基于神经网络的中文姓名抽取技术

吴芬芬，刘 磊
(吉林大学 计算机科学与技术学院，长春 130012)

摘要：设计了一个中文姓名抽取系统，该系统采用神经网络进行汉语句子的分词处理，根据姓名后置特征词进行姓名的抽取，成功解决了尾字和下文成词的姓名抽取问题. 以1998年1月份《人民日报》语料库中含有此类姓名的语句作为测试数据，结果表明，姓名抽取的召回率和精确度较现有方法都有很大提高.

关键词：姓名抽取；神经网络；特征提取
中图分类号：TP391 **文献标识码**：A **文章编号**：1671-5489(2006)03-0411-04

Extraction Technology of Chinese Names Based on Neural Network

WU Fen-fen, LIU Lei
(*College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*)

Abstract: An extraction system of Chinese names was designed. The system adopts the neural network to deal with the Chinese word segmentation, then carries on extraction of name by means of rearmounted characteristic word according to name. The system solves the question of extracting Chinese name successfully whose tail character and the following character construct a word. With the sentences with this kind of name in "People's Daily"'s corpus base of January of 1998 the testing data, the result shows that the recall and precision rates are all improved a lot compared with existing methods.

Key words: name extraction; neural network; character extraction

信息抽取是指从文本中抽取指定的某一类信息，将其形成结构化数据，它的主要任务是识别命名实体，确定语义关系. MUC(Message Understanding Conferences)将命名实体定义为人们感兴趣的专有名词和特定的数量词，它一般分为人名、地名、组织机构名、日期等类型.

针对中文姓名的识别问题，根据其使用方法的不同，现有的方案分为三种类型：规则法、统计法[1,2]以及规则统计相结合[3]的方法，这些方法都基于大规模语料库[4]. 在这些方法中，大量规则的制定，代价昂贵而且不易扩展；姓名候选字段大都是选取切分后的单字碎片，也有研究者将少量的二字或多字词纳入候选字段的选取范围，在这种选取机制的作用下，尾字与下文成词的姓名基本上无法识别，例如在文献[5]中，采用词语粗分和统计的方法对人名进行识别，就不能解决这个问题. 这些方法的统计数据都基于静态语料库，不具有动态性.

对于尾字和下文成词的姓名抽取问题，传统的基于词典分词和统计的方法不能很好地解决. 例如：王比学报道，姓名"王比学"的尾字"学"和下文组成词语"学报"，基于词典的分词结果为：王/比/学报/道，不能对姓名正确提取. 本文运用特征提取和神经网络相结合的技术，舍弃词典，改用神经网络分词，继而根据中文姓名后置特征词进行姓名抽取，成功地解决了这个问题. 我们使用的神经网络

收稿日期：2005-09-05.

作者简介：吴芬芬(1979～)，女，汉族，硕士研究生，从事信息抽取技术的研究. 联系人：刘 磊(1960～)，男，汉族，教授，博士生导师，从事程序理论、形式化方法、编译技术、语义网和计算语言学等研究，E-mail：liulei@jlu.edu.cn.

基金项目：吉林省科技发展计划项目基金(批准号：20050527).

方法，具备学习功能强、开放性好以及分词速度快、精确度高等特点，可以大大地提高姓名抽取的召回率和精确度.

1 神经网络简介

BP 模型的结构如图 1 所示.

BP 算法的执行步骤如下：若每层有 n 个神经元，即 $i=1,2,\cdots,n$；$j=1,2,\cdots,n$，对于第 k 层的第 i 个神经元则有 n 个权系数 $w_{i1},w_{i2},\cdots,w_{in}$，另外多取一个 $w_{i,n+1}$ 用于表示阈值 θ_i，并在输入样本 X 时，取 $X=(x_1,x_2,\cdots,x_n,1)$ 以及对应期望输出 $Y=(Y_1,Y_2,\cdots,Y_n)$.

