中 文 信 息 学 报
第9卷 第2期 JOURNAL OF CHINESE INFIOMATION PROCESSING Vol.9 No.2

中文姓名的自动辨识②

孙茂松 黄昌宁 高海燕+ 方 捷

(清华大学计算机科学系 +烟台大学计算机应用系)

【摘要】中文姓名的辨识对汉语自动分词研究具有重要意义。本文提出了一种在中文文本中自动辨识中文姓名的算法。我们从新华通讯社新闻语料库中随机抽取了300个包含中文姓名的句子作为测试样本，实验结果表明，召回率达到了99.77%。

关键词：中文姓名自动辨识，生词处理，汉语自动分词，中文信息处理

一、引言

汉语自动分词是中文信息处理的基础课题之一。虽已有十年的研究历史，但始终未见真正实用的系统面世。困扰此项研究的困难主要有二：歧义切分问题与生词处理问题。

本文所讨论的属于第二个问题的范畴之内。中文姓名一般来说具有任意性，或典雅浑成，或粗俗率真，或简单明了，或深奥晦涩，洋洋洒洒，繁若星沙，随心所欲，实难预期。无论分词词典如何庞大，都不可能用穷举的办法将它们囊括进去（著名人物除外）。中文姓名在文章中的出现频率虽然不高，但绝非可以忽视。由于中文姓名不象印欧语言那样可以通过大写字母来辨识，其中的姓氏和名字用字不少又可同时以普通词或普通词一部分的身分参与句子的活动，因此如果不予处理，将导致为数可观的分词错误。例如：

郑杰士来时遇见了林红。 (例1)
刘清楚楚动人。 (例2)

利用从左向右扫描的最大匹配法进行切分，得到：

郑 杰 士 来 时 遇见 了 林 红。

刘 清楚 楚 动人。

注意，例1中姓名"郑杰士""林红"被生硬地拆裂成字串。例2中因未认为"刘清"是一个姓名，导致分词错误蔓延。由此可见，中文姓名的自动辨识对建立一个健壮的自动分词系统具有重要意义。

①本文1994年9月2日收到
②清华大学智能技术与系统国家重点实验室开放基金资助项目

迄今为止，这方面研究见诸报道的，主要有文献[1,2,3]。[1]完全以中文姓名的概率作为辨识依据，并将之融入限制式满足求解过程中。虽然取得了令人注目的结果，但存在两点局限：(1) 姓名语料库偏小，仅包含 18541 个姓名样本。一般认为，这类语料库需要 10 万个以上姓名样本才算足够大；(2) 仅利用了关于姓名的统计信息，而没有充分挖掘其它各种行之有效的手段。[2],[3]则进一步从不同的角度进行了有益的尝试。

二、辨识中文姓名的当用资源

2.1. 姓氏频率表 XFL 与名字用字频率表 MCFL

中文姓名由姓氏和名字两个部分组成。表 XFL 和 MCFL 取自[4]，系根据 1982 年中国全国人口普查资料，使用计算机对 174，900 个中文姓名进行抽样综合统计的结果。姓名样本分别从中国六大区的七个省市抽取（北京、上海、辽宁、陕西、四川、广东和福建，各随机抽取 25，000 个左右），覆盖面广，代表性强，比较合理、科学。统计结果显示:

(1) 当今仍然使用、活跃的中文姓氏远没有某些姓氏典籍所列举的那么多。统计共得到 729 个姓氏，虽遗漏固属难免，但姓氏数目的量级大致应是这样。宋《姓解》收录了姓氏 2568 个，明《万姓统谱》收录了 3557 个，大陆出版的《中华大词典》收录了 1942 个，台湾出版的《中国姓氏集》收录了 5544 个，其中绝大部分已成"死"姓，不复流行；

