基于角色标注的中国人名自动识别研究

张华平[1),2)] 刘 群[1),3)]
1)(中国科学院计算技术研究所 北京 100080)
2)(中国科学院研究生院 北京 100080)
3)(北京大学信息学院计算机科学与技术系计算语言研究所 北京 100871)

摘 要 该文提出了一种基于角色标注的中国人名自动识别方法. 其基本思想是: 根据在人名识别中的作用, 采取 Viterbi 算法对切词结果进行角色标注, 在角色序列的基础上, 进行模式最大匹配, 最终实现中国人名的识别. 识别过程中只需要将某个词作为特定角色的概率以及角色之间的转移概率. 该方法的实用性还在于: 这些角色信息完全可以从真实语料库中自动抽取得到. 通过对 16M 字节真实语料库的封闭与开放测试, 该方法取得了接近 98% 的召回率. 文中介绍了计算所汉语词法分析系统 ICTCLAS, 集成人名识别算法之后, 词法分析的准确率提高了 1.41%, 同时人名识别的综合指标 F-1 值达到了 95.40%. 不同实验从各个角度表明: 基于角色标注的人名识别算法行之有效.

关键词 中国人名识别; 未登录词识别; 角色标注; Viterbi 算法
中图法分类号 TP391

Automatic Recognition of Chinese Personal Name Based on Role Tagging

ZHANG Hua-Ping[1),2)] LIU Qun[1),3)]
1)(*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)
2)(*Graduate School, Chinese Academy of Sciences, Beijing* 100039)
3)(*Institute of Computational Linguistics, Department of Computer Science and Technology, Information Science School, Peking University, Beijing* 100871)

Abstract Automatic recognition of Chinese personal name is emphasis and difficulty for unknown words recognition. Because of their inherent deficiencies, previous solutions are not satisfactory. This paper presents an approach for Chinese personal name recognition based on role tagging. That is: tokens after segmentation are tagged using Viterbi algorithm with different roles according to their functions in the generation of Chinese personal name; The possible names are recognized after maximum pattern matching on the roles sequence. During the recognition process, only the possibilities of tokens being specific roles and the transition possibilities between roles are required. The significance is that such lexical knowledge can be totally extracted from corpus automatically. In both close and open test on a 16-Mbyte realistic corpus, its recalling rate is nearly 98%. After combined with the algorithm for personal name recognition, authors' Chinese lexical analysis system ICTCLAS improves 1.41% in performance while the agglomerative evaluation argument F-1 value of person recognition achieve 95.40%. Various experiments show that: role-based algorithm proposed in this paper is effective for Chinese personal name recognition.

Keywords Chinese personal name recognition; unknown words recognition; role tagging; Viterbi algorithm.

收稿日期: 2002-04-08; 修改稿收到日期: 2003-02-18. 本课题得到国家"九七三"重点基础研究发展规划项目(G1998030507-4; G1998030510)和中国科学院计算技术研究所领域前沿青年基金项目(20026180-23)资助.张华平, 男, 1978 年生, 博士研究生, 主要研究方向为计算语言学、中文信息处理与信息抽取. E-mail: zhanghp@software.ict.ac.cn. 刘 群, 男, 1966 年生, 在职博士研究生, 副研究员, 主要研究方向为机器翻译、自然语言处理与中文信息处理.

1 引 言

词法分析是中文自然语言处理的前提和基础，目前中文词法分析的研究已经取得较大进展，但在处理含有未登录词的文本时，其结果很难满足实际需求.未登录词往往与其前后的字词交叉组合，不仅增加了自身切分的难度，而且严重地干扰了相邻词的正确切分，从而大大地降低了词法分析乃至整个句子分析的正确率.未登录词问题实际上已经成为了词法分析实用化的主要瓶颈.

中国人名①在未登录词中占有较大比重，也是未登录词识别的主要难点.在《人民日报》1998 年 1 月的语料库(共计 2305896 字)中，平均每 100 个字包含未登录词 1.192 个(不计数词、时间词)，其中 48.6%的未登录词是中国人名.国家“八六三”高技术研究发展计划 306 智能接口技术专家组 1998 年对国内自动分词软件的评测结果表明[1]：中国人名识别的召回率仅为 68.77%，其切分错误高达 50%以上，对所有分词错误进行统计，姓名错误占了将近 90%[2].因此中国人名的自动识别是未登录词识别的重点和关键，中国人名识别问题的解决必然会提高词法分析、句法分析乃至整个中文信息处理的质量.

