DOI:10.11992/tis.201507065
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160106.1555.002.html

词边界字向量的中文命名实体识别

姚霖[1,2,3],刘轶[1],李鑫鑫[4], 刘宏[2]
(1. 深港产学研基地,广东 深圳 518057; 2.北京大学 信息科学技术学院,北京 100871; 3.哈尔滨工业大学 软件学院,黑龙江 哈尔滨 150001; 4.哈尔滨工业大学深圳研究生院 计算机科学与技术学院,广东 深圳 518055)

摘　要:常见的基于机器学习的中文命名实体识别系统往往使用大量人工提取的特征,但特征提取费时费力,是一件十分繁琐的工作。为了减少中文命名实体识别对特征提取的依赖,构建了基于词边界字向量的中文命名实体识别系统。该方法利用神经元网络从大量未标注数据中,自动抽取出蕴含其中的特征信息,生成字特征向量。同时考虑到汉字不是中文语义的最基本单位,单纯的字向量会由于一字多义造成语义的混淆,因此根据同一个字在词中处于不同位置大多含义不同的特点,将单个字在词语中所处的位置信息加入到字特征向量中,形成词边界字向量,将其用于深度神经网络模型训练之中。在 Sighan Bakeoff-3(2006)语料中取得了 F_1 89.18%的效果,接近当前国际先进水平,说明了该系统不仅摆脱了对特征提取的依赖,也减少了汉字一字多义产生的语义混淆。

关键词:机器学习;中文命名体识别;深度神经网络;特征向量;特征提取
中图分类号:TP391.1　**文献标志码**:A　**文章编号**:1673-4785(2016)01-0037-06

中文引用格式:姚霖,刘轶,李鑫鑫,等.词边界字向量的中文命名实体识别[J]. 智能系统学报, 2016, 11(1): 37-42.
英文引用格式:YAO Lin, LIU Yi, LI Xinxin, et al. Chinese named entity recognition via word boundary based character embedding[J]. CAAI Transactions on Intelligent Systems, 2016, 11(1): 37-42.

Chinese named entity recognition via word boundary based character embedding

YAO Lin[1,2,3], LIU Yi[1], LI Xinxin[4], LIU Hong[2]
(1. Shenzhen High-Tech Industrial Park, Shenzhen 518057, China; 2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China; 3. School of Software, Harbin Institute of Technology, Harbin 150001, China; 4. School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China)

Abstract: Most Chinese named entity recognition systems based on machine learning are realized by applying a large amount of manual extracted features. Feature extraction is time-consuming and laborious. In order to remove the dependence on feature extraction, this paper presents a Chinese named entity recognition system via word boundary based character embedding. The method can automatically extract the feature information from a large number of unlabeled data and generate the word feature vector, which will be used in the training of neural network. Since the Chinese characters are not the most basic unit of the Chinese semantics, the simple word vector will be cause the semantics ambiguity problem. According to the same character on different position of the word might have different meanings, this paper proposes a character vector method with word boundary information, constructs a depth neural network system for the Chinese named entity recognition and achieves F_1 89.18% on Sighan Bakeoff-3 2006 MSRA corpus. The result is closed to the state-of-the-art performance and shows that the system can avoid relying on feature extraction and reduce the character ambiguity.

Keywords: machine learning; Chinese named entity recognition; deep neutral networks; feature vector; feature extraction

命名实体识别(named entity recognition,NER)是计算机理解自然语言信息的基础,其主要任务是从文本中识别出原子元素,并根据其所属类别,标注预定义的标记,如人名、地名、组织机构名等。由于其在自然语言处理领域中的重要作用,许多国际会议,如 MUC-6、MUC-7、Conll2002、Conll2003 等,将命名实体识别设为共享任务(share tasks)。英文命名实体识别具有相对较长的发展历史。许多机器学习方法,如最大熵[1-4]、隐马尔可夫模型[2,5-7]、支持向

收稿日期:2015-08-13.　网络出版日期:2016-01-06.
基金项目:原创项目研发与非遗产业化资助项目(YC2015057).
通信作者:姚霖. E-mail: 1250047487@qq.com.