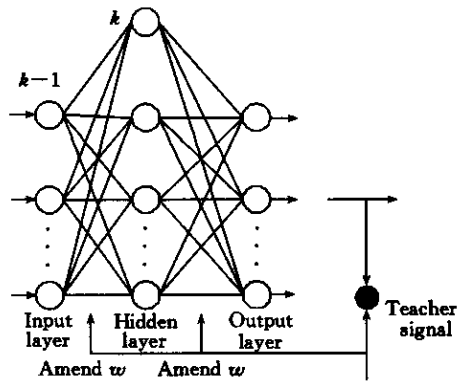

Fig. 1 BP model

(1) 对各层的权系数 w_{ij} 置一个较小的非零随机数，其中 $w_{i,n+1}=-\theta_i$；

(2) 读入一个向量对 (X,Y)；

(3) 计算各层的输出，对于第 k 层第 i 个神经元的输出 x_i^k，有：

$$u_i^k=\sum_{j=1}^{n+1}w_{ij}x_j^{k-1},\qquad x_i^k=f(u_i^k),\qquad f(x)=\frac{1}{1+\exp(-x)};$$

(4) 求各层的学习误差 d_i^k：对于输出层 $k=m$，有 $d_i^m=x_i^m(1-x_i^m)(x_i^m-Y_i)$；对于其他各层，有 $d_i^k=x_i^k(1-x_i^k)\sum_l w_{li}d_l^{k+1}$；

(5) 修正权系数 w_{ij} 和阈值 θ_i：$w_{ij}(t+1)=w_{ij}(t)-\eta d_i^k x_j^{k-1}$，其中 η 为学习效率，t 为时间标志；

(6) 当求出各层各个权系数之后，可按给定品质指标判断是否满足要求，如果满足要求，则算法结束；如果未满足要求，则返回(3)执行.

2 中文姓名抽取系统

2.1 分 词

以 1998 年 1 月份《人民日报》作为训练和测试的语料，对其进行统计，找出尾字和下文成词的姓名和其后置词组成的短语，放入表中，形成样本空间表. 例如：姓名“江泽民”的尾字“民”和下文“主”组成词“民主”，将短语“江泽民主席”放入样本空间表中，作为训练神经网络的语句. 用于训练神经网络的样本空间示例如下：①江泽民主席；②王继宣教授；③董现明主任；④吴邦国会见；⑤王比学报道；⑥郑必坚强调；⑦李长水上任；⑧李长水担任；⑨毛泽东部署；⑩周恩来生平.

2.1.1 语句预处理 为了使神经网络能够接受外部数据，首先要对读入的汉语句子进行预处理，把句子中的每个汉字映射成神经网络模型能够接受的数字化输入，即汉字编码. 具体的编码方式是利用汉字的机内码来构成多个汉字的输入神经元初值，一个汉字的机内码占两个字节，将其化成二进制形式，变成神经网络可以识别的格式.

2.1.2 建立样本表、特征词表和姓氏表 对样本空间表中的每个样本语句均进行汉字编码，并给出其切分的期望输出，形成输入输出向量对，存入样本表中；统计样本空间表中的每条语句，把姓名后置特征词“主席”、“教授”、“主任”等放入另一表中，组成姓名的后置特征词表；统计中华百家姓和语料库中的姓氏，组成姓氏表.

例如：王继宣教授的汉字编码为如下向量 X：

1100110111110101101111001100 1101000011111011011110111001100110010101011010，

对应的期望输出 Y 为：00101000. 把向量对 (X,Y) 存入样本表，作为神经网络训练的输入，将特征词“教授”存入特征词表，用于特征提取.

2.1.3　**训练神经网络**　BP 模型的主要参数如下：

（1）输入层节点数：每个汉字用 16 位表示，若限定句子的长度为 n 个字，则神经元的输入节点数为 $16n$，在本文样本空间中，最长句子的汉字个数为 8，因此输入层节点数为 16×8 个；

（2）隐层节点数：一般比输入层神经元数目少，但是不能过少，否则将限制神经网络存储各种模式的能力，可以考虑设为：$16na$（a 为 0.2～0.8），经过多次调试，在本系统中，取 a 为 0.5 即可学习所有的样本；

（3）输出层节点数：输出层节点数与输入句子的汉字最大个数相等. 本系统选 8 个节点.