1.1 中国人名自动识别的困难

中国人名数量众多，规律各异，有很大的随意性.对其进行识别的主要困难在于：(1)中国人名构成的多样性；(2)人名内部相互成词；(3)人名与其上下文组合成词；(4)歧义理解.

中国人名构成的形式有：(1)姓＋名，如：张华平、张浩、西门吹雪、诸葛亮；(2)有名无姓，如：“春花点点头”；“杰，你好吗？”(3)有姓无名，如：“刘称赵已离开江西”；(4)姓＋前后缀，如：刘总、张老、小李、邱某；(5)港澳台等地已婚妇女的姓名，如：范徐丽泰、彭张青.

人名内部相互成词，指的是姓与名、名与名之间本身就是一个已经被核心词典收录的词.如：“人散后宝玉回到怡香院”；[王国]维、[高峰]、[汪洋]、张[朝阳].根据我们对 8 万条人名的统计，内部成词的比例高达 6.89%.

人名与其上下文组合成词包括人名的首部(姓或名的首字)与人名的上文成词以及人名的尾部(姓或名的末字)与下文成词.例如：“这里[有关]天培的壮烈事迹”；“费孝通向人大常委会提交书面报告”.在《人民日报》1998 年 1 月的语料库中，这种情况接近 200 例.

歧义理解主要是由同源歧义冲突[3]引起的，例如：“河北省刘庄”中的“刘庄”存在中国人名与地名的两种歧义理解，“周鹏和同学”存在人名“周鹏”和“周鹏和”的歧义[4].

1.2 现有解决方案及其不足

针对中国人名的自动识别问题，人们已经作过深入的探索，并提出了多种解决方案.根据其使用的方法不同，这些方案大致可以分为三种：规则方法[2,4,5]、统计方法[6]以及规则与统计相结合的方法[1,7,8].

规则方法主要是利用两种信息：姓氏用字分类[5]和限制性成分[8].即：分析过程中，当扫描到具有明显特征的姓名用字时，开始触发姓名的识别过程，并采集姓名前后相关的成分，对姓名的前后位置进行限制.小规模测试的结果表明，其准确率可以高达 97%[4].在缺乏大规模熟语料库的时候，规则似乎是唯一可行的方法.统计方法主要是针对姓名语料库来训练某个字作为姓名组成部分的概率，并用它们来计算某个候选字段作为姓名的概率值，其中概率值大于某一阈值的字段识别为中国人名[6].规则与统计相结合的方法，一方面通过概率计算来减少规则方法的复杂性与盲目性，另一方面通过规则的复用，来降低统计方法对语料库规模的要求.目前的研究基本上都是采取规则与统计相结合的方法，不同之处仅在于规则与统计的不同侧重.

现有解决方案本身存在一些固有的不足：首先，它们一般都采取“单点(首或尾)激活”[4]的机制来触发人名的识别处理.即扫描到姓氏用字、职衔、称呼等具有明显姓名特征的字段时，才会将前后的几个字列为候选姓名字段进行人名的识别.这样往往会丢失那些不具备明显特征的姓名，如上文中提到的“有名无姓”的情况.其次，姓名候选字段大都是选取切分后的单字碎片[1,2,4,6]，也有部分研究者将少量的二字或多字词纳入候选字段的选取范围[4].在这种选取机制的作用下，内部成词以及与上下文成词的人名就很难召回.根据上文的统计数据，由于这两种机制所引起的召回率损失不小于 10%.再者，人名识别所用的规则往往琐碎，一般代价昂贵而且难以扩展.例如，文献[4]中，研究者就是从 10 万条人名库、2 亿字的真实语料库中将姓名用字分为了 9 类，并总结了 21 条识别规则.无论是收集规模巨大的人名库与真实语料库，还是提炼识别规则，都是一个费时费力的浩大工程，是一般机构难以承受的.一旦增加了新特征的人名，还必须增加相应的新规则，对以前的规则重新修订，因此规则方法很难扩展.我们也知道，规则可以保证很高的准确率，但是任何规

① 亚洲某些国家的非音译人名也类似.

则体系的覆盖范围都是有限的, 对于覆盖集之外的人名就完全无能为力.