(2) 姓氏分布很不均匀，但相对集中。729 个姓氏中，"王、陈、李、张、刘"这 5 大姓就占了姓名样本数的 32.0%，前 14 个姓占 49.5%，前 114 个姓占 90.0%，前 365 个姓占 99.0%。而其余 364 个姓氏仅占不到 1.0%。中国俗语所谓"张王李赵刘，普遍天下走""陈蔡李林王半天下"，即是这种分布趋势的形象反映；

(3) 某些姓氏可用作单字词，其中不乏高频单字词。常用的姓氏如"王、黄、马、高、于"等，不常用的姓氏如"是、过、来、从、那"等；

(4) 名字用字分布较姓氏要平缓、分散。共得到 3345 个名字用字，频率最高的前 6 个字("英、华、玉、秀、明、珍") 的覆盖率为 10.4%，前 28 个字为 30.1%，前 410 个为 90.0%，前 1141 字为 99.0%；

(5) 名字用字涉及范围很广。从所属的词类看，不仅有实词，也有各类虚词。如副词"再、常、太、必、就、非、最、更、也、极、又、仅、皆"，介词"以、向、从、于、把"，连词"而、虽、且、与"等。从感情色彩看，多使用褒义字和中性字，但也出现了一些贬义字或不太雅的字，如"虫、狗、鸡、狼、愚、暴、恶、悲"等；

(6) 某些汉字既可用作姓氏，又可用作名字用字。如"林、方、金、江、万、颜、童、柳"等。

上述各点，(1)、(2)、(4) 赋予中文姓名具有统计意义上的可区别性，(3)、(5) 使得部分姓名边界模糊，(6) 则导致相邻候选姓名之间产生交叉歧义。

2.2 中文姓名的概率分布

一般来说，中文姓名分单名 sn 和双名 pn 两类。

单名形如: sn = x(姓氏) m1(名字首字)

双名形如: pn = x(姓名) m1(名字首字) m2(名字末字)

姓氏又可分为单姓(如“赵、钱、孙、李”）和复姓（如“诸葛、司马、欧阳”）。冠夫性(如“李杨、刘张、陈王”）本文暂不予考虑。

单姓形如：x＝x1

复姓形如：x＝x1 x2　　(以上 x1,x2,m1,m2 均代表汉字)

令：fx(x0|x0∈姓氏)表示姓氏 x0 的使用频率;

fm(m0|m0∈名字用字)表示名字用字 m0 的使用频率

则根据表 XFL 及 MCFL，可给出姓名的概率估值;

p(sn)＝fx(x) * fm(m1)

p(pn)＝fx(x) * fm(m1) * fm(m2)

利用这两个公式，我们对清华大学近 10 年共 23，175 名学生的姓名（计单名 8，594 个，双名 14，581，单名与双名之比约为 37.2%:62.8%）进行了概率估值，从而得到中文姓名的概率分布曲线:

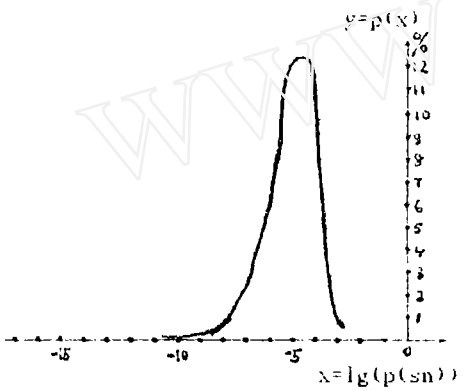

图 1　单名概率分布曲线

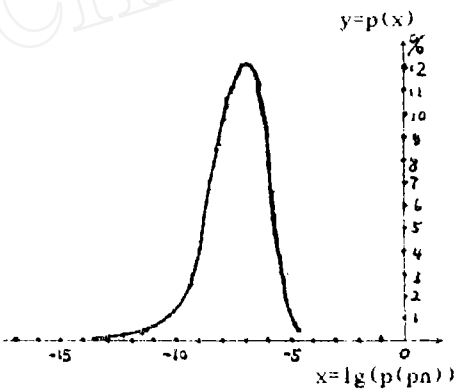

图 2　双名概率分布曲线

显然，这种概率分布是姓名辨识算法所乐见的。另，单名与双名的概率分布曲线非常相似。

2.3　绝对封闭式、相对封闭式和开放式名字用字

根据构词能力，名字用字可划分成三类:

