情报学报 ISSN 1000-0135
第28卷第1期 88-96，2009年2月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.28 No.1，88-96 February 2009

基于CRFs的角色标注人名识别模型在网络舆情分析中的应用[1)]

王 昊 苏新宁
（南京大学信息管理系，南京 210093）

摘要 本文在理论分析CRFs优于HMM和MEM等序列标注模型的基础上，提出一种基于CRFs的字角色标注人名识别模型。重点阐述了该模型的构建过程，包括角色定义、特征模板建立、特征函数生成及其参数训练、角色标注和基于模式的人名抽取等步骤，并通过实验验证模型的识别效果，探讨包括特征组合、字长窗口等在内的各种影响因素，探索模型的最佳识别条件，同时对CRFs和HMM在人名识别实验中进行了比较分析，认为CRFs在付出更大的实验复杂度的代价下，其人名识别效果明显优于HMM。论文最后通过实例探讨了CRFs_RL_PnR模型在网络舆情分析，包括新闻人物自动抽取、焦点人物时序分析等中的实践应用。

关键词 条件随机场 字角色 特征模板 模式匹配 网络舆情分析

Model for Person Name Recognition Based on Role Labeling Using CRFs and Its Application to Web Opinion Analysis

Wang Hao and Su Xinning
(*Information Management department of Nanjing University, Nanjing* 210093)

Abstract This paper raises a model for identifying person name using role auto-labeling based CRFs(CRFs_RL_PnR), based on the theoretical analysis of the advantage of CRFs comparing with HMM and MEM, which are the models for sequence labeling. It emphasizes the process of the model building, including role defining, feature template building, feature function creating and its parameter training, role labeling and person name extracting based patterns. It tests the identified effect for the model by experiment, and discusses all kinds of influencing factors including combined feature, word length window and so on, for exploring the best condition of the model, while this paper also compares CRFs with HMM by experiment for person name identification, and gains the conclusion that when CRFs pays out more experiment complexity, the effect of person name identification is more excellent then ones of HMM. At last, this paper discusses how CRFs_RL_PnR model applying in web public feelings analysis, including the news persons auto extraction, focus persons time analysis and so on.

Keywords CRFs, character role, feature template, pattern matching, Web public feelings analysis

收稿日期：2007年10月11日

作者简介：**王昊**，男，1981年生，毕业于南京大学信息管理系，情报学博士，现就职于南京大学信息管理系，讲师，主要从事信息智能处理与检索、本体自动构建及应用、引文分析和评价等方面的研究。E-mail:ywhaowang810710@sina.com。**苏新宁**，男，1955年生，毕业于武汉大学信息管理学院，情报学硕士，南京大学信息管理系教授，博士生导师，主要从事信息智能处理与检索、引文分析和科学评价等方向的研究。

1） 本文系国家自然科学基金项目“电子政务中信息资源管理对政府辅助决策的研究”(70373028)研究成果之一。

1 引 言

随着网络技术的迅猛发展，信息开始呈现电子化、数字化和海量化的趋势。如何从海量的网络信息中抽取出用户感兴趣信息已经成为人们普遍关注的热点。非结构化文本中命名实体特别是人名实体的抽取是信息抽取中重要而实际的研究基础，它对于信息采集、信息处理以及信息服务等都具有重要意义。例如，抽取网络舆情信息中的人名可以了解近期的新闻人物，实现对焦点人物的时序跟踪，使决策者充分掌握事件主体等。然而人名实体的少规则性和无限扩充性，使基于非结构化文本的人名抽取成为了自然语言处理的难点，目前在学术界也提出了各种实验方法[1]：基于规则、基于统计以及规则和统计相结合等方法。其中基于统计方法是当前人名实体抽取较为有效的方法，隐马尔可夫模型（Hidden Markov Model，HMM）[2]、最大熵模型（Max Entropy Model，MEM）[3]等都是其常用的概率统计模型。

HMM 可以为观察序列从状态集合中选择具有最大可能性的状态序列；MEM 则是在 E.T.Jaynes 等提出的最大熵原理[4]基础上发展起来，在已知事实知识的约束下，熵最大的概率分布即为最合理的分布。然而 HMM 存在生成模型所固有的独立性假设和无法融合多种特征的缺陷；MEM 则因为局部归一化导致标记偏置（label bias）问题。上述缺点使得基于这些统计模型的人名抽取无法达到最佳效果。条件随机场则可以解决上述问题。

条件随机场（Conditional Random Fields，CRFs）[5,6]是一种典型的序列标注判别模型，是在给定观察序列的条件下，计算整个观察序列状态标记的联合条件概率分布的无向图模型。从理论上讲 CRFs 优于 HMM 和 MEM：①CRFs 采用条件概率建模，避免了 HMM 采用联合概率[7]所带来的独立性假设；②CRFs 继承了 MEM 可无限制扩充特征的思想，突破了 HMM 依赖于指定特征的局限；③CRFs 在 MEM 特征定义的基础上加入了之前的标记状态对当前状态标记的影响，这使得特征的选择更为合理；④CRFs 是对整个标记序列求解联合概率，在整个序列范围内归一化，避免了 MEM 因求解单个或局部观察值概率所带来的标记偏置问题。相对于 HMM 和 MEM 而言，CRFs 更适合解决序列标注问题。