本文提供一种可以避免上述不足的解决方案——基于角色标注的中国人名自动识别方法. 该方法主要采用隐马尔可夫模型在分词结果上标注人名构成角色, 然后在标注出的角色序列基础上根据各个不同的角色, 进行最长模式串匹配, 最终识别出人名. 某个字词人名构成角色的制定依据是该字词在人名构成中所起的不同作用, 如: 姓、名、上文、下文等. 角色标注所需的字词作为不同角色的出现概率和角色间的转移概率, 都是在语料库训练过程中自动抽取的. 基于角色标注的人名识别算法的特色还在于: 自动学习、自动识别, 无需人工的直接干预, 改变训练样本, 就可以适应新的情况; 一次扫描, 无需回溯, 同时选取所有可能字段作为候选姓名, 识别处理作用在整个切分序列上, 无需激活. 在大规模真实语料库上的不同测试实验表明, 该方法能取得较高的召回率. 最后, 我们将角色标注的人名识别应用到了中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS 中, 应用人名识别前后的对比实验说明: 人名识别极大地提高了词法分析性能, 人名识别也取得了较好的综合性能.

本文下面将阐述该方法的理论依据, 随后给出具体的实现算法, 最后提供人名识别不同的评测实验并进行结果分析.

2　基于角色标注的中国人名自动识别方法

2.1　中国人名的构成角色

中国人名用字与上下文用词都比较集中, 有很强的规律性. 在 83077 条人名库中, 姓氏用字仅有 820 个, 其中王、张、李三大姓, 就占了 20%; 20631 个单名中, 单名用字为 1489 个; 双名的首字与末字数量均不到 2000 个. 人名的上下文用词范围也很有限. 上文一般是称呼、职衔和一些连词、动词等, 如: "总统"、"主任"、"打"、"向"等. 下文大多是"说"、"表示"、"同志"之类的动词和称谓词.

在这里, 我们将一个句子中所有的词划分为: 人名的内部组成、上下文、无关词, 并称之为中国人名的构成角色(为行文方便, 以下简称角色). 具体的角色分类见表 1, 根据该表, 对切分结果"馆/内/陈列/周/恩/来/和/邓/颖/超生/前/使用/过/的/物品"进行角色标注, 其结果为: "馆/A 内/A 陈列/K 周/B 恩/C 来/D 和/M 邓/B 颖/C 超生/V 前/A 使用/A 过/A 的/A 物品/A".

表 1　中国人名的构成角色表

编码	意义	例子
B	姓氏	张华平先生; 欧阳修
C	双名的首字	张华平先生
D	双名的末字	张华平先生
E	单名	张浩说: "我是一个好人"
F	前缀	老刘、小李
G	后缀	王总、刘老、肖氏、吴妈、叶帅
K	人名的上文	又来到于洪洋的家.
L	人名的下文	新华社记者黄文摄
M	两个中国人名之间的成分	编剧邵钧林和稽道青说
U	人名的上文与姓氏成词	现任主席为何鲁丽.
V	人名的末字与下文成词	龚学平等领导
X	姓与双名的首字成词	王国维、
Y	姓与单名成词	高峰、汪洋
Z	双名本身成词	张朝阳
A	其它无关词	全军 和 武警 官兵 深切 缅怀 邓 小平

2.2　角色自动标注与中国人名识别

既然含中国人名的句子包含姓、名、上下文等构成角色, 那么换一个角度说: 我们可以标注句子中词语的不同构成角色, 通过对角色序列进行简单的模式匹配来实现中国人名的识别. 实际上, 中国人名构成角色的标注本质上是一个简单的词类标注过程.

我们采用 Viterbi 算法[9] 来实现角色自动标注. 即: 从所有可能的标注序列中优选出概率最大者作为最终标注结果. 求解过程推导如下:

我们假定 W 是分词后的 Token 序列(即未登录词识别前的词语切分结果), T 是 W 某个可能的角色标注序列. 其中 $T^{\#}$ 为最终标注结果, 即概率最大的角色序列. 则有

$$W = (w_1, w_2, \cdots, w_m),$$
$$T = (t_1, t_2, \cdots, t_m), \quad m > 0,$$
$$T^{\#} = \arg\max_T P(T \mid W) \tag{1}$$