量机[8]和条件随机场[9]等都曾被应用于命名体识别任务,并取得了较好的精确度。英语中的人名、地名和组织机构名具有首字母大写等特点,因此英文命名实体识别相对简单。然而中文的语言特点与英文大不相同,字和词之间没有明确的界限,人名、地名和组织机构名也都没有首字大写的特点,而且对于外国人名和组织机构名,常常会有不同的翻译。上述特点使得中文命名实体识别任务更具有挑战性。前面提到的有监督机器学习方法,如隐马尔科夫模型(HMM),最大熵(ME)[10],支持向量机(SVM)和条件随机场(CRF)[11]算法等,也都曾被应用于解决中文命名体识别问题。这类监督学习方法训练过程中需要人工定义复杂不同的特征模板以获得较好的识别率。设计和比较模板的工作不仅要求开发人员具备坚实的语言学背景,还需要花费大量的时间通过实验进行筛选,是一件费时费力的工作。

为了减少中文命名实体识别任务对人工构建特性模板的过度依赖,在 Collobert[12] 工作的启发下,我们搭建了一个基于词边界字向量的中文命名实体识别系统。利用字在词语中所处位置不同,含义大不同的特点,提出了词边界字向量的概念。在一定程度上减少了一字多义产生的语义混淆。该系统在标准语料库 SIGHAN Bakeoff-3(2006)上取得了较好的识别效果。

本文首先介绍了中文命名实体识别领域的发展现状;接着详细描述了基于词边界字向量的中文命名体识别系统的架构;其后对系统的结果和性能进行了进一步的分析;最终进行了总结和展望。

1 系统架构

Bengio 首次将卷积神经网络架构(convolutional neural network,CNN)应用到概率语言模型中[13],并成功应用于自然语言处理(natural language processing,NLP)问题[14]。在本系统中,我们使用 CNN 来处理中文命名体识别问题,神经网络架构如图 1 所示。

首先将待测汉字通过系统的查找层,和滑动窗口中其他相邻汉字一起,被转化为实数字向量,再将该向量序列输入到下层神经网络中。如图 1 所示,在时刻 t,系统待测汉字为处于滑动窗口正中的汉字"的"。通过查找层后,汉字"的"以及相邻的汉字"香、港、繁、荣"(假设窗口大小为 5)被转换为实数向量,传输到线性转换层。经过线性层和 sigmoid 层的处理后,系统对字符"的"对应所有可能标记进行打分,标注概率越大的标记,分值越高。在下一时刻 $t+1$,系统接着处理句子中的下一个汉字"繁",以此类推。在句子解析层中,将针对整句生成一个由分值组成的网格。网格中第 t 列上的节点是 t 时刻待测汉字对应所有标记的分值。节点间连线上显示转移概率,用来描述标记间的转换可能性。转移概率在全局统计的基础上产生。最后应用维特比(Viterbi)算法,在分值网格中找到分值最高的路径,作为最终的标记序列。