本文试图从熟语料中提取中文人名的上下文特征，采用 CRFs 模型，基于字角色标注思想建立中文人名实体自动抽取模型，并通过实验探讨模型的合理性和有效性。为了验证模型的实用价值，论文特将其应用于网络舆情分析的实践中进行讨论。

2 基于 CRFs 的字角色标注人名实体识别

2.1 基于 CRFs 的字角色标注模型

基于 CRFs 的字角色标注人名实体识别模型（Model for Person-Name Recognition Based on Role Labeling Using CRFs，CRFs_RL_PnR，如图 1）的基本思路如下：首先以单汉字为标准对中文文本进行原子切分；其次基于“字”原始序列衍生相应的观察序列，根据观察序列的取值建立针对人名识别的上下文特征模板，并按这些模板规则从训练语料中提取特征函数；再次是根据训练语料中出现的人名构成

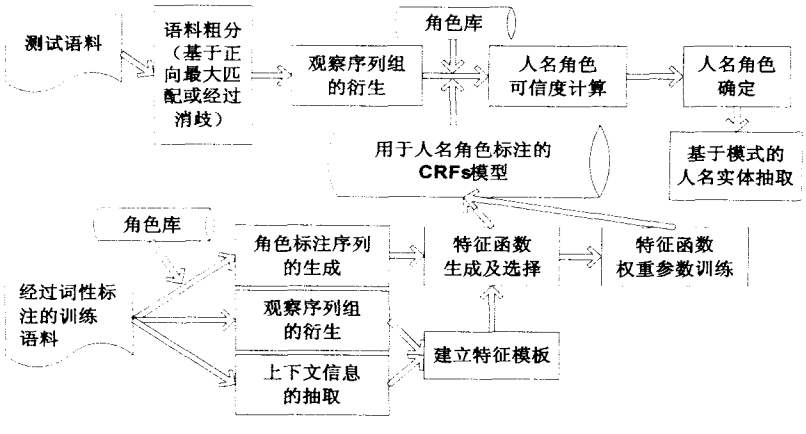


图 1 基于 CRFs 的字角色标注人名实体识别模型（CRFs_RL_PnR）

规则,定义其构成角色集合;然后利用训练语料中的观察序列和角色序列,基于 CRFs 方法训练特征函数权重以构建用于人名识别的 CRFs 字角色标注模型,通过调整观察序列取值、特征模板个数以及角色集合定义来调节特征函数权重以实现最合理的建模;根据建立的模型,计算测试语料中每一个(或组)观察值在其上下文环境中被识别为某种角色的联合可信度,并选择可信度最大的角色类别作为观察值的角色标注;最后根据人名实体的角色组合从原始观察序列中抽取出符合指定模式的人名实体。

本文采用的训练和测试语料来自《人民日报》1998年1月份经过标注的中文语料库,以1月份前24天共15 767句语料作为训练集,后7天共3717句的语料作为测试集。

2.2 人名角色的定义

考察中文语料库中的人名构成,发现中文人名在结构上具有一定的规律:①中国人名一般最多由四个字(姓+姓+名+名)组成;②中国人名的组合模式在数量上具有有限性;③音译外来人名在翻译用字上具有一定的重复性。根据上述特征,结合考虑本文对中文单"字"进行角色标注,可以总结出如表1所示的12种角色(R)集合。

需要说明的是,本文将在语料中出现的连续阿拉伯数字和连续英文字母作为原子单字处理。

2.3 特征模板的构建

仅以原子单字作为观察序列所能表现出来的特征,特别是能够表征人名角色特点的信息太少,需要考虑从原始观察序列中衍生出其他观察序列,同时结合上下文信息来反映出更多的特征。

(1)观察序列的衍生。笔者考虑到中文人名用字的重复性以及字词位信息等因素,从音译人名用字(F)、姓氏用字(S)、前(P)后(Q)临界字、前(B)后(E)缀以及词位用字(W)等5个特征来扩展观察序列。笔者先后收集了音译外来人名常用字集合、姓氏集合、人名前后临界字集合、前后缀集合,认为如果汉字在上述集合中出现,那么该汉字即具有上述相应特征;词位则是指当前"字"在包含其的"词"中所处的位置标记,常用的有2词位、4词位以及6词位等3种词位标注集合[8]。词位标记字符越多,表达能力越强,描述词语组合的精确度越高,词位标注是一种多值特征。通过观察序列的衍生,原有的单汉字序列被扩展为具有8个元素的复合观察序列组,为准确序列标注提供可靠的横向约束特征。

(2)上下文字长窗口特征的引入。除观察序列组外,当前对象的上下文信息也可以强化特征。常用的上下文信息有:局部上下文信息和远程上下文信息。局部上下文信息是指以当前观察对象为中心,向前或向后连续选取一定字长范围的上下文信息作为当前对象的特征,这个局部连续范围称为字长窗口,常用的有3字长和5字长窗口[8];远程上下文信息则是指与当前观察对象具有一定文本距离对象的特征,可以提供长距离约束,"词语触发对"就是典型的远程上下文信息[9]。本文采用字长窗口来纵向约束当前对象。

