# 中文人名自动识别的一种有效方法①

李建华② 王晓龙

（哈尔滨工业大学计算机科学与技术系 哈尔滨 150001）

摘 要 介绍了一种基于大量实验的有效的中文姓名自动识别方法。实验结果表明，该方法在兼顾准确率与召回率的同时获得了较好的识别效果。

关键词 中文姓名自动识别，自动分词，知识源

## 0 引言

中文信息计算机自动处理的研究已有几十年的历史，但至今仍有许多技术难题没有得到很好解决，中文姓名自动识别问题就是其中的一个。由于它与中文文本的自动分词一样，属于中文信息处理的基础研究领域，因而它的研究成果直接影响到中文信息的深层次研究。汉语的自身特点使得中文信息自动处理大多是先对要处理的文本进行自动分词（加入显式分割符），然后再在分词的基础上进行词法、语法、语义等方面的深入分析。而在分词阶段，文本中的人名、地名以及其它专有名词和生词大多被切分成单字词，在这种情形下如不能很好地解决汉语文本中专有名词生词的识别问题，将给其后的汉语文本的深入分析带来难以逾越的障碍。中文姓名的自动识别问题就是在这种背景下提出来的。对这一问题的研究目前采用的技术中主要利用以下几方面的信息：姓名用字的频率信息、上下文信息[1,2]、语料库统计信息[2]、词性信息等[3]。本文的方法是，首先对中文人名的构成、姓名用字的规律及上下文文本信息特征进行充分分析，在此基础上建立起两组规则集，将其作用于测试文本，获得初步识别结果，再利用大规模语料库的统计信息对初步识别结果进行概率筛选，设定合适的阈值，输出最终识别结果。经对 50 多万字的开放语料测试，系统自动识别出 1781 个中文人名，在不同的筛选阈值下获得 90%以上的识别准确率，而召回率高于 91%。

## 1 规则的施加与松弛

### 1.1 中文姓名的构成规律

中文姓名一般由二字或三字组成，第一字为姓氏字（复姓为前两字），其后的一到两个汉字为名用字。统计表明，中文姓名在用字上也有一定规律：一方面某些字频频出现在姓名中，如在姓氏用字中，虽然姓氏辞典中列举了几千个姓氏字，但目前实际使用的不过几百个，而张、王、李、赵、刘 5 个姓竟占了 32%[1]；另一方面，某些字又从不被用作姓名用字，如最、仅、紧、以、且等字。根据这一特性，首先从一个含有 1 万多个人名的数据库中抽取 303 个姓用字和 1047 个名用字，形成系统的知识源；然后根据姓名的构成原则制定了一组姓名构成规则集，其中的规则以姓氏字驱动。由于中文姓名的构成是严格遵守构成规则的，因而本文将姓名构成规则定义为一组必须匹配的严格规则。

### 1.2 中文姓名存在的上下文环境分析

中文姓名在文本中不是孤立存在的，其依存的上下文信息具有一定的特点。

（1）前置信息：姓名的前端多冠有对人的职业、职务及与说话人的关系的称谓，如“这是上海市副市长刘振元日前在与上海旅游记者协会座谈时介绍的。”、“我和妻子秦润英都是双目失明的盲人。”等。在上述句子中的“市长”和“妻子”就是人名“刘振元”和“秦润英”的前置提示信息。

（2）后置信息：姓名的后端多随有对此人的职业、职务及与说话人的关系的称谓，如“我国著名学者彭明教授访问前苏联时将书稿复印件全文带回。”，这里的“教授”就成为人名“彭明”的后置提示信息。