根据贝叶斯公式, 有

$$P(T \mid W) = P(T)P(W \mid T)/P(W) \tag{2}$$

对于一个特定的 Token 序列来说, $P(W)$ 是一个常数, 因此根据式(1)和式(2)我们可以得到

$$T^{\#} = \arg\max_{T} P(T)P(W \mid T) \quad (3)$$

如果把词 w_i视为观察值, 把角色 t_i视为状态值(其中 t_0为初始状态). 则 W 是观察序列, 而 T 为隐藏在 W 后的状态序列, 这是一个隐马尔可夫链. 那么, 我们可以引入隐马尔可夫模型[10] 来计算 $P(T)P(W|T)$. 即

$$P(T)\,P(W \mid T) \approx \prod_{i=1}^{m} p(w_i \mid t_i)p(t_i \mid t_{i-1})$$

所以,

$$T^{\#} = \arg\max_{T} \prod_{i=1}^{m} p(w_i \mid t_i)p(t_i \mid t_{i-1}) \quad (4)$$

为了计算的简便, 对式(4)中的概率取负对数, 则有

$$T^{\#} = \arg\min_{T}\left(-\sum_{i=1}^{m}[\ln p(w_i \mid t_i) + \ln p(t_i \mid t_{i-1})]\right) \quad (5)$$

最后, 角色自动标注问题就转换为式(5)的求解问题. Viterbi 算法[9] 是解决这类问题的经典算法, $T^{\#}$也就迎刃而解了.

为了解决人名与其上下文组合成词的问题, 在人名识别之前, 我们要对角色 U(人名的上文和姓成词)和 V(人名的末字和下文成词)进行分裂处理. 相应地分裂为 KB, DL 或者 EL. 而基于角色序列的人名识别可以经过最终角色序列基础上的模式串最大匹配实现. 我们使用到的人名识别模式集为{ BBCD, BBE, BBZ, BCD, BE, BG, BXD, BZ, CD, FB, Y, XD}. 一旦匹配到其中的一个最长模式串, 其对应的 Token 片段就识别为中国人名, 2.1 节中的例句"馆/内/陈列/周/恩/来/和/邓/颖/超生/前/使用/过/的/物品"经过 Viterbi 计算后, 对应的 $T^{\#}$为: "AAKBCDMBC-VAAAAAA". V 分裂处理后, 最终的角色序列为: "AAKBCDMBCDLAAAAAA". 模式最大匹配后, 我们识别出的人名是: "周恩来"和"邓颖超".

2.3 角色信息的自动抽取

$p(w_i \mid t_i)$ 和 $p(t_i \mid t_{i-1})$是式(5)中两个关键的角色信息参数. 其中 $p(w_i \mid t_i)$对应的实际意义是在给定角色 t_i的条件下, 该 Token 为 w_i的概率; $p(t_i \mid t_{i-1})$表示角色 t_{i-1}到角色 t_i的转移概率. 在大规模语料库训练的前提下, 根据大数定理, 我们可以得到

$$p(w_i \mid t_i) \approx C(w_i, t_i)/C(t_i) \quad (6)$$

其中 $C(w_i, t_i)$为 w_i作为角色 t_i出现的次数; $C(t_i)$为角色 t_i出现的次数.

$$p(t_i \mid t_{i-1}) \approx C(t_{i-1}, t_i)/C(t_{i-1}), \quad i > 1 \quad (7)$$

其中 $C(t_{i-1}, t_i)$为角色 t_{i-1}的下一个角色是 t_i的次数; $C(w_i, t_i)$, $C(t_i)$, $C(t_{i-1}, t_i)$均可通过对已经切分标注好的熟语料库进行学习训练、自动抽取得到.

3 自动识别的实现算法

基于角色标注的中国人名自动识别主要包括三个过程: 角色信息的自动抽取; 角色标注以及人名的最终识别. 角色标注实质上类似于一个小型的词性标注过程, 主要是从所有可能的角色标注中, 尽快选取满足式(5)的标注序列. Viterbi 算法专门解决这类问题, 已经非常成熟. 在此不作介绍. 下面分别给出角色信息自动抽取算法和整个中国人名自动识别流程.

3.1 角色信息自动抽取算法

我们的训练集来自于切分标注好的《人民日报》语料库, 该语料库采用的是北京大学计算语言所制定的词类标注集. 在北京大学的标准中, 中国人名的姓与名切分开, 但只采用同一个标注 nr, 不利于中国人名的甄别, 如图 1 所示. 为此, 我们保留原始的切分结果, 采用 ICTPOS (中国科学院计算技术研究所词类标注集)代替北京大学的标注集. ICTPOS 对姓名分别标注为 nf 和 nl, 并将整个人名标注为 nr, 如图 2 所示. 修正后的语料库更利于人名的定位和分析.