名字用字{ 可用作单字词→开放式名字用字
不可用作单字词{ 可构词→相对封闭式名字用字
不可构词→绝对封闭式名字用字

●绝对封闭式名字用字

一类名字用字，既不可作单字词，又不能作构词成分。若这类字落在某个姓氏的作用域内，则必为该姓名的一部分。进一步地，若落在姓名的 m2 位置，则该姓名的右边界定。如“李逵”中的“逵”，“郑筱云”中的“筱”，“刘仲藜”中的“藜”字等。

●相对封闭式名字用字

胡戎睿十分聪明。 (例3*a*)
胡戎睿智过人。 (例3*b*)

“睿”不是单字词，但可构词，如“睿智”。当不存在相应的构词语境时，这类字的作用与绝对封闭式名字用字完全相同（例 3a 中，“胡戎睿”应判成姓名）。否则，应与姓名脱离（例 3b 中，“睿”属“睿智”，“胡戎”为姓名）。

●开放式名字用字

张玉爱读小说。 (例4)

例 4 中的“爱”字，可成为姓名的一部分，同时又可以单字词的身份参与句子的活动。这将导致在形式上，句子有两种可能的解释：

张玉 爱读小说。

张玉爱 读小说。

从姓名的角度来看，如果这类字出现在姓名的 m2 位置上，那么该姓名的右边界是不确定的（即应保留“x m1”及“x m1 m2”两种可能）。故称之为开放式名字用字。

2.4 称谓表

称谓常与姓名联袂出现，因此对辨识姓名有指示作用。

省长李长春赶到了抗灾现场。 (例5)
这是王继宁教授的学生。 (例6)

例 5 中，由称谓“省长”指示姓名的左边界；例 6 中，由称谓“教授”指示姓名的右边界。

称谓具有三种属性：

●只能用于姓名之后，如“之流”“阁下”等；

●只能用于姓名之前，如“青年”“战士”“运动员”等；

●用于姓名前后均可，如“先生”“同志”“市长”“司令员”等。

与称谓表相配合，我们还设置了一个称谓前缀表，存放“副、总、代、代理、助理、常务、名誉、荣誉”等，以减少称谓表的冗余度。通过简单的组合即可处理“代理副总经理”之类较为复杂的称谓。

2.5 简单上下文

●指界动词

一些动词，如“说、是、指出、认为、表示、参加”等，常紧接姓名的后面，故可用来帮助判断姓名的右边界。如：

姬鹏飞指出，… (例7)

●匹配模式

某些模式，如：“…的＜姓名＞”、“以＜姓名＞{称谓}为＜称谓＞”（大括号表可缺省）等，亦具界定姓名左右的功效。如：

毕业于同济大学分校的叶冬梅1983年分到局机关后，……　　(例8)

以潘杜泉为团长的香港工会代表团和以欧少雄为团长的澳门工会代表团是应中华全国总工会邀请,…　　(例9)

三、算法设计

3.1 预切分、标志数组及概率初筛选

输人文本分割成句子后送人数组 SENT，并在常用词表的支持下，利用最大匹配法对 SENT 进行分词，分词结果仍存回 SENT，且保持 SENT 的长度不变（不妨认为切分出的词与词之间存在隐式的"空"分割符）。然后，根据上节所述资源，建立一个与经分词处理后的数组 SENT 具有一一对应关系的标志数组 FLAG。赋加标志遵循的规则是：

对句内任一位置 pt (0<pt<length(SENT)-1)

若 SENT[pt]为孤立字且该字不可作单字词　　FLAG[pt]←'0';

若 SENT[pt]为孤立字且该字可作单字词　　FLAG[pt]←'1';

若 SENT[pt]属于某个多字词(含双字词)　　FLAG[pt]←'2';

