一种中文人名识别的训练架构

王嘉文[1],王传栋[1],杨雁莹[2]
(1. 南京邮电大学 计算机学院,江苏 南京 210023;
2. 南京森林警察学院,江苏 南京 210023)

摘 要:中文人名识别作为中文语言处理的一项关键技术,广泛应用于文本挖掘、语义分析、机器翻译等领域。在数据日趋海量化和异构化的当今社会,对于中文人名进行命名实体识别已经成为现阶段中文自然语言处理的研究热点之一。由于现有方法大多依赖于先验的领域知识和工程化的特征,识别模型常需要研究人员的大量语言学知识。为了减少甚至忽略对这些工程化的特征的依赖,旨在建立一种较为灵活的深度神经网络架构,通过对大规模未标记语料的内部表示的学习,使得系统减少甚至忽略这些工程化特征的影响,采用无监督的方法进行中文人名识别。实验结果表明,该模型不但性能良好,而且不需要过多的计算资源,在中文人名识别的应用中具有良好的效果。

关键词:自然语言处理;深度学习;神经网络;中文人名识别
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2018)07-0053-05
doi:10.3969/j.issn.1673-629X.2018.07.012

A Training Framework for Chinese Name Recognition

WANG Jia-wen[1], WANG Chuan-dong[1], YANG Yan-ying[2]
(1. School of Computer and Software, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. Nanjing Forest Police College, Nanjing 210023, China)

Abstract: Chinese name recognition, as a key technology in Chinese language processing, is widely used in text mining, semantic analysis, machine translation and other fields. The data are becoming massive and heterogeneous in today's society, so the named entity recognition for Chinese names has become one of the hotspots of Chinese natural language processing at this stage. Identification model often requires a large number of linguistic knowledge of the researchers because most of the existing methods rely on transcendental domain knowledge and engineering characteristics. In order to reduce or even ignore the dependence on these engineering features, we aim to establish a more flexible deep neural network architecture which can be through the large-scale unmarked corpus of the internal representation of learning, making the system reduce or even ignore the impact of these engineering features and using the unsupervised method for Chinese name recognition. Experiment shows that the model not only has excellent performance but also does not need too much computing resources, with good effect in the Chinese name recognition application.

Key words: natural language processing; deep learning; neural networks; Chinese name recognition

0 引 言

当前大数据环境下,海量化数据的结构日趋呈现多样化和异构化。然而在许多问题中,数据事实因应用需求的不同,其数据结构的组织形式存在多方差异。为了满足不同数据挖掘任务对数据事实的统一理解,需要建立一种数据的统一内部表示。

在自然语言处理(natural language processing, NLP)[1]任务中,许多高度工程化的 NLP 系统应用,大都采取基于特定任务特征的线性统计模型,这些模型由应用背景激发、受领域知识的限制,通过面向工程的专用特征发现数据表示。这些特征通常由一些特征提取的辅助工具预处理而得到,是一种监督学习的训练方法。但是这种方法不仅会导致复杂的运行时依赖关系,而且要求研发人员必须拥有大量的语言学知识。

收稿日期:2017-06-13 **修回日期**:2017-10-18 **网络出版时间**:2018-02-24
基金项目:中央高校基本科研业务费专项资金项目(LGZD201502,LGYB201603)
作者简介:王嘉文(1993-),男,硕士研究生,研究方向为文本挖掘与自然语言处理;王传栋,副教授,硕导,研究方向为文本挖掘与自然语言处理、机器学习。

为了减少这种依赖,必须捕捉关于自然语言的更多的一般性,分析语言的元信息,如词性、实体、语法、句法等,以期获取一种更一般的描述方法,减少甚至忽略先验领域知识对模型的影响,用无监督学习的方式,尽量避免工程化特征,在大规模未标记数据上学习产生模型。