(3)特征模板的建立。根据观察序列组和字长窗口的特征约束,笔者建立了如表2所示的若干组 n 元(n-gram)特征模板(feature template)[10]。表中C表示当前字,F表示音译人名用字特征,S表示姓氏特征,P表示前临界字特征,Q表示后临界字特征,W(4)表示4词位标注特征,W(6)则表示6词位标注特征,R表示标注角色特征。

表2中,TMPT1、TMPT2、TMPT3旨在从观察序列组角度探讨观察序列个数变化对角色标注效果的影响;TMPT3和TMPT4则重点考察词位特征取值范围

表1 CRFs_RL_PnR中的中文单"字"角色集合

序	角色	解释	示例	序	角色	解释	示例
1	B	单姓字	梁国彪	7	F	姓前缀	小王,老张
2	X	复姓字	欧阳中石、司马徽	8	G	姓后缀	李总,王氏
3	C	双名首字	林华卿、王国栋	9	P	人名的上文	对克林顿来说
4	D	双名尾字	康新云、杨秀梅	10	Q	人名的下文	政委刘建新说
5	E	单名字	李茜、欧阳修	11	O	两个人名之间的字	刘伯承和邓小平
6	W	音译外来名	阿拉法特	12	Z	其他字	处理好人际关系

表 2　人名实体识别中的特征模板

模板名称	观察值特征	标注特征	特征模板
TMPT1	C	R	$C_n, n=-2,-1,0,1,2$
			$C_{n-1}C_n, n=-1,0,1,2; C_n\text{-}2C_n, n=0,1,2; R_{-1}R_0$
			$C_{n-2}C_{n-1}C_n, n=0,1,2$
TMPT2	C、F、S、W(4)	R	$C_n, F_n, S_n, W_n, C_nF_nS_nW_n, n=-2,-1,0,1,2$
			$C_{n-1}C_n, F_{n-1}F_n, S_{n-1}S_n, W_{n-1}W_n, n=-1,0,1,2; C_{n-2}C_n, F_{n-2}F_n, S_{n-2}S_n, W_{n-2}W_n, n=0,1,2; R_{-1}R_0$
			$C_{n-2}C_{n-1}C_n, F_{n-2}F_{n-1}F_n, S_{n-2}S_{n-1}S_n, W_{n-2}W_{n-1}W_n, n=0,1,2$
TMPT3	C、F、S、P、Q、B、E、W(4)	R	$C_n, F_n, S_n, P_n, Q_n, B_n, E_n, W_n, C_nF_nS_nP_nQ_nB_nE_nW_n, n=-2,-1,0,1,2$
			$C_{n-1}C_n, F_{n-1}F_n, S_{n-1}S_n, P_{n-1}P_n, Q_{n-1}Q_n, B_{n-1}B_n, E_{n-1}E_n, W_{n-1}W_n, n=-1,0,1,2; C_{n-2}C_n, F_{n-2}F_n, S_{n-2}S_n, P_{n-2}P_n, Q_{n-2}Q_n, B_{n-2}B_n, E_{n-2}E_n, W_{n-2}W_n, n=0,1,2; R_{-1}R_0$
			$C_{n-2}C_{n-1}C_n, F_{n-2}F_{n-1}F_n, S_{n-2}S_{n-1}S_n, P_{n-2}P_{n-1}P_n, Q_{n-2}Q_{n-1}Q_n, B_{n-2}B_{n-1}B_n, E_{n-2}E_{n-1}E_n, W_{n-2}W_{n-1}W_n, n=0,1,2$
TMPT4	C、F、S、P、Q、B、E、W(6)	R	同 TMPT3
TMPT5	C、F、S、P、Q、B、E、W(4)	R	$C_n, F_n, S_n, P_n, Q_n, B_n, E_n, W_n, C_nF_nS_nP_nQ_nB_nE_nW_n, n=-1,0,1$
			$C_{n-1}C_n, F_{n-1}F_n, S_{n-1}S_n, P_{n-1}P_n, Q_{n-1}Q_n, B_{n-1}B_n, E_{n-1}E_n, W_{n-1}W_n, n=0,1; C_{n-2}C_n, F_{n-2}F_n, S_{n-2}S_n, P_{n-2}P_n, Q_{n-2}Q_n, B_{n-2}B_n, E_{n-2}E_n, W_{n-2}W_n, n=1; R_{-1}R_0$
			$C_{n-2}C_{n-1}C_n, F_{n-2}F_{n-1}F_n, S_{n-2}S_{n-1}S_n, P_{n-2}P_{n-1}P_n, Q_{n-2}Q_{n-1}Q_n, B_{n-2}B_{n-1}B_n, E_{n-2}E_{n-1}E_n, W_{n-2}W_{n-1}W_n, n=1$
TMPT6	C、F、S、P、Q、B、E、W(4)	R	同 TMPT5，仅除去 R-1R0

的变化对角色标注效果的影响；TMPT3 和 TMPT5 的对比在于探讨字长窗口的大小对实验结果的影响；TMPT6 中去掉了 $R_{-1}R_0$ 约束，通过 TMPT5 和 TMPT6 的实验比较，可以探讨前字的角色标注对当前字标注结果的决定性作用。