（3）提示动词：某些动词多随在姓名和人称代词后，如“说、指出、告诉、通知...”，可充分利用这些词的提示作用。

根据这些特点，本文依据《同义词词林》第二

① 863 计划资助项目（863-306-ZT03-02-3）。

② 女，1965 年生，博士生；研究方向：人工智能，自然语言理解，文本校对；联系人。

（收稿日期：1998-12-17）

版[4] 建立起中文人名上下文环境资源信息表，及与此资源表对应的一组规则集，将其定义为姓名环境规则集，此规则集不是姓名识别过程中必须严格匹配的规则。

1.3 规则的施加与松弛

在进行自动识别中文姓名的工作之前，首先对人工识别姓名的过程进行了大量观察，发现人在对文本中的人名进行判定时，是以姓氏字驱动的，即最先找到姓氏用字，再利用其后跟随的局部信息，观察其是单字还是词，是单字则此字是否是常用的名用字，是词则看它的词长是否超过2，组成此词的字是否是常用的名用字；如仍无法断定，则只有增加信息量，依靠上下文的全部信息来判断。鉴于上述对人的认识过程的认识，本文利用在1．1和1．2中建立的知识源，首先将两组规则集同时施加到测试语料上，经过规则的严格匹配，将识别出的结果送入姓名第一候选集，留给后续的概率识别器，进行进一步识别。对在规则匹配过程中无法同时满足两组规则的文本，松弛姓名环境规则集的限制，而只对其施加姓名构成规则，得到一组识别结果，送入姓名第二候选集，供概率识别器进一步识别。实验结果表明，姓名第一候选集中的识别结果，姓名识别的准确率可高达96.81%，而姓名第二候选集的识别准确率在不经概率识别器进一步识别的情况下仅为64.24%（实验数据及结果参见本文第三部分）。因此，从整体上，看规则识别阶段产生的识别结果只能作为识别的粗选集，提高识别的准确率必须采取其它识别手段，本文采用的是概率识别器。

## 2 引入概率识别器

引入概率识别器就是充分利用姓名用字的规律性信息。通过对大规模语料库的统计，得到姓名用字规律的量化信息，使用这些量化后的规律信息，对由规则识别出的粗选结果进行最后识别，选定合适的阈值，输出最终的识别结果。实验证明，概率识别器的引入大大提高了粗选结果的识别准确率，在对50多万字的测试语料进行的实验中，系统的识别准确率平均可达90%以上。

在介绍系统使用的概率识别器的概率模型之前，首先进行如下定义：

令 *name* 代表一个可能成为姓名的字串，*name*＝*W*1、*W*2、*W*3，其中 *W*1、*W*2 和 *W*3 为组成该字串的单字（本文中暂不考虑复姓的人名识别问题）。*P*（*name*）为 *name* 字串成为中文姓名的概率，*p*（*Wi*）为单字 *Wi* 作为中文姓名用字出现的频率，则：

$$P(name) = p(W1) * p(W2) * p(W3);$$

$$p(Wi) = \lambda f(Wi) * d(Wi);$$

*f*（*Wi*）为 *Wi* 作为姓名用字出现的频率，*d*（*Wi*）为 *Wi* 在姓名中使用的频率，λ是概率调整系数。

*f*（*Wi*）＝*Wi* 作姓（或名）用字出现的次数/*Wi* 出现的总次数；

*d*（*Wi*）＝*Wi* 在姓（或名）中使用的次数/姓名中的所有姓（或名）使用的总次数。

考虑到姓名中大量存在单名情况下，此时的 *W*3 并不存在，因此完整的概率识别模型为：

$$P(name) = p(W1) * p(W2) * p(W3)$$

通过对40多万字的训练语料统计，“刘”字在语料中出现了208次，而这208次均是以姓氏字出现的，因此 *f*（“刘”）＝208/208＝100%；类似地，*f*（“李”）＝679/683＝99.41%；*f*（“雷”）＝44/114＝38.60%。*d*（“刘”）＝208/1528＝0.13612；*d*（“李”）＝0.44725；*d*（“雷”）＝0.02917。与以往的概率筛选模型[1,2] 不同的是，这里加入了 *f*（*Wi*）函数，即姓名用字的规律信息，实验结果表明，它的加入大大提高了识别准确率。实验中当对粗选结果应用单一概率模型（即仅使用 *f* 或 *d* 函数中的一项）进行筛选时，获得了约86%的识别准确率，而使用了上述概率模型（即同时使用 *f* 和 *d*）后，识别的准确率提高到91%。

