# 利用 CRF 实现中文人名性别的自动识别

赵晓凡[1], 赵 丹[2], 刘永革[1]
(1 安阳师范学院 计算机与信息工程学院, 河南 安阳 455002; 2 郑州大学 计算中心, 河南 郑州 455000)

**摘 要**: 受传统观念的影响, 中国人名最后一个或两个字的用法对性别判定通常具有一定的指示作用, 由此提出利用条件随机场模型来实现中文人名性别的自动识别. 该机器学习方法根据人名的结构和用字信息, 构建人名标注集, 选择 6 组不同的特征模板集, 利用条件随机场模型, 在 231 337 个人名数据库中经过封闭测试, 正确率可以达到 89.30%, 比采用朴素贝叶斯依赖人名用字进行性别识别的方法好将近 7 个百分点. 实验证明: 在人名库中识别性别, 名字尾字的作用要高于姓氏用字, 且女性人名性别识别的准确度要略高于男性, 一般是高 2 至 3 个百分点, 从机器学习的角度来说性别差异可以体现在人名用字中. 通过分析实验数据总结了适合人名识别的 CRF 特征模板设计的一般规律, 这为后续的研究工作提供了基础.

**关键词**: 性别识别; 中文人名要素; 命名实体识别; 特征选择; 条件随机场
**中图分类号**: TP391 **文献标识码**: A **文章编号**: 1000－7180(2011)10－0122－03

# The Automatic Gender Recognition of Chinese Name Using Conditional Random Fields

ZHAO Xiao-fan[1], ZHAO Dan[2], LIU Yong-ge[1]
(1 School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China;
2 The Computing Center of Zhengzhou University, Zhengzhou 455000, China)

**Abstract**: On the influence of traditional concept, the last one or two words of Chinese name usually has a certain instructions role to gender recognition. Gender recognition of person name can be used in natural language processing which is a specific application of Named Entity Recognition. Gender recognition method makes use of the structure and vocabulary information of Chinese personal name. The experiment on the basis of CRF is designed by constructing person name annotation set and selecting suitable feature model using NLP technology. Through the closed test on 231337 person names 89.30% accuracy is got which is about seven percentages higher than the bayes method. The experiment proves that the effect of the last name in gender recognition is higher than the role of the first name and the accuracy of gender recognition in female names is more higher than male names, about two or three percentages. Gender differences based on machine learning can be found from the names itself. The general principle of template design was proposed.

**Key words**: Gender recognition; Chinese name element; named entity recognition; feature selection; conditional random fields

## 1 引言

目前命名实体识别(Named Entity Recognition, NER)仍然是中文信息处理的难点和热点, 对未登录词的处理, 其结果往往很难满足需求. 人名的出现是影响未登录词识别正确率的关键[1]. 在《人民日报》1998 年 1 月的语料库(共计 2 305 896 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 61.34%的未登录词是人名[2].

文献[3] 指出语言对男性和女性的描写是不同

收稿日期: 2011－04－28; 修回日期: 2011－06－25
基金项目: 国家自然科学基金(60875081); 河南省教育厅高等学校青年骨干教师项目(2009GGJS－108)

的. 人名作为一个符号, 对单个的个体应该具有很高的辨识度. 按照日常生活的经验和人们的习惯, 起名用字往往与性别相关. 文献[4] 对 7 万中国人名的 90 个常用尾字进行非参数检验后认为: 男女人名用字有显著性别差异. 文献[2] 在基于人名用字的性别识别中, 用朴素贝叶斯方法对比了三种人名用字模型, 对 10 万人名的实验结果表明, 人名尾字对性别识别具有更好的应用能力, 开放测试准确率为 82.95%. 文章是命名实体识别的一个应用, 尝试用条件随机场(Conditional Random Fields)模型对人名用字的男女性别差异按照机器学习的方法进行训练, 分析由人名自动识别性别的可能性与可行性, 封闭测试结果要好于文献[2] 中的方法.

## 2　条件随机场理论

CRF 是一种无向图模型或者马尔可夫随机域, 它采用一阶链式无向图结构计算给定观察值条件下输出状态的条件概率[5]. 如图 1 所示.