2.4　特征函数权重训练及人名实体提取

特征模板是特征函数的抽象形式，需要从观察序列组和角色标注序列中生成特征函数。本文使用 CRF + + -0.49 工具包[11]来实现特征模板到特征函数的转化。CRF + + 提供了两个工具：①crf_learn 用于基于特征模板从观察标注序列组中生成 CRFs 特征函数，并对其权重参数进行训练，建立 CRFs 模型；②crf_test 使用建立的 CRFs 模型对输入的仅有观察序列组的测试语料进行角色标注，为每一观察值组选择最佳角色。基于特征模板可以从训练语料中抽取出成千上万个候选特征函数，可从中选取一定数量的高频或高效特征函数用于人名角色标注。本文选取的特征函数出现次数（即特征函数频次阈值，记为 f）在 10 次以上。

经过 CRF + + 的角色标注，可以获得测试语料中每一个“字”对象的最佳角色。考察表 1 所列的人名角色，结合中文文本中人名的一般构成规律，笔者认为 BB、BCD、BE、BG、BBCD、BBE、BBG、XXCD、XXE、XXG、CD、E、FB、FXX、FBB、W…W 等角色组合模式可以构成人名实体。图 2 显示了基于角色组合模式实现人名实体提取的算法过程。

3　CRFs_RL_PnR 模型的实验分析

文本采用实验方式证明 CRFs_RL_PnR 模型在人名实体识别中的有效性和实用性，笔者拟统计以下数据作为测评指标来评价模型的识别效果：

• 测试语料中的人名总数（$N_{totlenum}$）；• 准确率（P）= $N_{truenum}/N_{reconum} \times 100\%$；• 被识别的人名个数（$N_{reconum}$）；• 召回率（$R$）= $N_{truenum}/N_{totlenum} \times 100\%$；•

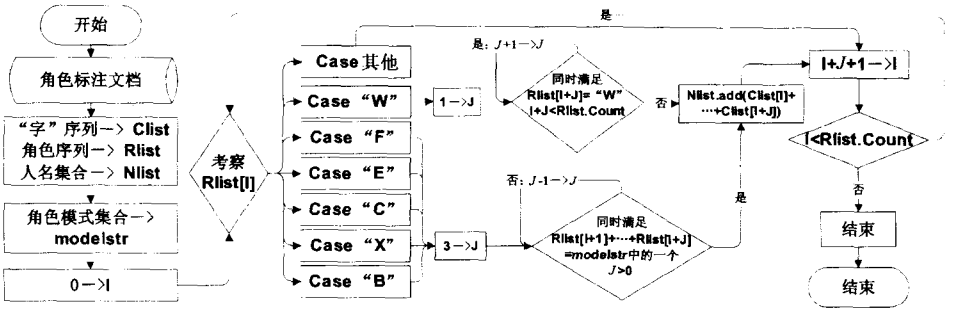
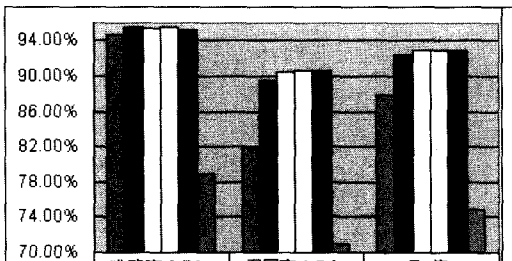
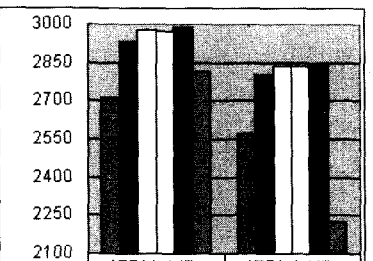


图2 基于角色组合模式从"字"序列和标注序列中抽取人名实体



	准确率(P)	召回率(R)	F_1 值
TMPT1	94.66%	82.06%	87.91%
TMPT2	95.57%	89.46%	92.41%
TMPT3	95.33%	90.55%	92.88%
TMPT4	95.46%	90.64%	92.99%
TMPT5	95.11%	90.68%	92.84%
TMPT6	79.01%	71.04%	74.81%



	识别人名数	识别正确数
TMPT1	2715	2570
TMPT2	2932	2802
TMPT3	2975	2836
TMPT4	2974	2839
TMPT5	2986	2840
TMPT6	2816	2225

图3 基于6组特征模板的开放性测试实验的测评指标结果比较

被正确识别的人名个数($N_{truenum}$)；• F_1 值 $=(\beta^2+1)\times P\times R/(\beta^2\times P+R)=2\times P\times R/(P+R)(\beta=1)$

此外，本文通过比较在不同特征模板作用下实验结果的变化情况以及不同概率统计模型对人名实体识别效果的影响来分析 CRFs_RL_PnR 的影响因素，以探索进一步改进模型的思路。