神经网络的训练过程如下：训练开始时，将内部连接权，阈值初始化，并根据实验结果确定隐含层数及 a 的值，网络各单元之间的连接权及单元阈值随机赋予. 每给网络提供一个输入输出向量对，首先要进行前向传播并计算出各单元的实际输出，求出各单元的参考误差. 当各单元的参考误差都求出后，进行连接权和各单元阈值的调整，从而完成一项迭代. 取下一向量对，重复上述过程. 当样本表中的所有向量对对各自的迭代都完成后，又重复对第一向量对的迭代. 这样循环下去，直到输出层单元的误差满足要求为止，并把这时的权值输出形成一个权值表（表 1）.

Table 1　Output of training of samples for 3 000 times

Sample 1	0.000 56	0.000 05	0.999 93	0.000 01	0.997 11	0.001 16	0.000 56	0.001 89
Sample 2	0.000 53	0.001 46	0.998 17	0.000 02	0.997 60	0.000 17	0.000 54	0.000 79
Sample 3	0.000 82	0.001 22	0.998 74	0.000 03	0.996 30	0.000 44	0.000 83	0.001 68
Sample 4	0.000 67	0.004 07	0.998 07	0.000 03	0.999 64	0.000 21	0.000 67	0.000 69
Sample 5	0.000 41	0.000 75	0.999 83	0.000 01	0.995 78	0.000 82	0.000 42	0.000 46
Sample 6	0.001 29	0.002 62	0.995 84	0.000 01	0.999 48	0.000 33	0.001 30	0.001 17
Sample 7	0.000 47	0.000 60	0.999 99	0.000 04	0.995 80	0.008 03	0.000 47	0.001 21
Sample 8	0.000 76	0.001 22	0.999 85	0.000 01	0.999 43	0.001 68	0.000 77	0.001 21
Sample 9	0.000 65	0.000 26	0.999 68	0.000 03	0.995 21	0.000 64	0.000 66	0.002 20
Sample 10	0.000 51	0.001 42	0.999 68	0.000 02	0.999 14	0.000 46	0.000 51	0.001 03

同时，对 BP 模型参数进行调整，并分析了其对分词结果的影响，结论如下：

（1）隐层的节点数增加时，要相应地增加网络的训练次数，否则分词可能出现错误，我们把隐层的节点数一般设置为输入层节点数的一半以下；

（2）隐层的层数增加时，也要相应地增加网络的训练次数，才不会影响分词的精度，但是这增加了网络的训练时间. 本系统选一层隐层，足够储存所有的输入模式；

（3）当加入新的未经训练的样本时，要重新对神经网络进行训练，这也说明神经网络可以通过重构实现自适应和自学习，分词模型是一个不断完善和发展的动态模型.

2.1.4　**切分过程和规范化**　根据姓氏触发策略，对读入的句子，先查询姓氏表，找到姓氏以后截取 3～8 个汉字，进行汉字编码，然后根据权值表和神经网络模型进行切分.

例如，截取后的句子为：江泽民主席. 汉字编码为：

1 0 1 1 1 1 0 1 1 0 1 0 1 1 0 1
1 1 0 1 0 1 0 0 1 1 1 1 0 0 1 1
1 1 0 0 0 0 1 1 1 1 1 1 0 0 0 1
1 1 0 1 0 1 1 0 1 1 1 1 0 1 1 1
1 1 0 0 1 1 1 1 1 0 1 0 1 1 1 1

切分的结果为：

0.000 621，0.000 064，0.999 881，0.000 018，0.996 612，0.000 731，0.000 548，0.002 006.

规范化是把 BP 模型得到的向量进行规整处理. 神经网络经过大量的学习训练，从某种意义上讲，具备一定的智能化，但它的结果不能用自然语言表达，其输出值为 0～1 之间的数据. 我们对其进行规整处理，即规整为 01 串，其中值大于 0.6 的视为 1，表示此节点对应的汉字与下一个相邻的汉字之间有切分标志，值小于 0.4 的视为 0，表示无切分标志，若值在 0.4～0.6 之间，则视为错误，语句不可用

神经网络学习. 依据上面的原则，切分结果的规范化输出为：00101000，0 为不切分，1 为切分，也就是江泽民/主席.