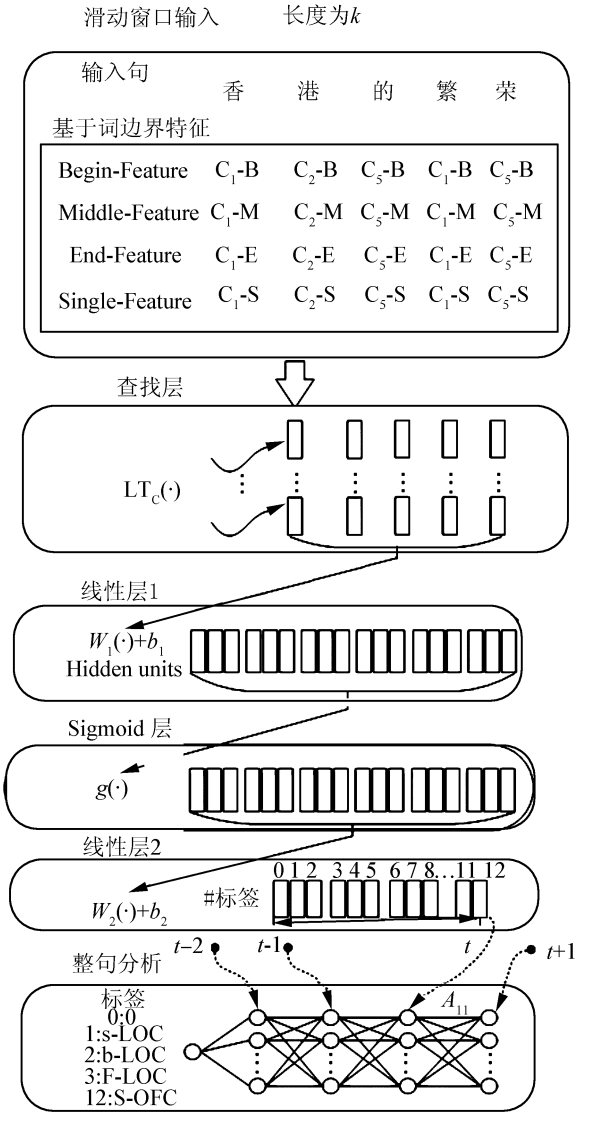

图 1 神经网络架构

Fig.1 The neural network architecture

1.1 字特征向量

生成特征向量序列的第一步要将全部汉字存储到查找层的字典 D 中,每个汉字由固定维度的实数向量表示。语句在通过查找层后,被转换为字向量序列,汉字 $x \in D$ 的特征向量可以通过方程 $\mathrm{LT}_{\mathrm{C}}(\cdot)$ 获取,方程定义如下:

$$\mathrm{LT}_{\mathrm{C}}(x) = C_x$$

式中:$C_x \in \mathrm{R}^{|s|}$ 是汉字 x 对应的字向量,s 表示向量的维度。查找表 $\boldsymbol{C}$ 为由字典 D 中的汉字以及其对应的向量组成的矩阵。该特征向量矩阵是通过深层神经网络模型,在海量未标注的中文数据上训练得到的。高维度的特征向量通过向量间的差值能够较为准确地捕捉到字/词间的句法和语义关系,自动包含汉字间所蕴藏的句法和语义信息。

利用神经网络模型获得中英文字词向量的工作已经有了一定的应用[13,15-17]。在本文中,我们采用相同的方式通过语言模型获得到向量矩阵。尽管采用了庞大的训练语料,但由于语言的复杂多变性,数据稀疏始终是中文名实体识别中存在的问题。通过对比不同的语言模型[12,17-20],我们最终选择 skip-gram 神经网络模型。

与 Collobert 和 Weston[12]、Turian[19]、Mnih 和 Hinton[17]采用的语言模型相比,Tomas Mikolov[21]的工作说明 skip-gram 模型在词类推任务(word analogy)能够获得较好的成绩,该模型虽然在训练速度上不占优势,但适合解决数据稀疏问题。skip-gram 模型使用当前的字/词向量来预测该字/词之前和之后各$(k-1)\div 2$ 个字/词的概率,如图 2 所示。该模型的优化目标是最大化训练数据的对数似然度:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{-k/2\leqslant j\leqslant k/2,j\neq 0}\log p(x_{i+j}\mid x_j)$$

式中:x_i 为训练语料中的汉字,k 为窗口大小。概率 $p(x_{i+j}|x_j)$ 由 softmax 公式得到,定义如下:

$$p(x_o\mid x_I)=\frac{\exp v_{x_o}v_{x_I}}{\sum_{x=1}^{D}\exp(v_x{}'v_{x_I})}$$

式中:v_x 为汉字 x 的向量初始值,$v_x{}'$表示输出向量。

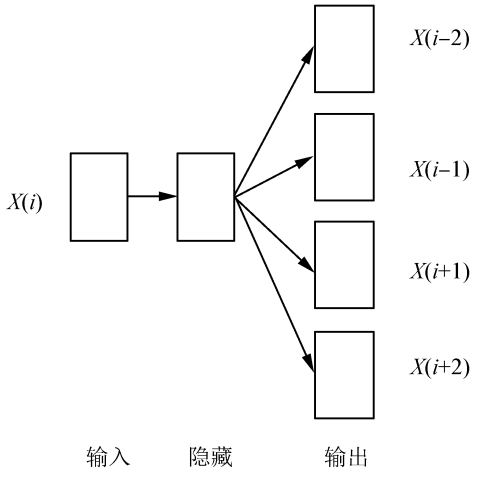

图 2　Skip-gram 神经网络语言模型

Fig.2　The skip-gram neural network language model

字向量虽然为分布式描述,但能够表达汉字间存在的相互关系,其泛化(generalization)的程度是其他传统的 N 元文法模型无法达到的。模型参数会依据属性相似的汉字出现的次数加以调整。