角色信息自动抽取的基本思路是: 在修正的语料库基础上, 将词类标注转换为角色并进行角色信息统计. 具体算法如下:

(1) 从切分标注好的熟语料库中依次读入按词性标注好的句子.

(2) 根据词性标注 nf(姓氏), nl(名)或者 nr(姓名)定位出中国人名, 标注将中国人名以外的词的标注换成角色 A.

(3) 若人名前面的片断 p 和人名首部 f 成为新词 pf, 将 pf 标注为 U, 否则将 p 标为 K (若 p 原来标注的角色是 A)或 M (若 p 原来标注的角色是 L).

(4) 若人名尾部 t 和人名后面的片断 n 成为新词 tn, 将 tn 标注为 V, 否则将 n 标为 L.

(5) 根据本文 1.1 节中人名的 5 种类别, 分别

```
本报/r 蚌埠/ns 1月/t 1日/t 电/n 记者/n 黄/nr 振中/nr 、/w 白/nr 剑峰/nr 报道/v
:/w 新年/t 的/u 钟声/n 刚刚/d 敲响/v ,/we 千/m 里/q 淮河/ns 传来/v 喜讯/n
```

图 1 采取北京大学计算语言研究所词类标注集的原始语料

```
本报/r 蚌埠/ns 1月/t 1日/t 电/n 记者/n [黄/nf 振中/nl]nr 、/we [白/nf 剑峰/nl]nr 报道/v
:/we 新年/t 的/uj 钟声/n 刚刚/d 敲响/v ,/we 千/m 里/q 淮河/ns 传来/v 喜讯/n
```

图 2 采取 ICTPOS(中国科学院计算技术研究所词类标注集)标注的原始语料

对姓、双名首字、双名末字、单名、前缀、后缀相应地标注为角色 B, C, D, E, F, G. 内部成词的情况, 相应地标注为 X, Y, Z.

(6) 在句子的角色序列中, 将角色不是 A 的词 w_i存入中国人名识别词典, 并统计 w_i作为t_i的出现次数 $C(w_i, t_i)$. 同时累计所有不同角色的出现次数 $C(t_i)$以及相邻角色的共现次数 $C(t_{i-1}, t_i)$.

通过上述过程, 转换后的角色语料如图 3 所示.

本报/A　蚌埠/A　1 月/A　1 日/A　电/A　记 者/K　黄/B　振/C 中/D、/M　白/B　剑/C 峰/D　报 道/L :/A　新年/A　的/A　钟声/A　刚刚/A　敲响/A ,/A　千/A　里/A　淮河/A 传来/A 喜讯/A

图 3　人名角色标注原料

3.2　中国人名的识别流程

(1) 对句子进行分词(我们采取的是基于 *N*-最短路径的汉语粗分模型[11]), 采取 Viterbi 算法进行角色标注, 求出概率最大的角色序列 $T^{\#}$.

(2) 将角色为 U 的片断 pf 分裂为 KB(若 f 为姓)、KC(若 f 为双名首字)或 KE(若 f 为单名).

(3) 将角色为 V 的片断 tn 分裂为 DL(若 t 为双名末字)或 EL(若 t 为单名).

(4) 对分裂处理后的角色序列在姓名识别模式集中进行模式串最大匹配, 输出对应片段组成人名, 同时记录它们在句子当中的位置.

(5) 对识别出来的结果加入一些限制规则, 排除错误的中国人名. 如中国人名前后不能是"·"(因为这种情况下, 往往是外国人的译名).

4　实验结果与分析

下面给出我们在实验过程中所采取的测试集和指标, 然后给出不同条件下的实验结果及相应分析.

4.1　测试集与评测指标

中国人名这类专名的识别评测, 往往需要一定规模的测试集. 测试集一般有两种: 只含专名的句子集和完全真实的语料库. 前者忽略了真实语言环境中大量的不含特定专名的句子, 测试结果往往偏高. 如在真实语料中, 不含中国人名的句子超过 90%, 而这些句子很可能被错误地识别出人名来. 例如: "吕梁的特点是贫困人口占全省的 1/3 左右." 中的"吕梁"一般都会被识别为人名. 但是在只含专名的句子集上进行的测试实验中, 这种可能被错误识别的句子事先均被人为地排除掉了. 另外, 根据已发表的文章显示, 目前用来测试的句子集包含人名的数量均不足 1000 个[2,4], 相比之下, 语料库规模往往比较大, 包含的人名数量和种类都很丰富, 在此基础上测试更具统计意义. 因此, 在不做任何筛选的大规模真实语料库上进行测试, 评测结果更符合真实的语言环境.