文中描述了一种基于深度神经网络的字词训练模型,通过发现其内部表示,尽量避免了工程特征对于模型的限制,采取一种无监督学习方式对中文人名识别进行了研究,最后通过实验验证该模型的合理性。

1 研究现状

20 世纪 90 年代初,国外就已经开始了对命名实体识别的研究。最早采用的大多是基于规则的方法,专家和学者在某些特定领域对于相关文本进行总结和归纳,提取一种易于理解和表达的规则进行命名实体识别,如 GATE 项目中的 ANNIE 系统和曾经参加过 MUC 评测的 FACILE 系统等,并在小规模文本上取得了较好的效果。但是这种方法需要极富经验的领域专家进行人工干预,会耗费大量的人工成本,在面对大数据量的命名实体识别任务时效率较为低下。

随着机器学习理论的发展以及计算机性能的提升,基于统计的方法开始不再受限于计算量不足而导致的识别率低的问题,逐渐被广大的专家和学者所接受和青睐。自 1997 年开始,国外的专家依次将隐马尔可夫[2]、支持向量机[3]、条件随机场[4]等模型应用到英文命名实体识别任务中,都取得了较好的效果。

因为中文的特殊性,中文命名实体识别拥有比英文更高的难度,国内专家和学者在借鉴国外研究成果的基础上进行了长期的研究。张华平等引入隐马尔可夫模型[5],根据人工制定的角色编码使用 Viterbi 算法在分词结果上标注人名构成角色,根据标注的角色序列进行最长模式串匹配,对中文人名识别进行了研究,在开放性测试集中的 F_1 值达到 95% 左右。张素香等使用专家知识对各类特征进行定义,利用条件随机场建立了相应的语言模型[6],在人民日报语料上进行实验,也获得了超过 95% 的 F_1 值。这些方法都可以看作是从句中提取一组依靠纯语言学经验的人工设计的特征,再经过不断地修正将其馈送到经典的浅分类算法中。特征选择因应用背景的不同,对每种 NLP 应用都需要进行额外的研究,任务较为复杂时(比如 SRL 语义角色标注[7])就需要进行大量关于特征的人工提取。由于受研究人员语言学知识的限制,这些模型随着语料规模的扩大,工作效率会出现显著下降。

命名实体识别应用为了规避这种限制,目前的研究多采用词的向量化[8]对自然语言进行数字化表示,以期刻画词的更一般性。1986 年 Hinton 提出了 Distributed Representation[9],将语料中每个词用一个 N 维向量表示,N 个维度代表词的 N 个不同特征,有效地避免 One-hot 所造成的维度灾难、数据稀疏等问题,并且能用向量的距离来计算词与词之间意思的远近,刻画了词与词之间所隐藏的关联性,避免了两个词之间的孤立。但是要从大量未标记文本中无监督地提取词向量,还必须借助语言模型。2003 年,Bengio 等在 n 元模型的基础上,提出神经概率语言模型-NNLM[10],用一种三层前馈神经网络建模词的上下文关系。后人将词向量和神经网络引入到命名实体识别工作中,极大地减少了维度灾难对模型性能的影响。2007 年,为了提高效率,降低算法时间复杂度,Mnih 等又提出 Log 双线性语言模型-LBL[11]。2010 年 Mikolov 等使用循环神经网络模型[12],通过迭代多个隐藏层保存更丰富的上下文信息,提高了命名实体识别的潜力。

2011 年 Collobert 等提出一种灵活的层叠式神经网络架构[13],将训练好的词向量应用于各种 NLP 任务,避开过多的面向特定任务的特征工程,在降低计算资源的同时保持了优秀的性能。Mikolov 在 2013 年提出 CBOW 和 Skip-gram 两个模型[14],以过渡的投影层代替隐藏层,减少了矩阵运算的次数,提高了训练速度,而且用上下文各词的词向量的平均值来刻画目标词,排除了词序对训练的影响。近年来,卷积神经网络(CNM)[15]、Glove 模型[16]、时间递归神经网络(LSTM)等模型相继被应用到命名实体识别中,在提高识别效果的同时也不断降低识别过程对语言学知识的依赖,使用无监督学习进行命名实体识别已经成为国外命名实体识别技术研究的新方向。但是由于中文的复杂性,中文命名实体识别在无监督学习方向上的研究目前仍然存在不足。