1.2　语句级特征抽取

中文命名实体识别对语句中的每个汉字,标注相应的实体类型,输出是针对整个句子产生的一串标记序列。由于 CNN 的输入端长度固定,与自然语言中句子变长的特点不符,因此我们采用了机器学习领域较为常见的滑动窗口方法,将待标注的句子切分成特定长度的片段,分批输入。图 1 中,在时刻 t,当前处理的汉字是 p 位置上“的”字,与该字距离在 $[(p-(k-1)/2),(p+(k-1)/2)]$ 范围内的相邻字也将一同输入到查找层,从而转换成字向量。通过字向量抽取出来的蕴含在这个范围内的句法和语义信息将会被传递到系统的下一层。人为设定窗口大小为 k。k 值对系统精度有一定影响,如果选择窗口过小,有利信息不能被覆盖;而窗口过大,因此带来的冗余信息对系统产生不必要的干扰。设字向量为 s-维,则线性层的输入大小为 $s\times k$。

1.3　标记预测

深层神经网络为多层结构,每一层都在前一层获得的特征基础上,进一步提取特征。根据设计,各层由不同的线性函数或其他转换函数实现。公式 $f_\theta(\cdot)$描述本系统深层神经网络的中间 3 层:

$$f(x)=M^2g(M^1C_x+b^1)+b^2$$

式中:$M^1\in\mathrm{R}^{H\times SK}$,$b^1\in\mathrm{R}^{1\times H}$,$M^2\in\mathrm{R}^{|L|\times H}$,$b^2\in\mathrm{R}^{1\times|L|}$,$g(\cdot)$代表 sigmoid 转换。$H$ 为隐藏节点的个数,可以通过调整该值,获得更高的精确度。$|L|$为标注集的大小。$\theta=(M^1,M^2,b_1,b_2)$代表系统中所有通过训练获得的参数。采用随机梯度法,在训练集 T 上,通过最大似然率 $\sum_{(x,y)\in T}\log p(y\mid x,\theta)$ 来训练得到 v-维的参数矩阵 $\boldsymbol{\theta}(\theta_1,\theta_2,,\cdots,\theta_v)$。

中文命名体识别属于多分类问题,用 $f(x,l,\theta)$ 表示汉字 x 标注为第 l 个标记的分值,通过条件概率 $p(l|x,\theta)$描述。应用 softmax 回归得到

$$p(l\mid x,\theta)=\frac{\mathrm{e}^{f(x,l,\theta)}}{\sum_j \mathrm{e}^{f(x,j,\theta)}}$$

为方便计算,定义操作 log-add 为

$$\log\operatorname*{add}_i z=\log\Big(\sum_i e^{z_i}\Big)$$

因此一个训练样本(x,y)的对数似然率为:$\log p(y|x,\theta)=f(x,y,\theta)-\log\operatorname*{add}_j f(x,j,\theta)$句子 $x_{[1:T]}$处在 t 时刻的汉字标注为 l 的分值为 $f(x_{[1:T]},l,t,\theta)$,

系统参数集为 θ。

作为一项序列标注任务,中文命名体识别要根据整条标记路径的分值情况,判断最终的标注结果。句子中的邻近标记间存在着关联性。例如,命名实体 B 类的内部标记不可能紧跟在类 A 的左边界标记后面出现;组织机构中间标记不可能紧跟在人名的起始标记后面。因此,我们将标记之间存在的依存关系引入到系统中来。Collobert 用 A_{lj}来描述这种依存关系,即标记 i 到标记 j 的转移概率。因此整句的标注过程可以转化为在由标记组成的图中找一条总体分值最高的路径。一个句子的标记路径分值是由两部分组成,第一部分为前面描述的分值 $f(x_{[1:T]},l,t,\theta)$;第二部分是 A_{ij}。从而系统的所有参数集 $\tilde{\theta}$ 包括 A_{ij}和 θ。为了加强序列中邻近标记之间的关联,本文将标记推导层的一阶标记转移矩阵扩展为二阶标记转移矩阵,用 A_{ijk}来表示标记 k 与前面标记(i, j)的关联关系。则对于句子的总体分值可以描述为

$$S(x_{[1:T]},l_{[1:T]},\tilde{\theta})=\sum_{t=1}^{T}(A_{l_{t-2}l_{t-1}l_t}+f(x_{[1:T]},l_{[t]},t,\theta))$$