2.2 特征提取

首先，运用神经网络对含有姓名的短语进行切分，在此结果上，查询姓名后置特征词表. 如果切分出的第二个词出现在特征词表中，则第一个词为姓名，将其提取出来，例如：江泽民/主席，在姓名后置特征词表中查询，找到“主席”，则第一个词“江泽民”即为姓名；若在特征词表中，未能找到切分出的第二个词，则根据姓氏触发策略，继续寻找句子中含有姓名的短语，重复以上的分词和特征提取过程，直到输入句子结束.

3 实验结果分析

用 1998 年 1 月份《人民日报》语料库中的语句作为测试对象，以召回率和精确度作为评测标准，对含有尾字和下文成词的姓名的语句，进行姓名抽取，实验结果显示：召回率达到 100%，精确度达到 96.36%. 召回率和精确度的计算公式如下：

$$\text{召回率} = \frac{\text{识别出正确的姓名个数}}{\text{语料中符合条件的姓名个数}}, \qquad \text{精确度} = \frac{\text{识别出正确的姓名个数}}{\text{识别出所有的姓名个数}}.$$

导致精确度还有一定误差的原因如下：例如外经贸部部长吴仪会见了哈马迪一行.“长”和“吴”都可以作为姓氏出现，于是截取了“长吴仪会见”和“吴仪会见”两个句子，根据神经网络分词和特征提取后，出现了两个姓名长吴仪和吴仪，其中长吴仪为一个错误姓名，导致精确度下降.

当然，想要保持高的召回率和精确度必须不断扩大样本表和特征词表，这样才能保证神经网络能够正确地分词进而正确地提取姓名.

神经网络的分词系统所具有的学习机制，使它可根据用户的要求随意地增添或删除某些权重链接，以达到维护知识库的目的. 在神经网络中，允许输入偏离学习样本，只要输入模式接近于某一学习样本的输入模式，则输出亦会接近学习样本的输出模式，这种性质使得神经网络系统具有联想记忆的能力. 神经网络学习的过程是一个由简到繁、逐渐完成知识积累的过程.

综上，本文运用神经网络进行汉语句子的分词处理，继而根据姓名后置特征词，进行尾字和下文成词的中文姓名抽取. 进一步，我们会尝试把神经网络和统计技术相结合进行姓名的抽取，利用神经网络分词精确的特性和中文姓名前置词、后置词等的概率统计，建立一个中文姓名的抽取系统，并进一步改进 BP 算法，提高中文姓名识别的召回率和精确度.

参 考 文 献

[1] LIU Bing-wei, HUANG Xuan-jing. Statistical Chinese Person Names Identification [J]. Journal of Chinese Information Processing, 1999, **14**(3): 16-24. (刘秉伟, 黄萱菁. 基于统计方法的中文姓名识别 [J]. 中文信息学报, 1999, **14**(3): 16-24.)

[2] ZHUANG Ming, LAO Song-yang, WU Ling-da. A Named Entity Discovery Using Statistics-based and Pos-based Method [J]. Computer Applications, 2004, **4**(1): 22-24. (庄 明, 老松杨, 吴玲达. 一种统计和词性相结合的命名实体发现方法 [J]. 计算机应用, 2004, **4**(1): 22-24.)

[3] JI Heng, LUO Zhen-sheng. A Chinese Name Identifying System Based on Probability Distribution and Rules [J]. Applied Linguistics, 2001, **2**(1): 14-18. (季 姮, 罗振声. 基于统计和规则的中文姓名自动辨识 [J]. 语言文字应用, 2001, **2**(1): 14-18.)

[4] ZHENG Jia-heng, LI Xin. The Research of Chinese Names Recognition Method Based on Corpus [J]. Journal of Chinese Information Processing, 1999, **14**(1): 7-12. (郑家恒, 李 鑫. 基于语料库的中文姓名识别方法研究 [J]. 中文信息学报, 1999, **14**(1): 7-12.)

[5] ZHANG Feng, FAN Xiao-zhong, XU Yun. The Research of Chinese Names Recognition Method Based on Statistics [J]. Computer Engineering and Applications, 2004, **40**(10): 53-54. (张 锋, 樊孝忠, 许 云. 基于统计的中文姓名识别方法研究 [J]. 计算机工程与应用, 2004, **40**(10): 53-54.)

(责任编辑：赵立芹)