3.1 基于不同特征模板的实验分析

根据上述 CRFs_RL_PnR 模型的人名实体识别过程，笔者基于表2所列的6类特征模板分别进行了开放性测试实验，获得的测评指标结果如图3所示。其中发现：

(1)原始特征和角色定义的决定性作用。TMPT1仅由训练语料的原始特征"字"序列及其角色特征所组成，却使 F 值达到了87.91%，距离实验中最高 F_1 值(TMPT4)仅差5.08个百分点；TMPT1的 R 值相对较低，在加入6个衍生特征后，R 值明显提高；通过TMPT5和TMPT6的实验比较发现，在除去角色二元特征后 P、R 值明显变小，F_1 值下降了18个百分点强，表明角色二元特征对人名实体识别意义重大。

(2)合理观察序列的衍生可提高 F 值。从TMPT1到TMPT2、TMPT3，观察序列从1元增长到4元、8元，实验 P 值变化不大，F 值则由于 R 值的增长而明显变大。观察序列的衍生特别是高效特征的加入使得识别效果发生显著改善；TMPT3的变化幅度没有TMPT2大，表明在TMPT2中新加入的衍生特征较之TMPT3具有更显著的人名角色提示作用。

(3)观察序列取值变化可影响识别效果。TMPT3和TMPT4的差别在于多值词位特征W的取值不同，TMPT4采用6词位，比TMPT3的4词位具有更强的语义表达能力，实验结果也表明6词位的特征模板在人名识别的 P、R、F_1 以及 $N_{truenum}$ 值上都有一定的提高；然而由于中文人名大多在3字左右，4词位的表达能力即可实现准确的词位标记，因此词位表达能力的提高对人名识别的影响不大，主要表现在对多于3字的音译人名的识别上。

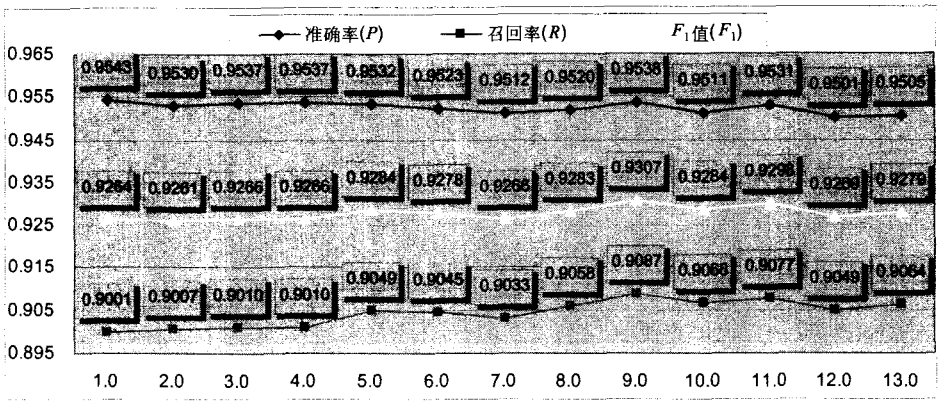


图 4　不同 c 值下的实验数据比较(TMPT5)

(4)字长窗口对人名识别效果影响不大。根据 TMPT3 和 TMPT5 的实验结果比较,发现在人名实体识别中增加字长窗口所附加的特征对人名识别没有起到有效作用,甚至在一定程度上阻碍了人名的召回:TMPT5 的 $N_{truenum}$ 略大于 TMPT3。

(5)软边界参数(soft margin parameter)c 对结果的影响。CRF++中的参数 c 用来调节 CRF 模型中的数据欠拟合(underfitting)和过拟合(overfitting)之间的平衡。图 4 为使用 TMPT5 在不同 c 值下的实验结果。随着 c 值的增大,P 值从总体上呈平缓下降趋势,而 R 则正好相反,在 $c=9.0$ 处似乎达到最高点,在 P、R 的共同作用下,F_1 略呈上升趋势,在 $c=9.0$ 处达到最高值。总体而言,c 值的变化对结果影响不大。本文在不做说明时,取 $c=10.0$。

(6)特征函数频次阈值 f 对结果的影响。以 TMPT1 为例,当选择的 f 值分别为 5、10、50、100 时,P 和 F_1 持续下降,而 R 则先增大后减小(如图 5 所示),说明随 f 值的增大,R 增大的速度始终低于 P 减小的速度;而 R 在增长达到一定值时回落。实验验证了:在强调 P 的应用中,使用低频特征函数可适当限制人名实体的召回以提高 P 值;反之,则可以使用较高频特征函数(一定范围内)提高 R 值。

(7)短句可提高人名识别的 F_1 值。在上述所有实验中不论训练还是测试均以语料库中一行作为一个句子,长度较长。笔者根据"。?!,;"等断句符对长句进行切分,使训练语料由原来的15 767句变为36 968句至115 318句。以 TMPT5 为例(实验结果如图 6 所示),缩短句子长度可以明显提高识别的 P 和 F_1 值,R 值也在一定范围内得到提升。

(8)考察根据 TMPT3 识别的人名集合,发现其中一些错误或没有识别的人名可以通过简单的规则

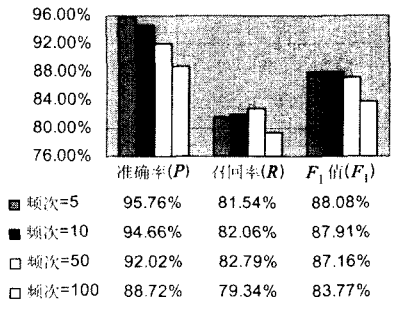


图 5　不同特征函数频次阈值下的实验数据比较(TMPT1)