式中:$l_{[1:T]}$ 为标记集。采用前面定义的 log-add 运算简化计算,归一化后得到:$\log p(y_{[1:T]}\mid x_{[1:T]},\tilde{\theta})=S(x_{[1:T]},y_{[1:T]},\tilde{\theta})-\underset{\forall l_{[1:T]}}{\log\text{add}}S(x_{[1:T]},l_{[1:T]},\tilde{\theta})$ 在训练过程中,使用全部的训练样本对参数集 $\tilde{\theta}$ 都会进行训练,即最大化 $\sum_{(x,y\in T)}\log p(y_{[1:T]}\mid x_{[1:T]},\tilde{\theta})$。在生成的网格中,我们使用 viterbi 算法选择最高分值路径,即满足$\underset{l_{[1:T]}}{\arg\max}S(x_{[1:T]},l_{[1:T]},\tilde{\theta})$的标记路径。

2 实验结果与分析

2.1 相关数据集

训练 skip-gram 模型的原始数据是包括人民日报和搜狗实验室提供的新闻语料库,总计超过 2GB 的中文文本。采用 ICTCLAS 工具进行分词。采用 Word2vec 工具训练 skip-gram 模型。

用来训练参数集 $\tilde{\theta}$ 的标注训练语料是 SIGHAN bakeoff-3 MSRA 语料(2006)。语料库的统计信息如表 1 所示。

使用上述数据,我们完成了两组实验。第 1 组实验只采用汉字特征向量一种特征,考察了深度神经网络模型本身的识别能力。该组实验滑动窗口大小为 3,隐藏层节点 500 个,不同特征向量维数下得到的实验结果如图 3 所示。

表 1 标注数据描述

Table 1 Labeled data descriptions

数据集	包含句子的数量	包含汉字的数量
训练集	136 621	2 170 848
测试集	11 504	172 602

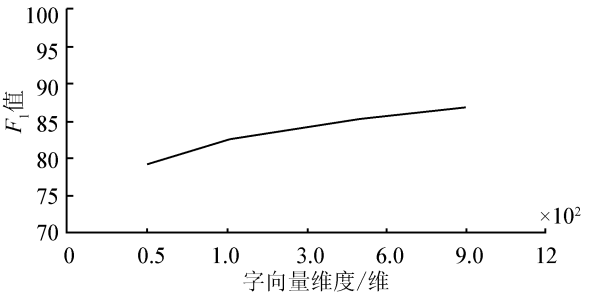

图 3 采用不同维度字向量的实验结果

Fig.3 Experimental Results with Different Vector Sizes

实验发现特征向量的维度对模型性能有较大的影响。增加维度可以提供更多信息,系统的性能也相对较好。实验表明即使没有其他词典或者词性特征参与,深层神经网络也能够较好地实现中文命名实体识别任务。

第 2 组实验中,将字典特征添加到汉字特征向量中。我们采用的字典如表 2 所示。

表 2 字典描述

Table 2 Dictionary descriptions

词典编号	词典名称
1	中文常用姓词典
2	中文不常用姓词典
3	中文常见人名词典
4	外国人名常用字词典
5	称谓词表
6	中国常见人名词典
7	地名后缀词典
8	地名词典
9	组织机构后缀词典
10	组织机构词典
11	单字词词典
12	常用多字词典

表 3 描述了字典特征的细节。三组特征分为:基础特征包含了字特征的基本信息;第二组包含了前后缀的信息;最后一组包含已知的名实体信息。

词是汉语句法和语义的基本单位。同一个字出现在单字词或者不同的组合词中,含义可能不同。例如,汉字“行”作为单字词的意思是表示赞同的意思,但是出现在“银行”和“行动”中意思就完全不同

了。通过观察发现字处于词语的不同位置时,通常会表现出不同的句法和语义属性。

基于以上观察,我们采用了基于词边界的字向量表示法。我们使用 4 种标记来描述单个字在词中所处的位置,汉字 x 可以转化为以下 4 种:x-B,x-M,x-E 和 x-S。例如,分词后的句子"去/哈尔滨/看/冰雕",采用词边界表示,可转化为"去 S/哈 B 尔 M 滨 E/看 S/冰 B 雕 E",我们基于上述带词边界信息的文本来构建字向量。对于汉字 x,其特征向量由基于 x-B、x-M、x-E 和 x-S 构建的特征向量连接起来,如下所示:

$$\mathrm{LT_c}(x) = [\mathrm{LT_c}(x-B), \mathrm{LT_c}(x-M), \mathrm{LT_c}(x-E), \mathrm{LT_c}(x-S)]$$