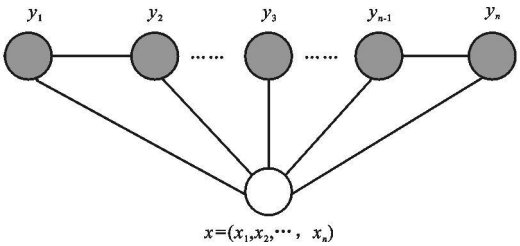


图 1　线链 CRFs 的图形结构

设 $O=\{o_1, o_2, \cdots, o_T\}$ 表示被观察的输入字串序列, $S=\{s_1, s_2, \cdots, s_T\}$ 表示将被预测的词位标记序列, 则在给定一个输入字串序列的情况下, 对参数为 $\Lambda=\{\lambda_1, \lambda_2, \cdots, \lambda_K\}$ 的线链 CRF, 其输出的词位序列的条件概率为

$$P_\Delta(S \mid O)=\frac{1}{Z_O}\exp\left[\sum_{t=1}^{r}\sum_{k=1}^{k} {}_k f_k(s_{t-1}, s_t, o, t)\right] \tag{1}$$

式中, $Z_O$ 是归一化因子, 定义为

$$Z_O=\sum_S \exp\left[\sum_{t=1}^{T}\sum_{k=1}^{K} {}_k f_k(s_{t-1}, s_t, o, t)\right] \tag{2}$$

$f_k(s_{t-1}, s_t, o, t)$ 是一个任意的特征函数, 用于表达上下文可能的语言特征, 通常是一个二值表征函数, 表示如下:

$$f_k(s_{t-1}, s_t, o, t)=\begin{cases}1, & \text{如果满足条件}\\ o, & \text{否则}\end{cases} \tag{3}$$

式中, ${}_k$ 是一个需要被学习的参数, 其对应于每一个特征函数的权值, 取值范围可以是 $-\infty$ 到 $+\infty$. 给定一个由式(1) 定义的条件随机场模型, 对任意的输入字串, 其最佳词位标记序列应满足式(4):

$$S^* = \arg\max_S P_\Delta(S/O) \tag{4}$$

要求出使得 $P_\Delta(S/O)$ 最大的标记序列 $S^*$, 可以使用维特比算法进行计算.

## 3　基于条件随机场的中文人名性别的自动识别

### 3.1　中文人名要素标注

基于字标注的命名实体识别方法将对词知识的学习转换成字串的标注过程, 由于每个字在构造一个特定的词语时都占据一个构词位置, 也就是字位, 因此可以把识别过程看成是学习这个字位信息的机器学习过程, 按字抽取特性, 最后对每一个字进行分类识别.

中文姓名一般由姓氏和名字两部分构成, 形式为"姓氏名字", 即姓氏在前, 名字在后, 姓氏和名字一般分别由一个或两个字构成, 标识出姓氏和名字中用字的不同位置, 组合之后可以得到四种形式[6]:

(1)单姓单名: 杨磊、王燕;

(2)单姓双名: 郑爱霞、李秋生;

(3)复姓单名: 诸葛亮、上官路;

(4)复姓双名: 完颜宏帆、上官宝珍.

文中使用的真实姓名只考虑以上四种情况, 所以中国姓名最长由四个字构成, 最短由两个字构成. 采用 B1、B2、I、E 四个标注符号组块的方法来标识人名要素. 即 B1 表示人名中的第一个字, B2 表示人名中的第二个字, I 表示除了第一个和第二个字之外的其他字, E 表示人名中的最后一个字. 其中男性用 M 表示, 女性用 F 表示, 加入到标注符号作为后缀, 形如: B1－F, E－M. 则古欣、孟亚男、端木顺阳应分别标识成古/B1－F 欣/E－F; 孟/B1－M 亚/B2－M 男/E－M; 端/B1－M 木/B2－M 顺/I－M 阳/E－M.

### 3.2　特征模板的选取

CRF 是一种有监督的机器学习方法, 因此设计能充分体现中文人名要素序列特性的特征模板是决定训练模型识别性能优劣的关键. 考虑人名的长度, 窗口不用开的太大, 一般选取大小为 3 的观察窗口, 即 $(w_{-1}, w_0, w_1)$. 根据人名本身结构和用字信息, 可用模板对特征进行筛选, 经过反复斟酌, 选定 Temp10B、Temp10、Temp6B、Temp6、Temp2B 和 Temp2 共 6 组特征模板集进行实验.

## 4　实验及结果分析

实验中所采用的全部是真实的人名数据, 取自

某家省级医院一年内的所有病患的姓名和性别. 由于人名语料取自医院, 基本格式不符合 CRF 训练的要求, 经专用的转换工具进行相应的预处理之后, 抽取其中符合实验要求的人名数据, 并且统计得到具体的实验用数据情况如表 1 所列, 实验结果如表 2 所示.