若 SENT[pt]属于某个称谓　　FLAG[pt]←'3';

若 SENT[pt]属于某个指界动词(单字词)　　FLAG[pt]←'4';

若 SENT[pt]属于某个指界动词(双字词)　　FLAG[pt]←'5';

若 SENT[pt]是分隔符(数字、字母、标点等非汉字)　　FLAG[pt]←'6';

赋加标志的优先权: '4'>'1'; '3'>'5'>'2'

接着，算法将寻找 SENT 中所有可能的潜在姓名。设 SENT 中任一连续汉字串 CSTR

C1C2:　　若有 C1(单姓)∈XFL,C2∈MCFL; 或　　/ * 单姓单名 * /

C1C2C3:　　若有 C1(单姓)∈XFL,C2,C3∈MCFL; 或　　/ * 单姓双名 * /

C1C2C3:　　若有 C1C2(复姓)∈XFL,C3∈MCFL; 或　　/ * 双姓单名 * /

C1C2C3C4:　若有 C1C2(复姓)∈XFL,C3,C4∈MCFL　　/ * 双姓双名 * /

则 CSTR 即被视作一潜在姓名 cn，并将之添加到潜在姓名表 CNL 中。

注意，此操作的作用范围并不受分词结果的制约，即使是已经切出来的多字词，仍可能成为姓名的一部分（这也正是引人术语"隐式分割符"的原因）。只不过在概率筛选阶段，将面临不同的阈值罢了。

若 cn 所属各字对应的 FLAG 值均为"0"、"1"或"4"。

则　　state(cn)←1;

否则　state(cn)←0。

下一步是概率初筛选:

对任一 cn∈CNL,做

state(cn)=1:

若(cn 为单名且 lg(p(cn))<ξ2)或(cn 为双名且 lg(p(cn))<ξ3)

则从 CNL 中删除 cn　　/ * 情形 1 * /

state(cn)=0:

若(cn 为单名且 lg(p(cn))＜ξ2)或(cn 为双名且 lg(p(cn))＜ξ3)

则从 CNL 中删除 cn　　/ * 情形 2 * /

取: ξ2＝−8.538992; ξ3＝−11.903766

ξ2＝−6.239499; ξ3＝−10.068238

情形 1 出现的机会远远大于情形 2。相应阈值的设置使得

p(lg(p(sn))＜ξ2)≤0.55%　p(lg(p(pn))＜ξ3)≤0.38%

总体上将保证辨识算法的召回率(定义见 4.3 节)不低于: 1−(0.55% * 37.2%+0.38% * 62.8%)＝99.56%.

情形 2 系针对姓名中含多字词(或多字词的一部分)之现象而设计。既要尽可能滤掉 CNL 表中"任何"、"香港"、"代表"、"记者"、"宣传"、"国科学"、"公司经"、"新华社"、"木材公"、"来参加"这样一些希望甚微的潜在姓名，又要尽可能保留"田间"、"汪洋"、"严肃"、"华丽"、"安然"、"杨万里"、"王朝闻"、"关山月"、"马胜利"、"盛世来"、"高海燕"之类极可能成为姓名者。

3.2 同源对表、互斥对表及其操作

[定义]同源对

由以句内同一位置为姓氏始点的单名和双名各一组成。记作[单名，双名];

[定义]互斥对

由以句内不同位置为姓氏始点，相互间有部分交叉的两个姓名组成。记作

＜姓名 1,姓名 2＞

于是，对每个输入句子，在潜在姓名表 CNL 的基础上，将产生一个同源对表 SSL 和一个互斥对表 CTL。SSL 及 CTL 体现了句子中潜在姓名之间的相互制约关系。

这两个表关涉的操作有:

●reject(cn)

若某一 cn 已被"否定"为姓名:

a.从互斥对表 CTL 中删除所有包含 cn 的互斥对＜cni,cn＞及＜cn,cnj＞;

b. 从同源对表 SSL 中删除所有包含 cn 的同源对[cni,cn]及[cn,cnj];

c. 从潜在姓名表 CNL 中删除 cn。

●cnfirm(cn)