2 网络架构

文中提出一种基于深度神经网络的中文人名训练架构,如图 1 所示,图中使用的训练语料和测试数据集均进行了分词预处理。训练可概述为 4 个模块:

(1)基本词面的向量化训练。仅基于训练语料统计词频,以词频为依据利用连续词袋 CBOW 模型迭代训练得到基本词面向量。

(2)特征提取及其特征项向量表示。获取包含每个人名特征项的目标词上下文的基本词面向量,构建共现矩阵,提取特征向量作为特征项向量。

(3)中文人名识别的模型参数训练。以目标词的基本词面向量与各特征向量的拼接作为输入,构建前向全连接的深度神经网络,使用梯度下降法迭代更新训练参数,并用 Viterbi 算法返回最佳路径,最终获得

稳定的模型参数。

(4)使用测试语料进行模型测试。测试模块使用训练好的稳定参数构建中文人名识别模型,根据 Viterbi 算法返回的最优路径,设定阈值提取标签结果并进行分析。

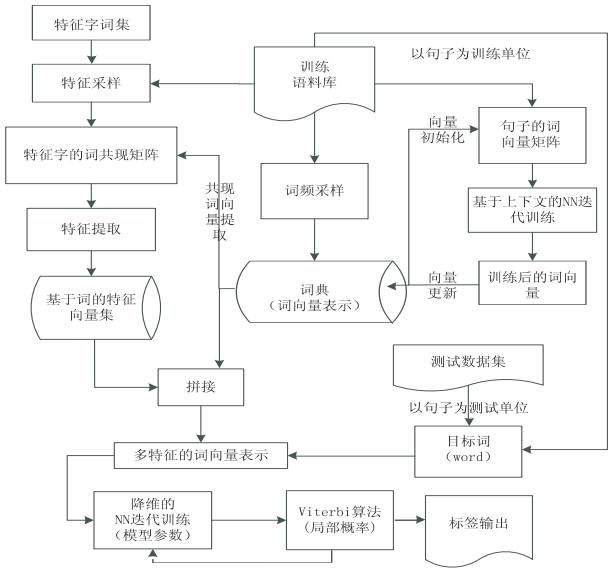

图 1　中文人名识别的训练框架

2.1　基本词面的向量化训练

训练以句子为单位迭代进行,训练的前期工作是遍历语料统计词频信息,以词频对所有词降序排序,并建立词典的 hash 索引以应对训练中的频繁查询。

对目标词 w 来说, Context(w) 表示其由 m 个词组成的上下文,训练中去除上下文中的词序信息,使用上下文中各个词向量的均值作为输入,具体表示为:

$$x = \text{Context}(w) = \frac{1}{m}\sum_{i=1}^{m} v(w_i) \tag{1}$$

模型为了提高训练速度,摈弃传统隐含层以减少矩阵运算,直接从神经网络结构转化为与 Logistic 回归一致的 Log 线性结构,用二分类的思想以式 2 所示的目标函数为优化目标:

$$E = \sum_{w \in D} \log p(w \mid x) \tag{2}$$

输出层是对于特征函数值进行一个二分类的处理以得出其发生的条件概率。在实际训练中,将任意一个目标词作为叶子节点,取其词频作为权值来构造 Huffman 树。从根节点到该目标词的叶子节点路径上的每个分支节点,都是一个影响目标词最终语义的隐式二分类器。训练中构造与每一个分支节点对应的权值向量,组成权值矩阵作为训练参数,这个参数矩阵在迭代训练中得到更新。为此,在模型构造中引入 Sigmoid 函数:

$$g(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

对于某一个目标词 w ,其 Huffman 树路径中的每一个分支节点,通过权值向量 $\boldsymbol{\theta}$ 计算其与上下文环境向量 $\boldsymbol{x}$ 语义距离 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$,依据 Sigmoid 函数,可将分支节点对目标节点的概率贡献归结为正类概率 $p(L=1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$ 或负类概率 $p(L=0 \mid \boldsymbol{x}, \boldsymbol{\theta}) = 1 - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$。其中 L 可视为最终的标签输出,取值为 1 表示正类,取值为 0 表示为负类,二式可统一写为:

$$p(L \mid \boldsymbol{x}, \boldsymbol{\theta}) = g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})^{L} \cdot [1 - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})]^{1-L} \tag{4}$$

由此,可得到目标词 w 受上下文的语义影响产生的隐含条件概率为:

$$p(w \mid \boldsymbol{x}) = \prod_{i=1}^{n} p(L \mid \boldsymbol{x}, \boldsymbol{\theta}) \tag{5}$$

其中, n 为目标词 Huffman 编码的长度,也就是从根节点到目标词节点路径中的父节点个数,由式 2、4 与 5 可推出目标函数:

$$E = \sum_{w \in D}\sum_{i=1}^{n} \{(L \cdot \log[g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})] + (1 - L) \cdot \log[1 - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})])\} \tag{6}$$

接下来使用随机梯度下降法对 E 更新,函数 E 关于参数 $\boldsymbol{\theta}$ 的梯度计算为: $\partial E/\partial \boldsymbol{\theta} = [L - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})]\boldsymbol{x}$,由此得到参数 $\boldsymbol{\theta}$ 的更新公式为: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta[L - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})]\boldsymbol{x}$。同理得到参数 $\boldsymbol{x}$ 的更新公式为: $\boldsymbol{x} \leftarrow \boldsymbol{x} + \eta[L - g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})]\boldsymbol{\theta}$ 。训练目标是使得目标词受上下文环境影响的编码概率最大,向量 $\boldsymbol{x}$ 与 $\boldsymbol{\theta}$ 的更新仅用于训练迭代计算,因为 $\boldsymbol{x}$ 是目标词 w 的上下文向量的累加的均值,因此目标词的向量 $\boldsymbol{v}(w)$ 的更新是一次迭代训练完成后,将梯度的累加更新到向量的每一个分量中,如式 7 所示:

$$\boldsymbol{v}(w) \leftarrow \boldsymbol{v}(w) + \eta\sum_{i=1}^{n}\frac{\partial E}{\partial x} \tag{7}$$

其中, η 表示学习率。训练中 η 随着模型处理而动态调整,直到达到阈值 $\eta_{\min}$ 为止,目的在于缓解梯度更新中的波动,使更新过程更加平稳。根据上述的迭代算法依次对所有的目标词和它们的词频进行训练,通过神经网络的反复迭代,网络对于输入的响应达到预定的目标范围后, $w(v)$ 就是训练好的最后的词向量。

2.2　特征提取及其特征项向量表示

在中文人名识别领域,特征的提取主要考虑人名的姓氏特征、称谓修饰特征与特殊人名特征,通过构建特征知识库,以特征的每一个特征项为单位,从训练语料中匹配其窗口上下文,从前一个模块的训练结果中提取拥有该特征项的目标词向量,构建特征项的共现矩阵 $\boldsymbol{R}$。对于特征项共现矩阵 $\boldsymbol{R}$, 求其协方差矩阵 $\boldsymbol{C}_R = 1/n * \boldsymbol{R}\boldsymbol{R}^{\mathrm{T}}$ 。

利用公式 $\boldsymbol{C}_R\boldsymbol{\mu} = \boldsymbol{\lambda}\boldsymbol{\mu}$ 求得协方差矩阵的所有特征值与所有特征向量 $\boldsymbol{\mu}$,对所有非 0 的特征值所对应的