表 1 实验相关的 23 万人名分散情况

| 实验数据 | 大小/kB | 男性人名 | 女性人名 | 二字人名 | 三字人名 | 四字人名 |
|---|---|---|---|---|---|---|
| 训练语料 | 161940/1528 | 80538 | 81402 | 27579 | 134278 | 0 81 |
| 测试语料 | 69397/648 | 31788 | 37609 | 15651 | 53724 | 22 |

表 2 基于 CRF 的人名性别识别实验结果

| 特征模板 | Accuracy% | Precision% | Recall% | FB1 |
|---|---|---|---|---|
| Temp10B | 89.00 | 88.64 | 88.64 | 88.64 |
| | F | 90.38 | 88.45 | 89.41 |
| | M | 86.68 | 88.86 | 87.75 |
| Temp10 | 89.91 | 81.81 | 87.21 | 84.42 |
| | F | 83.95 | 86.70 | 85.30 |
| | M | 79.45 | 87.81 | 83.42 |
| Temp6B | 89.06 | 88.70 | 88.70 | 88.70 |
| | F | 90.46 | 88.49 | 89.46 |
| | M | 86.72 | 88.96 | 87.83 |
| Temp6 | 86.56 | 63.56 | 75.96 | 69.21 |
| | F | 66.21 | 75.06 | 70.36 |
| | M | 60.77 | 77.02 | 67.94 |
| Temp2B | 89.30 | 88.92 | 88.92 | 88.92 |
| | F | 90.55 | 88.82 | 89.68 |
| | M | 87.07 | 89.03 | 88.04 |
| Temp2 | 90.05 | 82.13 | 87.08 | 84.53 |
| | F | 84.22 | 87.03 | 85.60 |
| | M | 79.79 | 87.15 | 83.31 |

从实验结果可以看出, 基于条件随机场的人名性别识别效果还是不错的, 大致和人们按照习惯由人名估计性别的正确率相同, 而且 6 组实验中均为女性人名的识别准确率要高于男性人名的识别准确率, 一般是高 2 至 3 个百分点, 综合评价采用 Temp2B 的效果是最好的, 准确率达到了 89.30%, 特别是对女性的识别准确率达到了 90.55%, 由此分析出来对人名的性别识别比较有用的是尾字, 这和用贝叶斯方法得到的结论相同, 并且姓名用字的下文较上文的有效范围更大些, 人名性别的自动识别是可以用这种机器学习的方法来实现的.

分析实验中错误的性别标注, 发现: (1)并非是人名中的所有的字都标错, 有相当一部分三字人名都是名字中的一个或者两个字标错, 比如: 魏清礼是男性, 正确标注为 B1—M, B2—M, E—M, 测试结果却把名字用字标成女性即 B2—F, E—F, 而姓氏的标注正确; (2)错误率较高的标注主要是出于一些中性字, 即男女名字中都有可能出现的, 比如说贾楠是女性, 由于名字用字较为中性化, 所以测试结果给标注成了男性. 以上两种错误的解决办法, 首先考虑是否可以通过修改模板来改进, 本次实验的特征模板窗口的移动是不对称的, 并且还未加入 Bigram 转移特征; 其次对于中性化字, 可以加入人名用字的统计概率做规则来提高标注的正确率. 这些都有待于进一步的实验.

## 5 结束语

基于条件随机场的中文人名的自动识别是中文文本处理的一个应用. 特别是现在网上有很多起名软件, 人名性别的自动识别可以辅助判断这个名字适合男性用还是女性用. 下一步的工作需要整合在实验过程中用到的语料处理和统计脚本程序, 达到一体化的应用程序界面, 更加方便以后实现具体的应用.

参考文献:

[1] 钱进. 姓名用字的性别差异统计分析[J]. 常州工学院学报, 2004, 17(5): 60—62.

[2] 郎君, 秦兵, 刘挺, 等. 中国人名性别自动识别[C] //第三届学生计算语言学研讨会, 沈阳: 2006.

[3] 于丽丽, 丁德鑫. 基于条件随机场的古汉语词义消歧研究[J]. 微电子学与计算机, 2009, 26(10): 45—48.

[4] 张华平, 刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报, 2004, 27(1): 85—91.

[5] Carreras X, Marquez L, Padro L. Named entity extraction using Ad— aBoost[C]//Proceedings of the Sixth Conference on Natural Language Learning, stroudsburg, RA, USA: Association for Computational Liquistics, 2002: 167—170.

[6] 钱进. 语言性别差异研究综述[J]. 甘肃社会科学, 2004(6): 47—50.

[7] 董银秀. 语言中的性别因素[J]. 兰州工业高等专科学校学报, 2004, 11(1): 66—71.