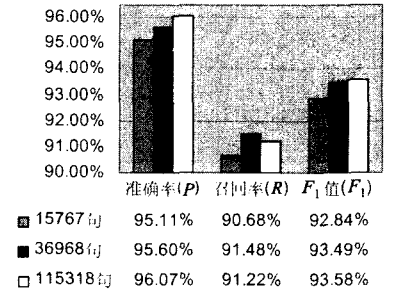


图 6　句子长度不同带来的实验结果差异(TMPT5)

加以排除或补充。①建立"人名停用字"集合,排除含有停用字符的错误人名;②收集"首尾停用字"集合,建立人名字符串首尾的用字规则;③对于之前一定范围内出现次数达到一定频率,而当前又无法识别的人名,可以通过建立"动态人名词表"加以补充识别等。在加入诸如此类的合理规则后,可以使 F_1 值可以达到更高的水平。

通过上述实证分析,可以得到如下结论:在获得足够多有效特征并尽可能将训练及测试语料切分成

合理短句(以不破坏语义为前提)的情况下,选择特定 f 值和 c 值(需要实验反复求证),通过 CRFs 字角色标注模型辅以一定规模的规则库,在训练语料充分情况下,可以使人名识别 F_1 值达到较高的水平。

3.2 CRFs 和 HMM 的实验比较

在本文引言中,笔者曾对 CRFs 和 HMM 两种概率统计模型进行比较,认为 CRFs 和 HMM 虽然都是根据观察序列特征进行序列标注,但依据的数学原理、建模的过程以及产生的结果都不相同。基于两种模型依赖的数学基础,笔者认为 CRFs 避免了 HMM 由于理论缺陷所必然存在的问题,较 HMM 更适合解决序列标注问题。为了验证理论探讨获得的结论,笔者对两种模型在人名识别中的应用进行了实验比较。基于 CRFs 和 HMM 的人名识别模型的开放性实验结果如图 7 所示。

(1)识别效果比较。CRFs 选择了表 2 中的 TMPT3 特征模板,f 值设定为 10,实验结果显示 CRFs 在识别准确率 P(95.33%)上明显优于 HMM(79.13%);两者召回率相差不多,但 CRFs 也比 HMM 高出 2 个多百分点;在 F_1 值比较中 CRFs 比 HMM 高出了近 10 个百分点,前者达到了 92.88%。HMM 仅在 $N_{reconum}$ 一项上比 CRFs 高,说明 HMM 比 CRFs 更善于发现人名,但这是以牺牲 P 值为代价。CRFs 能够融合多种特征的优势在人名识别效果比较中表现明显。

(2)实验复杂度比较。从图 7 中也可以明显发现 CRFs 的平均训练时间(23 小时)远大于 HMM(仅需 0.5 小时);而且在具体的特征学习过程中,CRFs 的实验环境要求也明显高于 HMM:CRFs 在使用 TMPT4 进行训练时,需要占用内存达 1G 以上。CRFs 的高实验复杂度的原因主要是其集成了众多的语言特征和标注序列全局范围内的归一化所造成的。基于更为复杂数学模型(CRFs)的人名识别系统虽然取得了理想的实验效果,但在效率方面明显低于 HMM。因此在有限的实验环境下,当识别效果要求不太高时,基于 HMM 的人名识别也是不错的选择。

(3)4 字以上人名识别比较。为了进一步分析 CRFs 和 HMM 的人名识别能力,笔者抽取了两次实验中 4 字以上(包括 4 字)人名进行比较,发现在 3132 个正确人名中共有 256 个人名为 4 字以上人名(基本上是音译外来人名),CRFs 共识别 251 个,其中 212 个为正确识别,P 值达到 84.46%,R 值为 82.81%,F_1 值为 83.63%,低于其平均水平近 10 个百分点(F_1 = 92.88%);而 HMM 则共识别 200 个人名,$N_{truenum}$ = 121,P 值仅为 60.5%,R 值仅为 47.27%,F_1 为 53.07%,低于其平均水平 30 个百分点强。实验表明:基于 HMM 的人名识别系统在对 4 字以上音译外来人名的识别中效果很不理想,这直接导致了其平均 F_1 值偏低;反之其在 4 字以下汉语人名识别中虽然仍逊色于 CRFs(F_1 值低于 CRFs),但具有一定的实用价值(R 值较为理想)。