特征向量取均值,即为所求的特征项向量。

2.3 中文人名识别的模型参数训练

训练以句子为单位,使用统一的模型架构,分三种方案组织输入向量 $\boldsymbol{S}_t(t \in (1,n))$ 训练三组模型参数,以评估人名识别的特征提取对于标签结果的影响。方案一,仅以词频训练的基本词面向量作为输入。方案二,在基本词面向量的基础上,融入姓氏特征和称谓修饰特征等人名的基本特征,携带更多的上下文信息。方案三,在方案二的基础上再引入特殊人名知识库,结合特殊人名特征进行强化学习,以期得到更佳的模型参数。特别需要指出的是,在训练中使用 UNKNOWN 关键词作为未识词和未识特征项的初始向量。将这个目标词组合向量作为训练模型输入,在后续的神经网络层中进行运算。

图 2 是中文人名识别的模型参数训练架构。

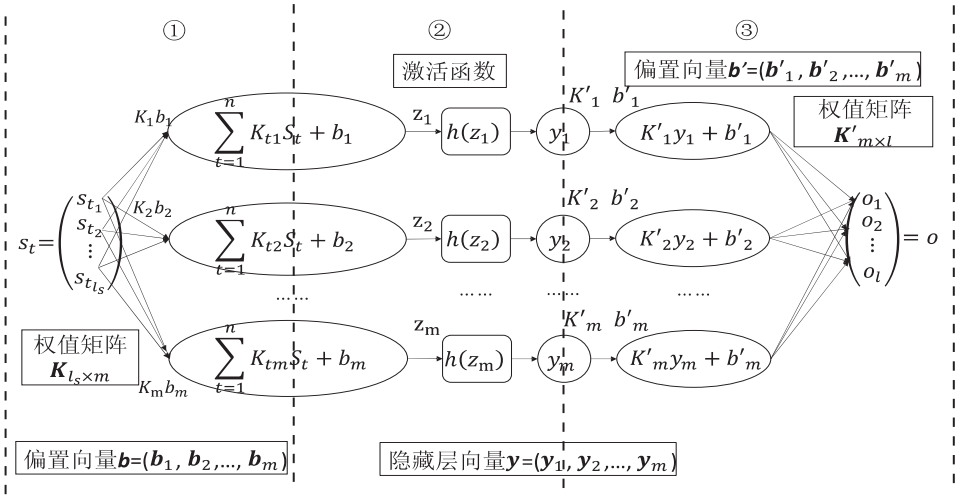

图 2 中文人名识别的模型参数训练架构

对于输入目标词向量 $\boldsymbol{S}_t$,在架构的一个或多个神经网络的隐藏层中做如下的映射变换进行降维操作,如图中的步骤①所示:

$$\boldsymbol{S}_t^{\tau} = \boldsymbol{K}^{\tau}\boldsymbol{S}_t^{\tau-1} + \boldsymbol{b}^{\tau} \tag{8}$$

其中,$\boldsymbol{K}$ 和 $\boldsymbol{b}$ 分别表示权值矩阵和偏置向量;τ 表示堆叠的隐藏层的层数。

将这多个隐藏层进行堆叠,对输入向量 $\boldsymbol{S}_t$ 进行降维,将其转化成一个向量 $\boldsymbol{z}$ 以获取非线性特征。这里要添加一个硬性的双曲正切函数 $h(z)$ 作为激活函数,用以保证模型的非线性,最终得到隐藏层向量 $\boldsymbol{y}$,如图中的步骤②所示。

$$h(z) = \mathrm{HardTanh}(z) = \begin{cases} -1 & z < -1 \\ z & -1 \leqslant z \leqslant 1 \\ 1 & z > 1 \end{cases} \tag{9}$$