进行大规模语料统计必然遇到的问题是数据的稀疏，在本系统中对稀疏数据采用的平滑策略是：将语料中未出现的数据的出现次数分配为1。

## 3 实验结果及分析

### 3.1 实验结果

实验使用 Visual C++6.0 在586微机上进行，

训练语料为人民日报新闻语料，训练规模为 40 余万字，测试语料为 50 余万字的人民日报新闻语料。实验先对测试语料按最大匹配法进行分词，在经过分词的语料上施加规则并引入概率识别器，结果，在未经概率筛选的情况下系统识别出 1781 个中文姓名，在不同的概率阈值下获得的识别准确率和召回率的结果如表 1 所示。令：

识别准确率＝识别出的姓名中真正为中文姓名的比例；

召回率＝文本中的中文姓名被识别到的比例；

表 1　不同概率阈值下获得的识别准确率和召回率

| 概率阈值 | 识别出的姓名个数 | 误识的中文姓名个数 | 误识率/% | 准确率/% | 没有识别出的姓名个数 | 召回率/% |
|---|---|---|---|---|---|---|
| 0.000125 | 1374 | 171 | 12.45 | 87.55 | 55 | 95.63 |
| 0.0003 | 1336 | 150 | 11.23 | 88.77 | 62 | 95.03 |
| 0.0006 | 1285 | 130 | 10.12 | 89.88 | 99 | 92.06 |
| 0.00085 | 1273 | 125 | 9.82 | 90.18 | 106 | 91.55 |
| 0.001 | 1264 | 123 | 9.73 | 90.27 | 112 | 91.06 |
| 0.002 | 1209 | 101 | 8.35 | 91.65 | 146 | 88.36 |

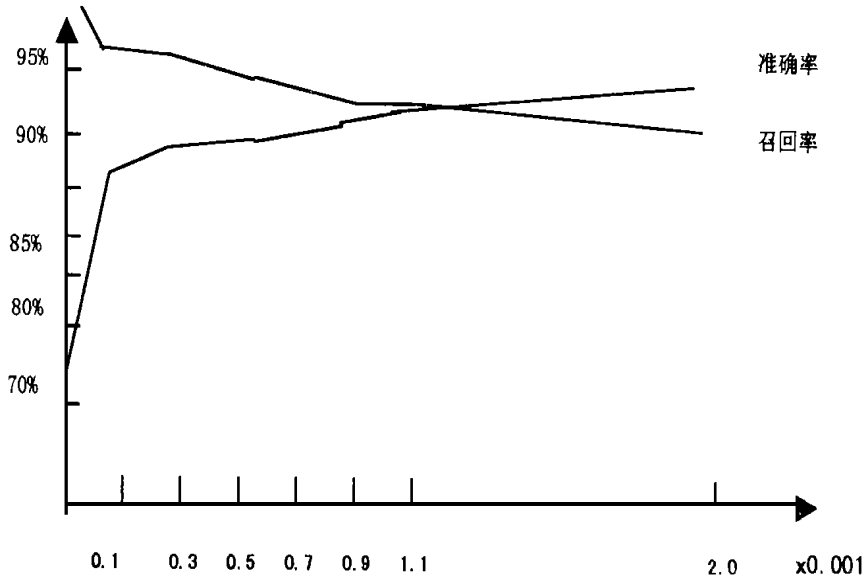


图 1　准确率与召回率比值曲线图

由表中数据可以看出，随着阈值的增大，系统识别的准确率在不断升高，而召回率却呈下降的趋势。准确率与召回率是一对互相矛盾的性能指标，为直观地反映它们之间的平衡关系，下面依据表 1 数据画出它们的曲线图（见图 1），找到二者的平衡点。