若某一 cn 已被"肯定"为姓名:

a. 缓存器 BUF 清零;

b. 从互斥对表 CTL 中删除所有包含 cn 的互斥对＜cni,cn＞及＜cn,cnj＞

且 BUF←BUF+{cni}; BUF←BUF+{cnj};

c. 从同源对表 SSL 中删除所有包含 cn 的同源对[cni,cn]及[cn,cnj]

且 BUF←BUF+{cni}; BUF←BUF+{cnj};

d. 对每一个 cnx∈BUF 做 reject(cnx).

3.3 姓名左右边界的确定

设某一潜在姓名 cn 的语境为:

context(cn)＝ zl cn zr　　(zl, zr 为汉字)

则有边界确定规则:

(1) 若 (FLAG(zl)='3')或(FLAG(zl)='6')或(cn 在句首),则
cn 的左边界确定，记作#cn
(2) 若 (FLAG(zr)='3')或(FLAG(zr)='5')或(FLAG(zr)='6'或
(cn 在句尾)或((cn 为双名)且(FLAG(zr)='4')),则
cn 的右边界确定,记作 cn#
(3) 若 (#cn)且(cn#),则 cn 被"确认"
并从潜在姓名表 CNL 中删除之，送入确认姓名表 OKL 中
(注意，一个姓名被"肯定"，仅意味着在同其它姓名的竞争中"生存"下来，
取得了继续保留在潜在姓名表中的资格."确认"则是百分之百的肯定)
匹配模式亦在此时动作.

3.4 两个特殊字表

在确定姓名的左右边界时，须作一些特殊处理。

●特殊字表 A-非常用姓氏兼最常用单字词

考察:
主帅是李广将军 (例10*a*)
司令官杜聿明将军… (例10*b*)

据 3.3 节之边界确定规则，"是李广""杜聿明"均被确认为姓名。对（例 10a)，有误。由于"是"可作姓氏（非常用)，但又可作最常用单字词，故对以之为姓氏的潜在姓名，左界确定规则不应发挥作用。类似的字还有:

百次从大道发国过和红还回家句可来老门那是树水同完位问新要有员远真种

特殊字表 A-非常用姓氏兼最常用单字词

●特殊字表 B-名字用字兼最常用单字词

考察:
博士生段恩和导师王松茂教授正在讨论问题 (例11)

"段恩和"不应被确认为姓名。即当潜在姓名为双名，且名字末字属"和"这类字（特殊字表 B）时，右界确定规则不应发挥作用。

爱八把百比边并不才长车成吃出船次从打大带党到道得的等地第点都对多而二发放风高个各给跟更国过好和河红后还回会家见讲进九就句开看可块快拉来老里两六路满每门们拿那内能年跑七起千钱前全却让人日三山上少身声十时使事是手受书树水四送所他它太提天条听同头外完万为位问我无五下先想向象小笑写新心性学眼要也一已以用由有又于与员远月再在早站找者真正之只中种住着走最作坐

特殊字表 B-名字用字兼最常用单字词

对特殊字表 B，尚需仔细甄别。譬如，通过遍历清华姓名库，发现"在"字仅可见于名字首字位置（即从不出现在名字末字位置)，如"于在河""王在明""陈在铁""金在荣""覃在林""谭在树""白在桥""尹在均""邓在军"等。

3.5 屏蔽与恢复

屏蔽与恢复操作均针对同源对

屏蔽发生在同源对中双名的名字末字位置之 FLAG 值为'0'时。如:

同源对[刘仲,刘仲黎]

"黎"的 FLAG 值为'0'，屏蔽有效."刘仲"被否定。

对某些双名，如果仅根据概率值，在初筛选阶段会被滤掉。但若满足条件:

名字末字位置的 FLAG 值为'0'且同源对中单名的概率值＞ξ2

则须予以保留(即恢复)。如:

lg(p(邵逸夫))＜ξ3

"邵逸夫"应被滤掉。然而:

"夫"的 FLAG 值为'0'且 lg(p(邵逸))＞ξ2

故应恢复"邵逸夫"，同时屏蔽"邵逸"。

3.6 同源对表、互斥对表的规则校正

[校正规则 1]同源对右界否定规则

若同源对形如[z1z2z3#,z1z2]则否定 z1z2;

[校正规则 2]互斥对左界否定规则

若互斥对形如＜#z1z2z3,z2z3＞则否定 z2z3;

[校正规则 3]互斥对落单字否定规则

设互斥对形如＜(cn1:PA+INTSECT) (cn2:INTSECT+PB)＞

/＊INTESECT 为 cn1,cn2 的交叉部分，且不为空＊/

若 cn1 之 PA 部分至少含一 FLAG 值为'0'的字，则否定 cn2;

若 cn2 之 PB 部分至少含一 FLAG 值为'0'的字，则否定 cn1;

[校正规则 4]互斥对等长概率否定规则

设互斥对＜cn1,cn2＞中 length(cn1)＝length(cn2)

若 p(cn1)＞p(cn2)则否定 cn2;

若 p(cn1)＜p(cn2)则否定 cn1;

[校正规则 5]互斥对不等长概率否定规则

设互斥对＜cn1,cn2＞中 length(cn1)≠length(cn2)

若 lg(p(cn1))／length(cn1)＞lg(p(cn2))／length(cn2) 则否定 cn2;

若 lg(p(cn1))／length(cn1)＜lg(p(cn2))／length(cn2) 则否定 cn1;

/＊关于 lg(概率值几何平均)之比较＊/

校正规则依所列顺序递次调用。校正规则 5 导致互斥对表 CTL 终必变化为空。

3.7 概率再筛选

对经以上各步余下的潜在姓名表 CNL，再进行一轮概率筛选。新的阈值定为

ξ2＝−7.904649;　ξ3＝−11.072278

ξ2＝−5.129399;　ξ3＝−8.881645

仅视乎此阈值，召回率也将不低于 98.69%。但由于透过多层处理后，OKL 表中已有相当积累，故辨识系统的实际召回率比这个指标要高。

四、实验系统与实验结果

4.1 系统结构

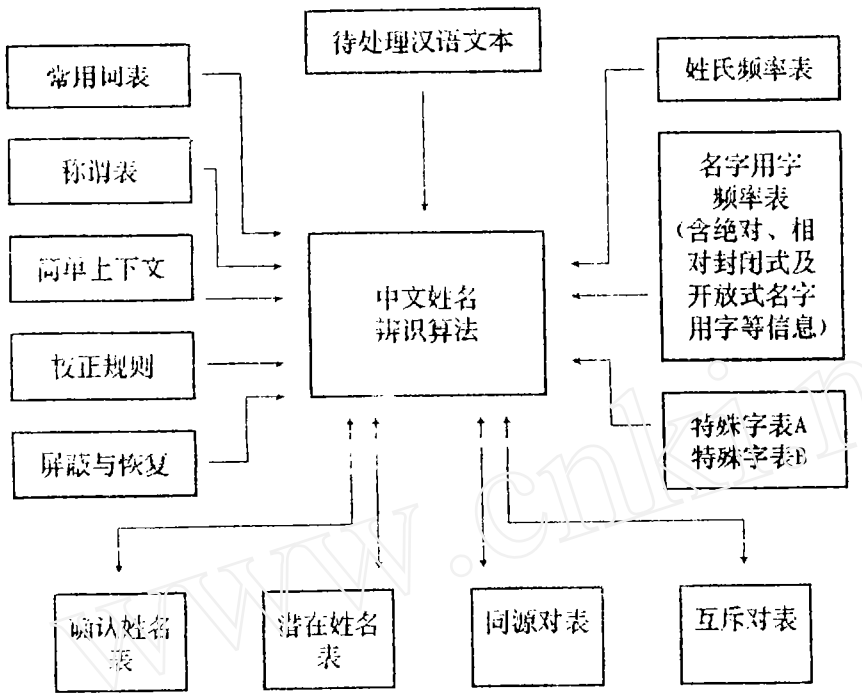

图 3　系统结构

4.2 分析例示

设输入句子为:

湖北柳大华和广东吕钦,　　(例12)

则算法分析过程如下:

a. 经最大匹配切分后，得标志数组:

SENT:　湖北柳大华和广东吕钦,

FLAG:　22110122006

并给出潜在姓名表 CNL:

(柳大华−7.523170) (大华和−9.348916) (大华−6.632542)

(华和广−8.837197) (华和−6.101831) (和广东−8.609877)

(和广−6.200005) (广东吕−11.244786) (广东−6.807564)

(东吕钦−10.978300) (东吕−7.771527) (吕钦−5.019326)

这里，潜在姓名"柳大"已被"柳大华"所屏蔽。

b. 概率初筛选后，CNL 变成:

(柳大华−7.523170) (大华和−9.348916) (大华−6.632542)

(华和广−8.837197) (华和−6.101831) (和广东−8.609877)

(和广−6.200005) (吕钦−5.019326)

互斥对表 CTL 为:

＜柳大华,大华和＞ ＜柳大华,大华＞＜柳大华,华和广＞

＜大华和,华和广＞ ＜大华,华和广＞ ＜柳大华,华和＞

＜大华和,华和＞ ＜大华,华和＞ ＜大华和,和广东＞ ＜华和广,和广东＞

＜华和,和广东＞ ＜大华和,和广＞ ＜华和广,和广＞ ＜华和,和广＞

同源对表 SSL 为:

[大华,大华和] [华和,华和广] [和广,和广东]

c. 运用同源对表、互斥对表的校正规则:

由[校正规则 4] 对互斥对＜柳大华,大华和＞否定“大华和”;

由[校正规则 4] 对互斥对＜柳大华,华和广＞否定“华和广”;

由[校正规则 4] 对互斥对＜大华,华和＞ 否定“大华”;

由[校正规则 3] 对互斥对＜华和广,和广东＞ 否定“和广东”;

由[校正规则 3] 对互斥对＜大华和,和广＞ 否定“和广”

注意，每一次运用规则后，CNL、CTL 和 SSL 要作相应调整。此时有:

CNL (柳大华−7.523170) (华和−6.101831 (吕钦−5.019326)

CTL ＜柳大华,华和＞

SSL 空表

d. 进一步,由[校正规则 5] 对互斥对 ＜柳大华,华和＞ 否定“华和”;

e. 最终结果是:

CNL (柳大华−7.523170) (吕钦−5.019326)

CTL 空表

SSL 空表

4.3 实验结果及其讨论

实验系统在 486 微机上用 Borland C 实现。为了验证算法的有效性，我们从新华通讯社的新闻语料库中随机抽取了 300 个包含中文姓名的句子作为测试样本。这 300 个句子共含 8144 个中文字符，434 个中文姓名。系统运行后，辨识出“中文姓名”618 个，其中 433 个为真正正确者（确认姓名表 OKL 中有 207 个）。设:

召回率＝文本中的中文姓名被辨识出的比例

精确率＝辨识为中文姓名者真正为中文姓名的比例

则本系统的召回率和精确率分别为:

召回率＝433／434＝99.77%

精确率＝433／618＝70.06%

以下给出若干辨识结果:

[输入]

管理系硕士生阎尔坤提出， (例13)

[输出]

※OKL: 阎尔坤

[输入]

全国政协副主席洪学智，第二炮兵司令员李旭阁、政治委员刘安元等

出席开幕式 (例14)

[输出]

※OKL: 洪学智　李旭阁　刘安元

[输入]

表彰林军、戴久高、朱荣昌、杨文禄四位用鲜血甚至生命保卫国家财产的农村金融工作者， (例15)