根据测试集和训练集的不同关系, 我们也可以将评测分为封闭测试和开放测试. 其中封闭测试的测试集是训练集的子集, 而开放测试的测试集与训练集不存在包含或者被包含的关系. 在以下的实验中, 我们将在《人民日报》切分标注好的真实语料库上进行各种条件下的封闭和开放评测.

针对中国人名识别, 我们采取了常用的 3 个评测指标, 即准确率(P)、召回率(R)和综合指标 F 值(F). 其定义如下:

准确率 P=

正确识别出的人名数/识别出的人名数×100%.

召回率 R=

正确识别出的人名数/实际人名数×100%.

$$F=\frac{R\times P\times(1+\beta^2)}{R+P\times\beta^2}.$$

其中 β 是准确率P 和召回率R 之间的权衡因子, 在这里, 我们认为 P 和R 同等重要, 因此 β 取 1, 此时 F 称为F-1 值.

4.2　未登录人名的识别评测实验

一方面, 中国人名的识别关键在于识别算法, 另一方面, 用来分词的核心词典收录的词条数目以及收录人名的数量也直接影响最终的识别效果. 对词典中已收录人名进行识别, 实际上就是简单的模式串匹配, 识别准确性和召回率往往接近于 100%. 对于一个核心词典中收录了大量人名的系统来说, 最终的人名识别性能更大程度上归功于核心词典. 为了客观评价基于角色标注的中国人名识别算法, 我们做了三组只考虑未登录人名的识别实验, 即在人名的统计中, 只统计词典中没有收录人名的识别性能. 前两组实验是封闭测试, 训练集和测试集相同; 第三组为开放测试, 训练集为《人民日报》1998 年 1 月 1 日至 2 月 19 日的新闻语料. 三组实验的结果如表 2 所示.

表 **2**　未登录人名识别的实验结果

测试类型	新闻日期	测试集大小(KB)	实际人名数	识别出的人名数	正确识别数	准确率(%)	召回率(%)	F-1 值(%)
封闭测试 1	98.1.1~98.1.31	8621	13722	17167	13376	77.92	97.48	86.61
封闭测试 2	98.2.1~98.2.20	6185	7534	10646	7489	70.35	99.29	82.35
开放测试	98.2.20~98.2.28	2605	3149	4130	2886	69.88	91.65	79.30

注: 语料库均来自于《人民日报》.

从表 2 中，我们可以看到：基于角色标注人名识别的召回率在封闭测试的情况下能达到 97.48%，甚至是 99.29%，开放测试的召回率也接近 92%. 而现在的一些解决方案仅为 68.77%[1,11]，最近一些方法小规模测试的召回率一般也很难达到 90%[2,3,5,9]. 对人名识别来说，召回率比准确率更加重要，因为低召回率就意味着没有办法再作后续的补救措施，而准确率完全可以通过限制条件或者后续处理（如词性标注、句法分析等）等手段将错误的人名排除掉，从而提高最终的准确率. 目前我们采用的限制规则仅仅是淘汰部分外国人名，如果再增加一些有效的消除规则，或者集成到整个词法分析中，准确率还有很大的上升空间.

训练集无论有多大，它毕竟是个封闭的有限集. 因此我们引入了数据平滑处理机制，使得基于角色标注的方法还可以成功识别出训练样本集涵盖范围之外的很多人名. 基本思路是：针对中国人名识别词典中没有收录的一些字词，我们猜测其可能的角色，然后赋予一个较小的概率，进行角色标注并最终识别人名. 例如："珺"在中国人名识别词典中并没有被收录，但我们会对"记者夏珺"中"珺"的角色进行各种可能的猜测，最后"珺"作为角色 D（单名）的前提下，该句才能取得最大概率的角色序列. 此时，"夏珺"的角色为"BD"（姓＋单名），因此可以被识别为单姓单名的人名.

