汉语人名识别：考虑上下文信息的识别方法

张华平、刘群，2004，基于角色标注的中国人名自动识别研究，《计算机学报》Vol.27, No.1，pp.85-91

詹卫东

zwd@pku.edu.cn

提 纲

1. 引言
2. 基于角色标注的中国人名自动识别方法
3. 自动识别的实现算法
4. 实验结果与分析
5. 结论

1. 引言

1.1. 中国人名自动识别的困难

1.1.1. 人名构成模式的多样性

1.1.2. 人名内部包含了普通词汇

1.1.3. 人名与其上下文字符可组合成普通词汇

1.1.4. 人名的歧义理解

1.2. 现有解决方案及其不足

1.1.1. 人名构成模式的多样性

人名的构成模式	示例
姓 + 名	张无忌、乔峰、西门吹雪、诸葛亮
有名无姓	“春花点点头”；“杰，你好吗？”
有姓无名	刘称赵已离开江西
姓 + 前后缀	邵公、沈叔、小李、邱某
港澳台等地已婚妇女姓名	范徐丽泰、彭张青

1.1.2. 人名内部包含了普通词汇

姓与名、名与名形成的字符串是一个已经被词典收录的词：

ex.1 “人散后[宝玉]回到怡香院”；

ex.2 [王国] 维、[高峰] 、[汪洋] 、张[朝阳]

张华平、刘群（2004）统计了 8 万 条人名，其中 6.89% 的人名内部包含普通词汇。

1.1.3. 人名与其上下文字符可组合成普通词汇

人名的开头（姓或名的首字）与其上文成词：

ex.1 “这里[有关]天培的壮烈事迹”

人名的结尾（姓或名的末字）与其下文成词：

ex.2 “费孝[通向]人大常委会提交书面报告”

1998年1月份《人民日报》标注语料库（200万字）中，上述情况有大约200 例。

1.1.4. 人名的歧义理解

同源歧义：

ex.1 “河北省刘庄” “刘庄” 中国人名 | 地名

ex.2 “周鹏和同学” “周鹏”和“周鹏和”

1.2.1. 现有解决方案

1. **规则方法：** 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。

2. **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，其中大于某一阈值的字段识别为中国人名。（主要是针对人名内部用字的统计规律来猜测一个字符串作为人名的可能性）

3. **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。

1.2.2. 不足

1. 单点(首或尾）激活机制的不足：

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识别。这样往往会**丢失那些不具备明显特征的姓名**，如 “有名无姓”的情况。

2. 将单字碎片作为姓名候选的不足：

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很难召回**。

3. 采用规则机制进行人名识别的不足：

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**。

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

2.2. 角色自动标注与中国人名识别

2.3. 角色信息的自动抽取

2.1. 中国人名的构成角色：15 种

标记	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、邵公、沈叔
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

人名模式（12 种）

模式	意义	例子
BBCD	复姓 + 双名	欧 阳 一 休
BBE	复姓 + 单名	欧 阳 休
BBZ	复姓 + 双名（成词）	欧 阳 胜 利
BCD	单姓 + 双名	张 浩 平
BE	单姓 + 单名	刘 邦
BG	单姓 + 后缀	吴 妈
BXD	姓1 + 姓2 + 名1 + 名2	欧 阳 平 南
BZ	单姓 + 双名（成词）	张 朝 阳
CD	名1 + 名2	志 雄
FB	前缀 + 姓	老 何
Y	单姓 + 名（成词）	汪 洋
XD	（成词）姓 + 名1 + 名2	王 国 维

2.2. 角色自动标注与中国人名识别

2.2.1. 角色自动标注
2.2.2. 中国人名识别

} 计算模型 与 计算过程

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。

采用Viterbi 算法来实现角色自动标注：

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。

$$T^{\#} = \operatorname{argmin}_T \left(-\sum_{i=1}^{m} \left(\ln p\left(w_i|T_i\right) + \ln p\left(T_i|T_{i-1}\right)\right)\right)$$