激活函数的优点在于能够在不增加过多的计算量的同时保证泛化性能基本不变。

类似输入层到隐藏层的处理,对于隐藏层向量 $\boldsymbol{y}$,同样对它在后续的一个或多个神经网络的隐藏层中做映射变换,从而进行降维,最终获得输出层向量 $\boldsymbol{o}$,如图中的步骤③所示。这里的权值矩阵和偏置向量为 $\boldsymbol{K}'$ 和 $\boldsymbol{b}'$。

根据逻辑回归的二分类可以得到条件概率 $p(L \mid \boldsymbol{S}_t,\varphi)$。

$$p(L \mid \boldsymbol{S}_t,\varphi) = [g(\boldsymbol{S}_t \cdot \varphi)]^L \cdot [1 - g(\boldsymbol{S}_t \cdot \varphi)]^{1-L} \tag{10}$$

以输出向量 $\boldsymbol{o}$ 为观察序列,用随机梯度下降法对所有的参数 φ 进行梯度更新,在更新的同时用一个反向指针保存更新的路径:

$$\varphi \leftarrow \varphi + \lambda \frac{\partial p(L \mid \boldsymbol{S}_t,\varphi)}{\partial \varphi} \tag{11}$$

其中,φ 是参数更新过程中所有参数组成的集合;L 是二分类的标签。与前面一样,引入一个学习率 λ 来减少梯度更新中的波动。

Viterbi 算法是解决分类标签标注问题的理想选择。使用式 11 每进行一次更新,都能通过反向指针,对 Viterbi 算法的初始概率矩阵和状态转移概率矩阵进行更新,参数更新终止条件是 Viterbi 算法中最终局部概率达到最大,即人名识别的条件概率达到局部最优。达到局部最优时的参数集就是模型最佳时的稳定参数集 φ^*,$\varphi^* = \{\boldsymbol{K},\boldsymbol{b},\boldsymbol{K}',\boldsymbol{b}',\pi,A\}$,通过 φ 确定最佳的中文人名识别模型,根据输入向量 $\boldsymbol{S}_t$ 的三种不同组织方案最终获得三种最佳模型。

2.4 使用测试语料进行模型测试

确定最佳的中文人名识别模型后,引入测试数据集,以句子为单位进行人名识别测试。首先将测试句进行分词等预处理操作,随后获取句子中每个词的基本词面向量,根据 2.3 节提到的三种方案训练得到的三种最佳模型分别获得句子中每个词的输出向量。

对于每一种方案,将所有词的输出向量序列作为 Viterbi 算法的观察序列进行输入,根据 φ^* 中最优模

型的初始概率矩阵和状态转移概率矩阵,得到一条最优路径,最优路径上的每个概率值都是对应目标词是人名的最优概率,随后人工设置一个阈值对其进行判断。对每个词来说,是人名的概率大于这个阈值即判断其为人名,否则判断其不是人名。

最后根据每种方案的不同结果对三种模型的效果分别进行一个评价。

3　实验与分析

3.1　实验语料

使用的训练语料是来自 2008 年的搜狐新闻总计超过 1.2 GB 的中文文本,总计文本的句子数目共计 26 964 条,使用 20 223 条进行训练,6 741 条用于测试,采用中科院 ICTCLAS 分词系统进行分词操作,分词后统计获得的词面数共有 89 464 个。

3.2　评价标准

对于实验结果,采用识别率 P 、召回率 R 和它们的调和平均值 F_1 值作为评价标准。识别率表示正确识别出的人名占总计识别出的人名的百分比,衡量模型排除非相关信息干扰的能力;召回率表示正确识别出的人名占语料中所有人名的百分比,衡量模型获得相关信息的能力; F_1 值作为识别率和召回率的调和平均值,是对模型性能的综合考量。