4.3 ICTCLAS 与人名识别

在我们研制的计算所汉语词法分析系统 ICTCLAS[①]（Institute of Computing Technology Chinese Lexical Analysis System）中，我们应用了基于角色标注的中国人名识别方法. 为了评测人名识别与词法分析的关系. 在 11.08049 万词的开放测试集上，我们对 ICTCLAS 进行了一组对比实验：

（1）BASE：没有应用人名识别的 ICTCLAS.

（2）PERSON：应用基色标注人名识别之后的 ICTCLAS.

对比实验结果如表 3 所示.

表 3 ICTCLAS 应用中国人名识别前后对比实验

类别	切分正确率(%)	词类标注正确率(%)	人名识别正确率(%)	人名识别召回率(%)	人名识别 F-1 值(%)
BASE	96.55	93.93	16.32	94.91	27.85
PERSON	97.96	95.34	95.57	95.23	95.40

注：测试集大小：11.08049 万词；人名：15888 个；
切分正确率＝切分正确的词数/总词数×100%；
词类标注正确率＝词类标注正确数/总词数×100%；
这里提供的人名识别指标包含了核心词典中收录的人名.

表 3 的数据表明：

（1）人名识别实际地提高了汉语分词和词类标注的正确率. 在我们的实验中，中国人名识别应用后，ICTCLAS 的切分正确率和词类标注正确率同时提高了 1.41%.

（2）词法分析提高了人名识别的最终性能. 中国人名识别应用在 ICTCLAS 之后，人名识别的 F-1 值提高了差不多 15%，一部分原因是核心词典中已收录人名的加入，但更主要的因素是词法分析帮助人名识别排除了很多不合理的候选结果，提高了正确率. 例如："刘庄的水很甜"，在人名识别阶段很可能将"刘庄"错误的识别为人名，而在整个词法分析候选结果集当中，"刘庄"作为地名时，整个分析结果的概率大于其为人名的情况，所以最终的词法分析结果会将这种错误的情况排除掉.

5 结 论

本文系统地研究了中国人名的多种构成形式以及交叉成词的各种情况，分析了目前解决方案中激活机制和候选姓名选取的本质缺陷. 针对实际问题与已有方法的不足，作者提出了一种基于角色标注的中国人名识别方法. 即采用 Viterbi 算法，利用中国人名构成角色表及其相关统计信息，对句子中的不同成分进行角色标注，在角色序列的基础上，进行模式最大匹配，从而识别出中国人名. 中国人名构成角色指的是各个分词片断在人名识别过程中所扮演的不同角色，如姓、单名、上下文等. 某个词作为特定角色的概率以及角色之间的转移概率，全部从训练语料库中自动抽取，从而降低了人工总结规则的高成本与内在缺陷. 角色的标注过程就是选取最大概率的角色序列过程，避免了以前方法盲目触发的不足. 通过对大规模完全真实语料库的封闭与开放测试，该方法取得了相当好的效果，我们还将基于角色标注的人名识别算法集成到计算所汉语词法分析系统 ICTCLAS 中，较大地提高了词法分析效果，同时也促进了人名识别的综合性能. 各种对比实验表明基于角色标注的人名识别算法是行之有效的，能满足实际的需求. 我们下一步的研究工作是将基于角色标注的方法推广到对中国地名、译名、缩略语等其它未登录词的识别.

致 谢 中国科学院计算技术研究所软件室的程学

① 该系统全部的源码和文档均可在中文自然语言处理开放平台（www.nlp.org.cn）上自由下载.

旗主任和首席科学家白硕研究员对本文的工作给予了悉心的指导.知识挖掘组的张浩、李继锋、俞鸿魁、邹纲、 国栋等同事对本文的完成提出了很多有益的建议,在此一并表示感谢.最后,特别感谢评审老师细致耐心的审查.