4 CRFs_RL_PnR 模型在网络舆情分析中的应用

网络舆情[12]是指在一定社会空间内,围绕中介

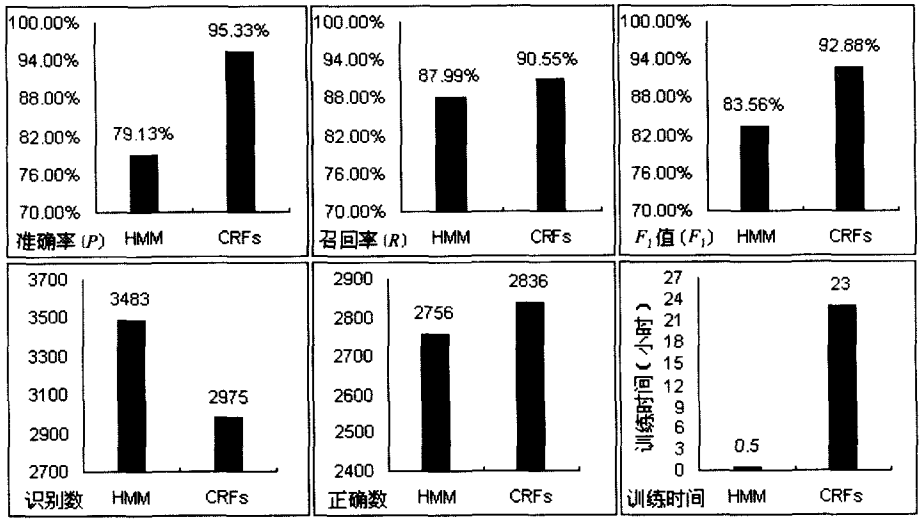


图 7 基于 HMM 和 CRFs 人名识别模型在开放性测试中的实验比较

性社会事件的发生、发展和变化，民众对社会管理者产生和持有的社会政治态度，是较多群众关于社会中各种现象、问题所表达的信念、态度、意见和情绪等表现的总和，并通过网络进行传递交流。网络舆情分析则是对网络中反映社会舆情的新闻、报道、评论等信息进行深入挖掘，从而为用户决策提供服务。而新闻、报道以及评论等信息中经常涉及各种人物，或作为事件主体，或作为评论对象。本文拟从新闻人物抽取及时序分析、基于人名的信息采集和全文检索等信息服务中探讨 CRFs_RL_PnR 模型的实际应用。

4.1 新闻人物的自动提取

通过对南京大学小百合 BBS 的信息采集，笔者获得其中百合论坛(Forum)和历史(History)两个版自创版以来到 2007 年 9 月的所有现存主题标题。采用 CRFs_RL_PnR 模型对主题标题进行人名自动抽取，获得这段时间内小百合 BBS 中最热门两个版块所涉及的新闻或焦点人物：在 Forum 版，讨论次数达到 5 次的有 49 人；在 History 版则有 61 人。

如图 8 所示，在这两个版面中，讨论次数都较多的人物有毛泽东、鲁迅和李慎之，或是我国开国领袖，或是著名文学家、学者，这从某个方面反映了当代大学生关心时政，热爱知识；从图中也可以发现两个版面探讨内容的差异：Forum 版多讨论当今社会的新闻人物，如孙志刚、马加爵等，而 History 版则多涉及历史著名人物，如李鸿章、孙中山、希特勒等。

4.2 焦点人物的时序变化分析

对上例 Forum 版中出现次数最多的五个人分别进行时序跟踪，以年为单位，使用 CRFs_RL_PnR 模型获得他们在每年(2002～2007 年)中出现的次数，结果如图 9 所示。

毛泽东、鲁迅和陈水扁是当代大学生一向比较关注的人物，因此在各年代中均有一定程度的出现；而孙志刚和马加爵作为某一时期的新闻人物，则集中出现在某一年代：孙志刚事件发生在 2003 年，在 2004 年及 2005 年也有零星关注，至 2006 年、2007 年就没有再涉及，而轰动一时的马加爵事件发生在 2004 年；关于陈水扁，由于其在 2007 年多次发表“台独宣言”，一时间也成为了国内大学生普遍关注的焦点人物；此外，如安倍晋三、福田康夫等人，均在 2007 年才崭露头角，引起人们关注，之前都没有涉及他们的讨论主题。

4.3 在网络舆情分析中的其他应用

(1)限制信息采集的内容。信息采集是网络舆情分析的三大内容之一，网络信息量的巨大使得盲目采集必然带来大量无关信息的堆积，有目的的采集是目前的研究热点之一。以当前互联网中广泛应用的 RSS(Really Simple Syndication)为例，仅以 RSS 种子(RSSSeed)作为采集标准实现内容聚合显然具有一定盲目性，进行内容限制是较好的方法，即采集在 RSS 提要中含有指定关键词的网页。在这种情况下，就需要对用户输入字符串进行主题词提取，如用户采集含有“南京大学冯端”内容的信息，首先需要提取主题词“南京大学”和“冯端”，而其中人名实体的提取在一定程度上决定了主题词自动提取的正确性，也决定了所采集信息的可用性。

(2)作为全文检索服务的分词基础。全文检索是网络舆情分析提供的信息服务之一。全文检索服务的关键在于全文索引创建和检索表达式切分，而中文分词是两者的基础。人名实体的抽取是中文分词的难点，同时在网络舆情中存在着大量的人物作为事件主体或评论对象，这就决定了在网络舆情分析系统的全文检索服务中，中文人名的识别是一项重要而有意义的工作。本文提出的 CRFs_RL_PnR