记正确识别出的人名的个数为 α ,总计识别出的人名的个数为 β ,语料中所有人名的个数为 γ ,则具体公式如下:

$$P = \frac{\alpha}{\beta} \times 100\% \tag{12}$$

$$R = \frac{\alpha}{\gamma} \times 100\% \tag{13}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{14}$$

3.3　实验结果

根据 2.3 节中提到的三种方案训练得到的三种最佳模型进行人名识别,共获得三个实验结果。使用文献[9]的隐马尔可夫模型和文献[10]的条件随机场,在文中所使用的训练语料中进行实验所得到的结果作为对照,进行对比分析。具体实验结果如表 1 所示。

表 1　各模型的人名识别结果　　%

方法	P	R	F_1
文献[9]	87.73	78.19	82.69
文献[10]	90.10	81.89	85.80
方案 1	73.67	68.35	70.91
方案 2	90.78	83.26	86.85
方案 3	92.36	86.01	89.07

通过表 1 不难得出,方案 1 在只考虑词频特征时,识别率、召回率、F_1 值都十分低下,甚至劣于隐马尔可夫模型和条件随机场的识别效果。但方案 2 在添加了人名的基本特征后,识别效果得到很大的提升,优于传统的识别方法。而方案 3 在引入了人工离散特征之后,识别效果获得了进一步的提升,其 F_1 值甚至能达到 89.07% 。说明文中提出的模型在中文人名识别中,能够在尽量避免工程化特征影响的同时保持较好的识别效果,并能够在添加少量人工特征时获得更好的性能,在高度工程化的实际应用中也非常有效。

4　结束语

以词向量为媒介,设计深度神经网络架构,将复杂的命名实体识别工作转移到多层的前向全连接的神经网络中实现,尽可能避免工程化特征的影响,有效提高了识别效率。但随着计算机技术的发展,递归神经网络等更先进的技术也开始逐渐应用到命名实体识别任务之中,笔者在对这些先进技术的调研中发现这些新技术在命名实体识别领域的应用尚处在开始阶段,效果并不能达到预期的要求。下一步的工作中,将对递归神经网络做进一步的研究,用其对现在的前向全连接网络进行改造,以期望获得更好的识别效果。

参考文献:

[1] 袁　琦. 中文信息技术和自然语言处理[J]. 中文信息学报,1986(1):33-36.

[2] BIKEL D M, MILLER S, SCHWARTZ R, et al. Nymble: a high-performance learning name-finder[C]//Proceedings of the 15th conference on applied natural language processing. Washington, DC: [s. n.], 1997: 194-201.

[3] SEKINE S. NYU: description of the Japanese NE system used for MET-2[C]//Proceedings of the 7th message understanding conference. [s. l.]: [s. n.], 1998: 1-6.

[4] ASAHARA M, MATSUMOTO Y. Japanese named entity extraction with redundant morphological analysis[C]//Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology. Edmonton, Canada: [s. n.], 2003: 8-15.

[5] 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(1):85-91.

[6] 张素香,高国洋,戚银城. 基于条件随机场的中国人名识别方法[J]. 郑州大学学报:理学版,2009,41(2):40-43.

[7] 刘　挺,车万翔,李　生. 基于最大熵分类器的语义角色标注[J]. 软件学报,2007,18(3):565-573.

[8] 张　剑,屈　丹,李　真. 基于词向量特征的循环神经网络语言模型[J]. 模式识别与人工智能,2015,28(4):299-305.

(下转第 62 页)

换的手机身份证字符分割速度很快;实验中字符分割基本正确,正确率达到 99% 。

7 结束语

针对安卓手机上复杂背景中的身份证字符分割,设计了一种基于透视变换的算法,并设计了一种目标区域内连接同一条直线的方法。实验结果表明,该方法具有较为良好的分割能力。另一方面,该方法对全白背景中的身份证字符分割效果较差,还需要进一步的探索。

参考文献:

[1] 严 曲. 身份证识别系统的原理及算法研究[D]. 长沙:中南大学,2005.

