基于规则与统计的维吾尔族人名识别研究*

赛依旦·阿不力米提，吐尔根·依布拉音
新疆大学 信息科学与工程学院，乌鲁木齐 830046
E-mail: sayida723@gmail.com

摘　要： 本文提出了一种基于规则与统计相结合的维吾尔族人名识别算法。我们从语料中提取人名左右边界词语，人名边界频度作为特征。识别过程是首先利用维吾尔族人名的后缀特点进行基于词典查找，然后应用带有频度的边界模型识别出可能的人名，并用几条排除规则对识别结果进行边界校正。系统采用真实语料进行测试的结果表明，正确率为88%，召回率为90%。

关键词： 维吾尔族人名识别；人名词典；边界模型；规则

On the Identifying System for Uyghur Names Based on Statistic Analysis and Rules

Sayida · Ablimiti, Turgun · Ibrayim
School of Information Science & Engineering, Xinjiang University, Urumqi 830046
E-mail: sayida723@gmail.com

Abstract: This paper proposes a recognition method of Uyghur name based on the statistics and regular. Person name's left and right boundary words and person name's character frequency are extracted from corpus, which will be used as character. First we use of Uighur name suffix features search based dictionary, Then application of the boundary model with frequency Identify the possible names and use a few heuristic rules to check and correct the name boundaries. System uses real corpus , the test results show that, The precision of the test is about 88%, and the rate o f recall is around 90%.

Keywords: identify Uyghur names; names dictionary; boundary model; rules

1　前言

少数民族自然语言处理目前也成了中文信息处理的一个及其活跃的研究领域，维吾尔族人名识别是维吾尔语自然语言处理中的一项重要工作，但目前要找到一个完善的维吾尔族人名识别系统还相当地困难。在机器翻译、信息提取、还是在自动文摘、信息检索等实际应用中，都需要进行词法分析，而人名识别是提高词法分析正确率的诸多有效方法之一。所以在词法分析阶段就利用上下文识别出人名即可消除歧义、提高分析精度。根据这个想法我们在本文中提出基于规则和统计相结合的维吾尔族人名识别方法。

2　维吾尔族人名识别的难点

维吾尔族人名数量多，规律性小、随意性大，维吾尔族人名识别的过程中我们遇到的困难，主要表现在以下两个方面。

（1）人名的构成。维吾尔族人名可以是单名，也可以是双名（双名构成为一人名）。如：mamat（买买提）和eli（艾力）各代表一个单名，而mamat eli是一个人名加父名。这样本属于一个人的名字被分成两个，会直接影响到维吾尔语人名识别的正确率。

（2）能当人名的词性很多。不仅有名词，还可以有动词、形容词等词性也充当人名。如：alim（阿里木，人名），alim （科学家），yalkhun（亚力坤，人名），yalkhun（火焰），surat（苏热特，

* 本文承国家社会科学基金重点项目[10AYY006]，国家自然科学基金项目[60663006]，国家工信部电子发展基金项目（批准号：工信部财（2009）453）的资助。

人名），surat（速度）这些特征对维吾尔族人名的识别带来歧义。

3 相关的研究

目前有关英文和中文人名识别问题的研究已经比较深入，主要方法可分为基于规则的和基于统计的方法[1]。基于规则的方法主要有根据人名的构成特征及上下文信息特征进行分析归纳，建立规则集[2]。该方法具有很高的准确率，但主要的信息依赖于现在系统的人名辞典，具有一定的局限性。基于统计的方法的特点是：建立统计模型[3]，对语料库进行统计训练，得到人名以及上下文信息的规律，设定阈值从而判断是否为人名，例如：隐马尔科夫模型（HMM）[4]、最大熵模型(ME)[5-6]。支持向量机模型（SVM）[7]，这些模拟在中英文人名识别中都得到了很好的应用。

虽然前人研究中文人名识别，并且已建立了一些识别方法，但这些方法不能直接用于维吾尔族人名识别任务中，原因在于维吾尔族人名的构成特点与中文名字完全不同，所以在人名识别中产生的歧义不同[8-9]，消除歧义的对象不同和对识别有贡献的特征存在较大的区别（在内容 1 中所介绍的维吾尔族人名识别的困难里举过例子）等。

针对以上问题，本文采用规则与统计相结合的混合策略，初步实现了一种维吾尔族人名的自动识别。该策略运用统计方法减少规则方法的复杂性，结合规则方法降低统计方法对语料库规模的要求。

4 基于规则和统计的识别算法

维吾尔语语料中经常遇到人名，而人名周围的信息有一定的规律性，如：

“يوللىغۇچى : ئابباس ياسىن”，“شىختيارنى مۇخبىرىمىز كامال جامال خەۋىرى”