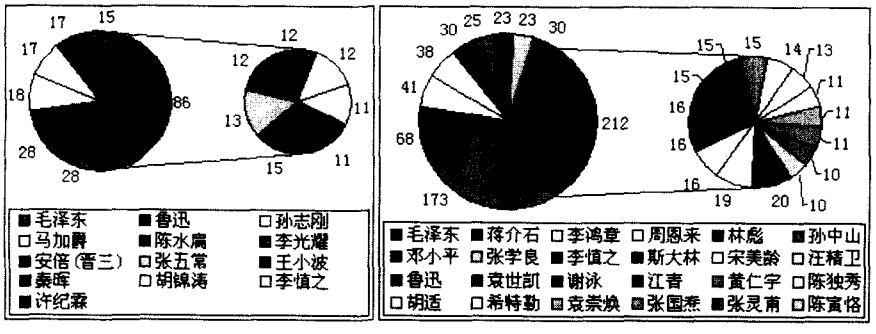
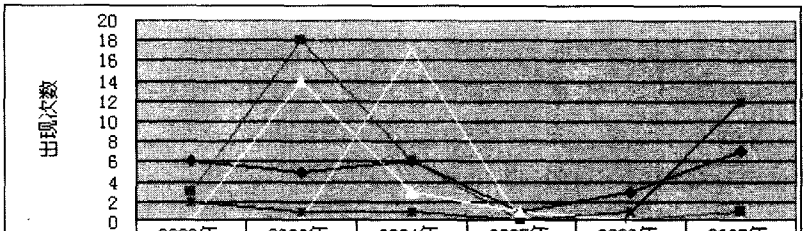


图 8 小百合 BBS Forum(左)和 History(右)版中讨论达到 10 次的新闻人物比例



	2002年	2003年	2004年	2005年	2006年	2007年
毛泽东	6	5	6	1	3	7
鲁迅	3	18	6	0	0	1
孙志刚	0	14	3	1	0	0
马加爵	0	0	17	0	0	0
陈水扁	2	1	1	0	1	12

图9 Forum版出现次数最多五个人的时序变化分析

模型为上述应用提供了研究的基础和方向。

5 结束语

本文对中文文本中人名实体的自动化抽取方法及其在网络舆情分析中的具体应用进行了系统研究，认为CRFs_RL_PnR模型能够对人名这一专名实体实现有效识别，F值可以达到93%以上，具有一定的实用价值。命名实体是中文文本中常见的概念，尤其是在具体应用领域，例如竞争情报系统研制中涉及的公司名，产品名等，各类命名实体的识别对于剖析文本语义，挖掘隐含信息具有重要作用。进一步提高人名实体的识别率（如在CRFs_RL_PnR模型中加入人名识别规则），促进基于CRFs的其他类型命名实体（如地名、机构名、企业名等）的正确识别是今后命名实体识别研究的重点。此外，命名实体识别在实践领域的具体应用，如网络舆情中除人名外其他命名实体的识别、竞争情报系统中命名实体抽取以及情报监测中的信息抽取[13]等，以及其作为其他领域，如本体概念抽取、文本语义分析、本体自动化构建等的研究基础所产生的影响和发挥的作用等都有待于进一步研究和探讨。

参 考 文 献

[1] 王昊. 基于层次模式匹配的命名实体识别模型[J]. 现代图书情报技术，2007(5)：62-68.

[2] Zhou G D, Su J. Named Entity Recognition using an HMM-based Chunk Tagger[C]// Proc. of the 40th Annual Meeting of the ACL, Philadelphia, 2002: 473-480.

[3] Bender O, Och F J, Ney H. Maximum Entropy Models for Named Entity Recognition[C]// Proc. of the Conference on Computational Natural Language Learning. Edmonton, Canada, 2003: 148-151.

[4] 钱晶，张杰，张涛. 基于最大熵的汉语人名地名识别方法研究[J]. 小型微型计算机系统，2006，27(9)：1761-1765.

[5] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]// Proc. of the International Joint Workshop on Natural Language Processing in Biomedicine and its Application (NLPBA). Geneva, Switzerland, 2004: 107-110.

[6] 向晓雯. 基于条件随机场的中文命名实体识别[D]. 厦门大学，2006.

[7] 姜维，关毅，王晓龙. 基于条件随机域的词性标注模型[J]. 计算机工程与应用，2006，42(21)：13-16，42.

[8] 黄昌宁，赵海. 由字构词——中文分词新方法[C]. 中国中文信息学会第六次全国会员代表大会暨成立二十五周年学术会议报告，2006.

[9] 赵健，王晓龙，关毅，等. 中文名实体识别：基于词触发对的条件随机域方法[J]. 高技术通讯，2006，16(8)：795-801.

[10] 郭家清，蔡东风，王智超，等. 一种基于条件随机场的人名识别方法[J]. 通讯与计算机，2007，4(27)：22-25.

[11] CRFs + + 0.49[OL]. [2007-03-06]. http://sourceforge.net/projects/crfpp/.

[12] 北大方正技术研究院. 以科技手段辅助网络舆情突发事件的监测分析——方正智思舆情辅助决策支持系统[J]. 信息化建设，2005(10)：50-52.

[13] 刘剑兰，朱东华. 信息抽取技术在情报监测中的应用[J]. 情报学报，2004，23(6)：661-666.

（责任编辑 芮国章）