(下转第 128 页)

式中，$\boldsymbol{C}$ 表示特征向量协方差矩阵.

马氏距离注意到样品的统计特性，排除了样品间的相关性影响，例如，若某一模式的 9 个分量反映了特征 A，而仅有 1 个分量反映了特征 B. 若用欧氏距离，则反映了特征 A 的距离，而用马氏距离，则可避免此缺点. 当协方差矩阵 $\boldsymbol{C}$ 为对角矩阵时，各特征分量就互相独立. 只有当 $\boldsymbol{C}$ 为单位矩阵 $\boldsymbol{I}$ 时，马氏距离才与欧氏距离相等.

当特征向量的各分量间没有相关性，马氏距离还可以进一步简化，因为这时只需要计算每个分量的方差，简化后的马氏距离如下所示：

$$D(I,J)=\sum_{i=1}^{N}(F_I-F_J)^2/c_i \qquad (3)$$

图 9 所示为系统实现车牌识别的功能界面.

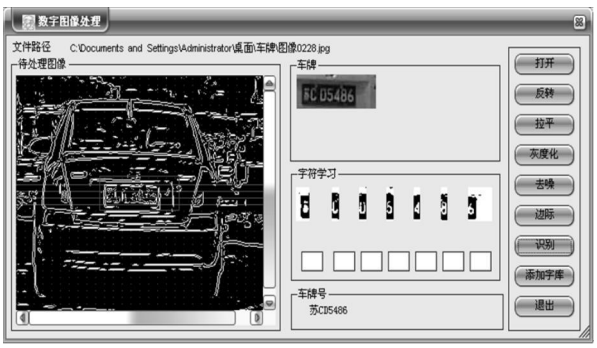


图 9 车牌识别功能界面

## 7 结束语

在广泛阅读了国内外现有的关于图像无线传输、车牌自动识别和字符识别的文献，比较和借鉴现有成功的车牌识别和字符识别所采用的方法后，本文研究实现了基于移动通信和图像处理技术的车辆牌照手机自动识别系统，主要研究实现了以下功能：

（1）手机抓拍的车牌图像以彩信的方式通用 GSM 网络发送到图像处理服务器.

（2）利用基于先验知识的投影法实现车牌字符分割.

（3）应用基于距离的分类器的方法实现了车牌字符特征的选择与提取、分类与学习.

（4）正常情况下本系统识别率可达 85%，系统响应时间小于 20 s.

参考文献：

[1] 陈振学，汪国有，刘成云，等. 一种新的车牌图像字符分割与识别算法[J]. 微电子学与计算机，2007，24(2)：42－44.

[2] 郭金发，张龙. 短信与 BREW 开发技术及实践[M]. 西安：西安电子科技大学出版社，2005.

[3] Chen Xilin，Yang Jie，Zhang Jing，et al. Automatic detection and recognition of signs from natural scenes[J]. IEEE TRANSACTIONS ON Image PRO cessing，2004 13(1)：87－99.

[4] 路小波，张光华. 基于二值图像的车牌精确定位方法[J]. 东南大学学报：自然科学版，2005，35(6)：972－974.

[6] 李波，曾致远，周建中，等. 一种自适应车牌识别系统设计方法[J]. 微电子学与计算机，2009，26(4)：218－221.

[7] Garris M D，Wilson C L. Neural network－based systems for handprinted OCR applications[J]. IEEE Tran Image Processing，2006，7(8)：1097－1112.

[8] Chen Xilin，Yang Jie，Zhang Jing，et al. Automatic detection and recognition of signs from natural Scenes[J]. IEEE Transactions on images processing，2004，13(1)：87－99.

[9] 袁方，杨柳，张红霞. 基于 K－近邻方法的渐进式中文文本分类技术[J]. 计算机工程，2004，32(11)：88－91.

作者简介：

邵晓根　男，(1963－)，硕士，副教授. 研究方向为智能计算、智能信息处理.

(上接第 124 页)

[8] 张仰森，徐波，曹元大，等. 基于姓氏驱动的中国姓氏自动识别方法[J]. 计算机工程与应用，2003(4)：62－65.

[9] Sun Jian，Gao Jianfeng，Zhang Lei，et al. Chinese named entity i－dentification using class－based language model[C]// proceedings of 19th international conference on Computational for Computational Linquistics. Stroudsburg，PA，USA：Association for Computational 2002.

[10] 黄德根，杨元生，王省，等. 基于统计方法的中文姓名识别[J]. 中文信息学报，2000，15(2)：31－36.

作者简介：

赵晓凡　女，(1981－)，硕士研究生. 研究方向为自然语言处理，信息安全.