参 考 文 献

1 Ji Heng, Luo Zhen-Shen. Inverse name frequency model and rules based on Chinese name identifying. In: Huang Chang-Ning, Zhang Pu ed.. Natural Language Understanding and Machine Translation. Beijing: Tsinghua University Press, 2001, 123~128(in Chinese)
(季 罗振声. 基于反比概率模型和规则的中文姓名自动辨识系统. 见:黄昌宁,张普编. 自然语言理解与机器翻译. 北京:清华大学出版社, 2001, 123~128)

2 Lv Ya-Juan, Zhao Tie-Jun *et al*.. Levelled unknown Chinese words resolution by dynamic programming. Journal of Chinese Information Processing, 2001, 15(1): 28~33(in Chinese)
(吕雅娟 赵铁军等. 基于分解与动态规划策略的汉语未登录词识别. 中文信息学报, 2001, 15(1): 123~128)

3 Sun Mao-Song *et al*.. Automatic recognition of Chinese personal names. Journal of Chinese Information Processing, 1994, 8(2)(in Chinese)
(孙茂松等. 中文姓名的自动辨识. 中文信息学报, 1994, 8(2))

4 Luo Zhi-Yong, Song Rou. Integrated and fast recognition of proper noun in modern Chinese word segmentation. In: Ji Dong-Hong ed.. In: Proceedings of International Conference on Chinese Computing, Singapore, 2001, 323~328(in Chinese)
(罗智勇,宋 柔. 现代汉语自动分词中专名的一体化、快速识别方法. 见: Ji Dong-Hong 等编. 国际中文电脑学术会议, 新加坡, 2001, 323~328)

5 Zhen Jia-Heng, Liu Kai-Ying. Discussion on strategy of surname and personal name processing in Chinese word segmentation. In: Chen Li-Wei ed.. Research and Application of Computational Linguistics. Beijing: Beijing Institute of Linguistics and Culture Press, 1993(in Chinese)
(郑家恒 刘开瑛. 自动分词系统中姓氏人名的处理策略探讨. 见:陈力为编. 计算语言研究与应用. 北京:北京语言学院出版社, 1993)

6 Song Rou, Zhu Hong *et al*.. Approach of personal name recognition based on corpus and rules. In: Chen Li-Wei ed.. Research and Application of Computational Linguistics. Beijing: Beijing Institute of Linguistics and Culture Press, 1993(in Chinese)
(宋 柔,朱 宏等. 基于语料库和规则库的人名识别法. 见:陈力为编. 计算语言研究与应用. 北京:北京语言学院出版社, 1993)

7 Wang Sheng, Huang De-Gen, Yang Yuan-Sheng. Chinese person name recognition based on mixture of statistics and rules. In: Huang Chang-Ning, Dong Zhen-Dong ed.. Corpora of Computational Linguistics. Beijing: Tsinghua University Press, 1999 (in Chinese)
(王 省,黄德根,杨元生. 基于统计和规则相结合的中文姓名识别. 见:黄昌宁,董振东编. 计算语言学文集. 北京:清华大学出版社, 1999)

8 Chen Xiao-He. Automatic Analysis of Modern Chinese. Beijing: Beijing University Linguistics and Culture Press, 2000, 104~114 (in Chinese)
(陈小荷. 现代汉语自动分析. 北京:北京语言文化大学出版社, 2000, 104~114)

9 Rabiner L. R.. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE, 1989, 77(2): 257~286

10 Rabiner L. R., Juang B. H. An introduction to hidden Markov models. IEEE Acoustics, Speech & Signal Processing Magazine, 1986, 3: 4~166

11 Zhang Hua-Ping, Liu Qun. Model of Chinese words rough segmentation based on N-shortest-paths method. Journal of Chinese Information Processing, 2002, 16(5): 1~7(in Chinese)
(张华平,刘 群. 基于 N-最短路径的中文词语粗分模型. 中文信息学报, 2002, 16(5): 1~7)

ZHANG Hua-Ping, born in 1978, Ph. D. candidate. His research interests include computational linguistics, Chinese information processing and information extraction.

LIU Qun, born in 1966, associate professor, Ph. D. candidate. His research interests include machine translation, natural language processing and Chinese information processing.

Background

The project "Unified and self-adaptive Chinese lexical analysis based on Cascaded HMM" aims to utilize a general model to accomplish all steps in lexical analysis, including word segmentation, disambiguation, unknown words recognition and part-of speech (POS) tagging. They have accomplished a Chinese lexical analysis system ICTCLAS. It ranked top in the official evaluation. It achieved two first positions in the First International Word Segmentation Bakeoff held in 2003 by the 41st Annual Meeting of the Association for Computational Linguistics. As a free project in Chinese NLP platform and a free product in Institute of Computing Technology, CAS, ICTCLAS was popular with research and industry. Over 3,000 copies were licensed in China, Japan, Singapore, and other nations. This paper solved the problem with unknown Chinese personal name recognition in lexical processing.