从图 1 可见，当阈值大于 0.001 时，识别准确率的升高和召回率的下降有一汇合处，汇合处的准确率和召回率均可超过 90%，此点的阈值可作为系统的最佳阈值。

3.2　部分实验数据

许贵元－0.083333－0.388350－0.027650－1.841244－4.765309－3.573981－－0.028066

黄翊明－0.277457－1.000000－0.139216－9.819967－0.238265－8.458423－－0.764433

江泽民－0.446708－0.997050－0.139130－58.306056－40.266857－36.216345－－5268.979243

贾西平－0.850000－0.037037－0.334656－3.477905－2.740052－30.140577－－3.025956

吴铁洪－0.941176－0.107527－0.393443－22.913257－3.573981－2.859185－－9.322875

韩天喜－0.346154－0.018895－0.032258－1.841244－1.548725－0.357398－－0.000215

易定刚－0.046154－0.020606－0.102564－1.841244－2.025256－0.476531－－0.000174

金利来－0.007117－0.013986－0.016637－0.818331－1.191327－2.263522－－0.000004

……

3.3 部分识别正确的实例

李鹏/同志/最近/分别/为/测绘/职工/题词/，/

新华社/记者/刘彦武/摄/

高山/堡/乡/天/兴/元/村/村民/马军/夫妇/抚养/着/刚/出生/D/个/月/的/婴儿/

下/半/时/广州/队/分别/由/黄伟雄/和/彭伟国/在/第/D/分 钟/和/D/分 钟/各/攻/入/一/球/，/

崔光日/是/这个/队/的/队员/。/

……

3.4 部分识别错误的实例

你们/地地道道/的/高水平/生物/工程/成果/却/在/研究室/里/睡/大/觉/。/

有/一/次/孩子/说/要/送给/老师/一/张贺年/卡/。/

……

3.5 错误识别产生的原因

中文姓名的误识主要由以下几个原因造成:

(1) 知识源的不完备，这主要是指在1.1中建立的姓氏用字和名用字知识源的不完备。对于在姓氏字表中不包含的姓氏，系统则无法识别出以此字为姓氏字的中文姓名。名用字也有类似的情况。

(2) 在姓氏字后紧随的单字或二字词是一个常被用作人名的字，因此无论通过规则还是通过概率筛选均无法将其排除掉。如句子“有/一/次/孩子/说/要/送给/老师/一/张贺年/卡/。/”，“张贺年”的概率识别结果为:

“张贺年－0.725857－0.021053－0.003142－47.667758－0.238265－1.548725－－0.000844”

③ 人名连续书写，之间没有明显的切分标志，因此会产生类似于自动分词中的切分歧义问题。如在句子“记者李明安纪才报道”中，可将“李明安”和“纪才”各切为一个人名，也可将“李明”和“安纪才”各切为一个人名，这样就可能造成识别错误。但这种误识并非真正的系统错误，因为即使是用人工来识别，也未必能将它很好地区分开。

## 4 结论

本文描述了一种有效的中文人名识别方法。它的基本原理是在大规模语料统计的基础上，利用知识源在文本上进行规则的施加与松弛，并引入概率分析器来提高识别的准确率和召回率。实验结果表明，在兼顾识别的准确率与召回率的情况下，系统取得了良好的效果。

参考文献:

[1] 孙茂松, 黄昌宁, 高海燕等. 中文姓名的自动辨识. 中文信息学报, 1995, 19(2)

[2] 张俊盛, 陈舜德, 郑萦. 多语料库做法之中文姓名辨识. 中文信息学报, 1992, 16(3)

[3] 郑家恒, 谭红叶. 基于变换的中文姓名识别技术探讨. 中文信息协会第八次国际会议论文集, 1998

[4] 梅家驹, 竺一鸣, 高蕴琦等. 同义词词林. 上海: 上海辞书出版社, 1996

# An Effective Method on Automatic Identification of Chinese Name

Li Jianhua, Wang Xiaolong

(Department of Computer Science, Harbin Institute of Technology, Harbin 150001)

**Abstract**

An effective method on automatic identification of the Chines name, which is on the basis of a large amount of experiments, is introduced. It is shown that a satisfied result has been achieved considering the accuracy and the recall rate at the same time.

**Key words**: Automatic identification of Chines name, Automatic word segmentation, Resource of knowledge