannlp16

[11] Ma X, Hovy E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 1064-1074.

[12] He H, Sun X. F-score driven max margin neural network for named entity recognition in Chinese social media [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 2017: 713-718.

[13] Lannoy G de, Francois D, Delbeke J, et al. Weighted conditional random fields for supervised interpatient heartbeat classification [J]. IEEE Transactions on Biomedical Engineering, 2012, 59(1): 241-247.

[14] 张海楠, 伍大勇, 刘悦,等. 基于深度神经网络的中文命名实体识别 [J]. 中文信息学报, 2017, 31(4): 28-15.

[15] 李明扬,孔芳. 融入自注意力机制的社交媒体命名实体识别 [J/OL]. 清华大学学报(自然科学版): 2019, 59(6): 461-467.

[16] 王蕾,谢云,周俊生,等.基于神经网络的片段级中文命名实体识别[J].中文信息学报,2018,32(03): 84-90,100.

[17] Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition [C]//Proceedings of the 13th Conference on Computational Natural Language Learning, CoNLL'09, PA, USA, 2009: 147-155.

[18] Zhang Q, Liu X Y, Fu J L. Neural networks incorporating dictionaries for Chinese word segmentation [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 5682-5689.

[19] 冯艳红,于红,孙庚,等.基于 BiLSTM 的命名实体识别方法[J].计算机科学,2018,45(02): 261-268.

[20] 石春丹,秦岭.基于 BGRU-CRF 的中文命名实体识别方法[J].计算机科学,2019,46(09): 237-242.

[21] 冀相冰,朱艳辉,徐啸,等.基于注意力机制的包装命名实体识别[J].包装工程,2019,40(15): 24-29.

[22] Liu T, Yao J-G, Lin C-Y. Towards improving neural named entity recognition with gazetteers [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 5301-5307.

[23] Lin H Y, Lu Y J, Han X P, et al. GEANN: Gazetteer-enhanced attentive neural networks for named entity recognition [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing.

[24] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012.

[25] Yu S W, Duan H M, Wu Y F. Corpus of multi-level processing for modern Chinese [DB/OL]. [2019-09-09]. https://doi.org/10.18170/DVN/SEYRX5.

[26] Devlinj, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minnesota, USA, 2019: 4171-4186.

成于思(1983—),博士,讲师,主要研究领域为文本挖掘与工程法律。
E-mail: xchengyusi@163.com

施云涛(1987—),硕士,高级工程师,主要研究领域为自然语言处理。
E-mail: shiyuntao@suning.com