表 3　中文命名实体识别的字典特征

Table 3　Dictionary features for Chinese named entity recognition

特征集	特征
基本特征	$X_n(n=-2,-1,0,1,2)$, $X_0 \in D_{11}$
	$X_nX_{n+1} \in D_{12}(n=-1,0)$,
	$X_{n-1}X_nX_{n+1} \in D_{12}(n=-1,0,1)$
	$X_{n-2}X_{n-1}X_nX_{n+1} \in D_{12}(n=-1,0)$,
	$X_n \in D_1(n=-2,-1,0)$
	$X_n \in D_2(n=-2,-1,0)$,
	$X_n \in D_3(n=-2,-1,0,1,2)$
	$X_n \in D_4(n=-2,-1,0,1,2)$
	$X_n \in D_7(n=0,1,2)$,
	$X_n \in D_9(n=0,1,2)$
	$X_{n-1}X_n \in D_5(n=-1,0,1,2)$,
	$X_nX_{n+1} \in D_7(n=0,1)$
前后缀特征	$X_nX_{n+1} \in D_9(n=0,1)$,
	$X_{n-2}X_{n-1}X_n \in D_5(n=0,1,2)$
	$X_nX_{n+1}X_{n+2} \in D_7(n=0)$,
	$X_nX_{n+1}X_{n+2} \in D_9(n=0)$
	$X_{-2}X_{-1}X_0X_1 \in D_5$,
	$X_{-2}X_{-1}X_0X_1X_2 \in D_5$
名实体特征	$X_{-i}\cdots X_0\cdots X_j \in D_6$,
	$X_{-i}\cdots X_0\cdots X_j \in D_8$
	$X_{-i}\cdots X_0\cdots X_j \in D_{10}$

表 4 显示了基于词边界字向量的模型使用不同词典特征的实验结果。采用的字典特征与第一组实验中的字典特征相同。可以明显看出右侧一列基于词边界向量的实验结果,相较于左侧基于字向量的实验结果有了明显的提高。增加了字典特征以后,我们发现对于基于词边界模型的实验结果提高并不明显。说明部分词典特征已经在词边界模型中自动提取,因此独立加入词典特征对于模型的影响较小。

表 4　中文命名实体识别的实验结果

Table 4　Experimental Results of Chinese Named Entity Recognition

模型	中文命名实体识别 F_1/%
条件随机场(字特征)	84.60
SIGHAN 2006 封闭测试[22]	86.51
SIGHAN 2006 开放测试[22]	91.18
Deep CNN	89.18

对比其他实验模型,如表 5 所示,分析第二组实验结果可以看出,在不包含任何词典特征的情况下,基于词边界字向量的深度神经网络模型的 F_1 值比基于基本字向量的深度神经网络模型提高了 1.5%,比条件随机场模型提高了 3.1%,优于 SINHAN 2006 封闭测试的最优结果。加入词典特征以后,模型的预测性能得到提高,其中第一组基本特征的作用最大,但是词典特征对于深度神经网络模型的作用不如条件随机场模型(如 SINHAN 2006 开放测试的最优模型)。分析原因是由于深度神经网络将两种不同类型的特征(字向量为高维的实数向量,词典特征为 101 维的布尔向量)直接串联作为输入,特征不能较好的融合;特别是词典特征加入到基于词边界字向量对系统的提高,没有单纯使用字向量的时候显著。究其原因,是词边界字向量的维度过高,对词典特征有较大的稀释作用。而条件随机场模型使用的是字、词和字典特征等离散特征的组合,由人工选择并通过实验进行了调整。

表 5　不同模型中文命名实体识别的实验结果

Table 5　Experimental Results of Chinese Named Entity Recognition Based on different model

Deep CNN	基本字向量	基于词边界的字向量
+字向量	86.72	88.31
+ Basic Features	87.76	88.87
+Prefix and Suffix Features	87.97	89.10
+ Named Entity Features	88.14	89.18