从 角色标注 到 人名识别

角色U和V要进行分裂处理

U → K B　　现任/A 主席/A 为何/U 鲁/C 丽/D → 为/K 何/B 鲁/C 丽/D

V → D L　　和/M 邓/B 颖/C 超生/V 前/A → 邓/B 颖/C 超/D 生/L 前/A

V → E L　　王/B 平等/V 领导/A → 王/B 平/E 等/L 领导/A

经过分裂处理后得到的角色序列，在人名模式集中进行模式匹配，最大匹配的子串即为人名识别结果。

上述3例的匹配结果为：K + BCD
BCD + L
BE + L

从 角色标注 到 人名识别 示例

馆/ 内/ 陈列/ 周/ 恩/ 来/ 和/ 邓/ 颖/ 超生/ 前/ 使用/ 过/ 的/ 物品/

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

馆/A 内/A 陈列/K 周/B 恩/C 来/D 和/M 邓/B 颖/C 超/D 生/L 前/A 使用/A 过/A 的/A 物品/A

标记	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、邵公、沈叔
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

2.3. 角色信息的自动抽取

在大规模语料库训练的前提下，根据大数定理，可以得到：

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

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

以上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

3.1. 角色信息自动抽取算法

3.2. 中国人名的识别流程

3.1. 角色信息自动抽取算法

在ICTPOS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。

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，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）。

ICTPOS：中科院计算所词性标记集
训练语料：北大计算语言所开发的人民日报词性标注语料库（北大ICL-POS）

3.1. 角色信息自动抽取算法

在ICTPOS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。

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. 根据人名的5种类别，分别对姓、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。对于人名内部成词的情况，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$C(t_{i-1}, t_i)$。

角色信息自动抽取示例

1

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 , /w 千/m 里/q 淮河/ns 传来/v 喜讯/n

2

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

3

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

1 PKU-ICL 标注语料　　2 ICT 标注语料　　3 角色信息标注语料

3.2. 中国人名的识别流程

1. 对句子进行分词，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

4.1. 测试集与评测指标

4.2. 未登录人名的识别评测实验

4.3. ICTCLAS 与人名识别

4.1. 测试集与评测指标

4.1.1. 测试集

类型一	只含人名的句子集	人名模式少，误判机会（干扰）少。 测试成绩偏高
类型二	完全真实的语料	人名模式多，90%的句子不含人名（误判机会大）。**测试成绩更能说明问题**

封闭测试	测试集是训练集的子集
开放测试	测试集与训练集之间没有包含和被包含的关系

4.1.2. 评测指标

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 %

综合指标（F-measure）：

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

其中β是准确率P和召回率R之间的权衡因子，β大于1，重视P，反之，重视R。如果认为P和R同等重要，则β取1，此时F称为F-1值。

$$F = \frac{2PR}{P+R}$$

4.2. 未登录人名的识别评测实验

为了客观评价基于角色标注的中国人名识别算法，做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。结果均为中准确率、高召回率。

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

语料均来自《人民日报》分词和词性标注语料库（1998年1、2月份）
开放测试的训练集为《人民日报》1998 年1 月1 日至2月19 日的新闻语料

4.3. ICTCLAS 与人名识别

为了评测人名识别与词法分析的关系，在计算所汉语词法分析系统ICTCLAS中应用基于角色标注的中国人名识别方法进行测试。

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

- BASE : 没有应用人名识别的ICTCLAS.
- PERSON : 应用基色标注人名识别之后的ICTCLAS.
- 测试集大小: 11. 08049 万词;人名:15888 个;

- 人名识别对ICTCLAS的影响： 96.55% → 97.96% 93.93% → 95.34%
  切词正确率 词性标注正确率

5 结论

- **把人名识别（分词问题）转化为角色标注问题。**
- **利用现有的语料库资源和成熟的Viterbi算法实现标注过程。**
- **这种方法可以推广到其他未登录词（地名、机构名、译名等）的识别。**