[2] 许新征,丁世飞,史忠植,等. 图像分割的新理论和新方法[J]. 电子学报,2010,38(2A):76-82.

[3] PAL N R,PAL S K. A review on image segmentation techniques[J]. Pattern Recognition,1993,26(9):1277-1294.

[4] 李 静,卢凯旋. 二代身份证的自动分割方法研究[J]. 计算机工程与应用,2015,51(14):165-169.

[5] ALVAREZ L,LIONS P L,MOREL J M. Image selective smoothing and edge detection by nonlinear diffusion. II[J]. SIAM Journal on Numerical Analysis,1992,29(3):845-866.

[6] 赵高长,张 磊,武风波. 改进的中值滤波算法在图像去噪中的应用[J]. 应用光学,2011,32(4):678-682.

[7] 许宏科,秦严严,陈会茹. 一种基于改进 Canny 的边缘检测算法[J]. 红外技术,2014,36(3):210-214.

[8] 金 刚. 自适应 Canny 算法研究及其在图像边缘检测中的应用[D]. 杭州:浙江大学,2009.

[9] 段瑞玲,李庆祥,李玉和. 图像边缘检测方法研究综述[J]. 光学技术,2005,31(3):415-419.

[10] 张 帆,彭中伟,蒙水金. 基于自适应阈值的改进 Canny 边缘检测方法[J]. 计算机应用,2012,32(8):2296-2298.

[11] 杜 奇,向健勇,袁胜春. 一种改进的最大类间方差法[J]. 红外技术,2003,25(5):33-36.

[12] 段汝娇,赵 伟,黄松岭,等. 一种基于改进 Hough 变换的直线快速检测算法[J]. 仪器仪表学报,2010,31(12):2774-2780.

[13] ILLINGWORTH J,KITTLER J. The adaptive Hough transform[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,1987,9(5):690-698.

[14] 王燕清,辛柯俊,陈德运,等. 基于启发式概率 Hough 变换的道路边缘检测方法[J]. 计算机科学,2013,40(9):279-283.

[15] 牛 彦. 关于透视变换的研究[J]. 计算机辅助设计与图形学学报,2001,13(6):549-551.

[16] 代 勤,王延杰,韩广良. 基于改进 Hough 变换和透视变换的透视图像矫正[J]. 液晶与显示,2012,27(4):552-556.

[17] 肖西华,江志兴,梁 旭,等. 移动平台下的身份证图像字符分割方法研究[J]. 计算机工程与应用,2015,51(24):201-204.

[18] 苗红霞,张 龙,徐文杰,等. 一种身份证图像字符分割的改进方法[J]. 微处理机,2016,37(3):51-55.

(上接第 57 页)

[9] HINTON G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science. [s. l.]:[s. n.],1986:1-12.

[10] BENGIO Y,DUCHARME R,VINCENT P,et al. A neural probabilistic language model[J]. Journal of Machine Learning Research,2003,3:1137-1155.

[11] MNIH A,HINTON G. Three new graphical models for statistical language modelling[C]//Proceedings of the twenty-fourth international conference on machine learning. Corvalis,Oregon,USA:ACM,2007:641-648.

[12] MIKOLOV T,KARAFIAT M,BURGET L,et al. Recurrent neural network based language model[C]//11th annual conference of the international speech communication association. Makuhari,Chiba,Japan:[s. n.],2010:131-138.

[13] COLLOBERT R,WESTON J,BOTTON L,et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011,12(1):2493-2537.

[14] MIKOLOV T,CHEN Kai,CORRADO G,et al. Efficient estimation of word representations in vector space[C]//International conference on learning representations workshop track. [s. l.]:[s. n.],2013:1301-1313.

[15] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. [s. l.]:[s. n.],2014:1746-1751.

[16] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]//Proceedings of the empirical methods in natural language processing. [s. l.]:[s. n.],2014:1532-1543.