3　结束语

针对传统序列标注方法普遍需要人工选择特征的问题,提出了一种基于词边界字向量的深度神经网络模型,并将其应用于中文命名实体识别问题。中文命名实体识别问题的实验结果表明:基于词边界字向量的深度神经网络模型优于基于基本字向量的深度神经网络模型和 SINHAN 2006 封闭测试的最优模型。

未来的工作中将从以下几方面考虑如何进一步提高深度神经网络的识别性能。首先,考虑如何能够较好地使用其他统计特征与字向量特征相融合。其次,目前的方法是从句首开始将每个字输入到神

经网络中,一旦命名实体的左边界词出现识别错误,会对整句的识别带来较大影响。下一步,我们考虑结合反向输入,以避免对命名实体的左边界词的识别率要求较高的问题。

参考文献:

[1] BENDER O, OCH F J, NEY H. Maximum entropy models for named entity recognition[C]//Proceedings of 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg, USA, 2003, 4: 148-151.

[2] WHITELAW C, PATRICK J. Named entity recognition using a character-based probabilistic approach[C]//Proceedings of CoNLL-2003. Edmonton, Canada, 2003: 196-199.

[3] CURRAN J R, CLARK S. Language independent NER using a maximum entropy tagger[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg, USA, 2003, 4: 164-167.

[4] CHIEU H L, NG H T. Named entity recognition: a maximum entropy approach using global information[C]//Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg, USA, 2002, 1: 1-7.

[5] KLEIN D, SMARR J, NGUYEN H, et al. Named entity recognition with character-level models[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL. Stroudsburg, USA, 2003, 4: 180-183.

[6] FLORIAN R, ITTYCHERIAH A, JING Hongyan, et al. Named entity recognition through classifier combination [C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg, USA, 2003, 4: 168-171.

[7] MAYFIELD J, MCNAMEE P, PIATKO C. Named entity recognition using hundreds of thousands of features[C]// Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg, USA, 2003, 4: 184-187.

[8] KAZAMA J, MAKINO T, OHTA Y, et al. Tuning support vector machines for biomedical named entity recognition [C]//Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain at ACL. Stroudsburg, USA, 2002, 3: 1-8.

[9] SETTLES B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA). Geneva, Switzerland, 2004: 104-107.

[10] WONG F, CHAO S, HAO C C, et al. A Maximum Entropy (ME) based translation model for Chinese characters conversion[J]. Journal of advances in computational linguistics, research in computer science, 2009, 41: 267-276.

[11] YAO Lin, SUN Chengjie, WANG Xiaolong, et al. Combining self learning and active learning for Chinese named entity recognition[J]. Journal of software, 2010, 5(5): 530-537.

[12] COLLOBERT R. Deep learning for efficient discriminative parsing[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Lauderdale, USA, 2011: 224-232.

[13] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(6): 1137-1155.

[14] 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.

[15] SCHWENK H. Continuous space language models[J]. Computer speech & language, 2007, 21(3): 492-518.

[16] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model[C]//Proceedings of 11th Annual Conference of the International Speech Communication Association (INTERSPEECH). Makuhari, Chiba, Japan, 2010, 4: 1045-1048.

[17] MNIH A, TEH Y W. A fast and simple algorithm for training neural probabilistic language models[C]//Proceedings of the 29th International Conference on Machine Learning (ICML-12). Edinburgh, Scotland, UK, 2012: 1751-1758.

[18] BOTTOU L. Stochastic gradient learning in neural networks [C]//Proceedings of Neuro-Nîmes 91. Nimes, France, 1991.

[19] TURIAN J, RATINOV L, BENGIO Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010: 384-394.

[20] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia, 2013: 746-751.

[21] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. California, USA, 2013.

[22] LEVOW G A. The third international Chinese language processing bakeoff: word segmentation and named entity recognition[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. Sydney, Australia, 2006: 108-117.

作者简介:

姚霖,1975年生,高级工程师,主要研究方向为生物信息、自然语言处理。主持和参与多项科研项目。发表学术论文20余篇。

刘轶,1972年生,研究员,主要研究方向为语音识别、多媒体信息处理、嵌入式软件及系统,主持和参与国家自然科学基金等项目几十项。发表学术论文50余篇,其中被SCI检索6篇,EI检索22篇。

刘宏,1967年生,教授,博士生导师,主要研究方向为软硬件协同设计、计算机视觉与智能机器人、图像处理与模式识别。发表学术论文50余篇。