[输出]

※CNL: 戴久高　卫国家

※OKL: 林军　朱荣昌　杨文禄

[输入]

广东省连南瑶族自治县板坳村的共产党员易顺， (例16)

[输出]

※CNL: 连南

※OKL: 易顺

[输入]

武钢团委书记铁红英说， (例17)

[输出]

※CNL: 武钢

※OKL: 铁红英

[输入]

河南会员冯俊发愿无偿赠送百日红1000株 (例18)

[输出]

※CNL: 冯俊发　冯俊

※SSL: [冯俊,冯俊发]

[输入]

当王文清的妻子韩英红接过奖章和证书时， (例19)

[输出]

※CNL: 王文清　王文　韩英红　韩英

※SSL: [王文,王文清] [韩英,韩英红]

在召回率和精确率之间，我们优先考虑召回率。召回率尽量高，即使精确率不十分理想，尚可有其它后续手段予以进一步的处理（对这一问题将另文讨论）。反过来，如果真正的姓名并不存在于潜在（或者叫侯选）姓名表中，则无米之炊，势必难为矣。实验结果表明，算法在召回率方面的表现极为突出。

与召回率相比较，精确率稍嫌逊色。这是为了保证召回率付出的必要代价。算法实际上给出了全部具有一定可能性的"形式"姓名。在不对句子进行深层次"理解"的条件下，很难讲这种策略是不合理的。如例 15 至例 19 中的"连南""武钢""王文""韩英""卫国家"等，孤立地看确实可作为姓名，纵便联系到句子内，句法分析也并不能可靠、有效地把它们从姓名表中排除出去。再如，对例 18，CNL 表中的"冯俊"及"冯俊发"均能通过句法甚至语义分析：

河南 会员 冯俊 发愿 无偿 赠送 百日红 1000 株。

河南 会员 冯俊发 愿 无偿 赠送 百日红 1000 株。

藉助其它手段，精确率尚可望提高：(i)应用中文姓名的某些构造规律，如兄弟排行字"伯""仲""叔""季"，或叠字名"媛媛""强强""毛毛""潇潇"；(ii)对某些特殊的常用字(如"在"字)，姓名首字与姓名末字这两个位置应予区分；(iii)外文译名中的某一部分会被误认作中文姓名，则增加外文译名辨识机制，并使两者结合起来，即可妥善处置之。

参考文献

[1] 张俊盛，陈舜德，郑崇，刘显仲，柯淑津，"多语料库作法之中文姓名辨识"，《中文信息学报》，Vol.6,No.3,1992

[2] 郑家恒，刘开瑛，"自动分词系统中姓氏人名处理策略探讨"，陈力为主编《计算语言学研究与应用》，北京语言学院出版社，1993

[3] 宋柔，朱宏，潘维桂，尹振海，"基于语料库和规则库的人名识别法"，陈力为主编《计算语言学研究与应用》，北京语言学院出版社，1993

[4] 中国社会科学院语言文字应用研究所汉字整理研究室，《姓氏人名用字分析统计》，语文出版社，1990

[5] 孙茂松，张维杰，"英语姓名译名的自动辨识"，陈力为主编《计算语言学研究与应用》，北京语言学院出版社，1993

Identifying Chinese Names in Unrestricted Texts

Sun Maosong, Huang Changning, Gao Haiyan+, Fang Jie

(Department of Computer Science, Tsinghua University

+Department of Computer Application, Yantai University)

Abstract

The processing of Chinese names is significant to the approach of Chinese word segmentation. This paper presents an effective algorithm for automatically identifying this sort of proper nouns in Chinese texts. The testing sample, involving 300 sentences each of which contains at least one Chinese names, is extracted at random from the Xinhua News Corpus. The preliminary experiment shows that the recall rate of this algorithm reaches 99.77%.

Keywords: Chinese name identification, Unknown word processing,
Chinese word segmentation, Chinese information processing