以上例子中黑色部分是人名，而前后的红色部分是经常出现在人名周边的词以及形式。这特征启发我们建立识别人名的规则。

4.1 规则中使用的术语定义

定义 1（**边界词语**） 对维语句子中抽取的任一单词序列 $w_{-1}w_0w_1$，其中 w_0 为人名，w_1 称为人名右边界，w_{-1} 为人名左边界，w_{-1} 和 w_1 均可为空。当 w_1 为空时，记 w_0= Start，表示句子的开始；w_{-1} 为空时，记 w_0= End，表示句子的结束。

定义 2（**边界模板**） 称四元组 $q = (w_{-1}, FL(w_{-1}), w_1, FR(w_1))$为人名边界模板，其中 w_{-1}、w_1 分别为人名的左、右边界词语，$FL(w_{-1})$为在训练语料库中 w_{-1} 作为人名左边界的频度，$FR(w_1)$为在训练语料库中 w_1 作为人名右边界的频度[10]。定义边界模板 q 的频度 F(q)为：

$$F(q)=\begin{cases}0 & FL(w_{-1})=FR(w_1)=0\\ FL(w_{-1})+FR(w_1) & otherwise\end{cases}$$

定义 3（**扩散操作**） 在识别一篇维吾尔语文本时，如果某位置上的字符串 w_i 被识别为人名，并且人名词典中有收录，则把篇章中所有的字符串 w_i 都识别为人名。

4.2 特征提取及其知识库的构建

我们根据维吾尔族人名特征提取了人名后缀（共含 68 种）其中抽取了出现频率最高的 9 个后缀来建立一个规则表。9个后缀分别为：

表 1 后缀频度示例

词缀	نىڭ	نى	غا	لار	گە	ەر	دىن	لاردىن	ەردىن
次数	497	388	246	195	134	95	88	72	60

另外，我们收集了语料中人名左右边界词：

例如，对下述出现的人名“مەھبۇبە”：

شىنجاڭ تىببىي ئۇنىۋېرسىتېتىنىڭ ئوقۇغۇچىسى مەھبۇبە مۇنداق دېدى.

可从中提取左边界“مۇنداق دېدى”及右边界“ئوقۇغۇچىسى”。

照此，我们一共提取了698个左边界，723个右边界。出现频度最高的前10个左、右边界词语分别是：

表2　左边界频度示例

编号	左边界	频度
1	مۇنداق دېدى	352
2	خەۋىرى	311
3	مۇنۇلارنى بىلدۈردى	298
4	فوتوسى	223
5	قاتارلىقلار	198
6	دەپ تەكىتلىدى	172
7	تەرجىمىسى	166
8	قاتناشتى	149
9	ئەپەندى	120
10	خانىم	114

表3　右边界频度示例

编号	右边界	频度
1	مۇخبىرىمىز	431
2	يازغۇچى	383
3	رەھبىرى	368
4	ئۇجىسى	364
5	رەئىسى	326
6	پىراكتىكانت	318
7	شائىر	301
8	پروفېسسور	276
9	ئوقۇتقۇچىسى	158
10	ئوقۇغۇچىسى	143

5　识别方法的具体步骤

5.1　维吾尔族人名识别流程图

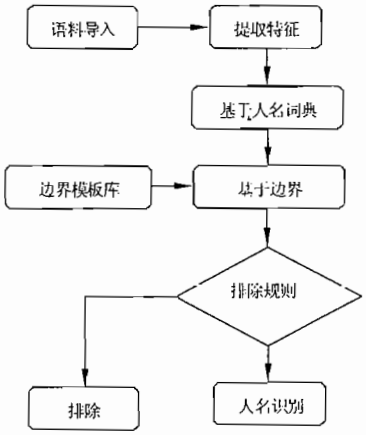

图1　识别流程图

5.2　人名识别方法的主要算法

[1] 篇章预处理：以句子为单位，把句子切分词。

[2] 第一次扫描：以句子为单位，逐词扫描。扫描时所遇到的词基于“维吾尔族人名词典”来识别。第一次扫描中识别出来的人名用##人名##来表示。

[3] 第二次扫描：用边界模板再次扫描，目的是通过边界模板方式找出没被人名辞典发现的人名。（扫描时跳过已经识别的人名）

[4] 第三次扫描：根据统计量对识别边界进行校正。

[5] 第四次扫描：应用排除规则对错误的识别进行排除操作。

使用的排除规则为：

规则 1 如果位置 i 识别为人名，但 i 的左右边界词语没有出现在边界规则库里，则排除位置 i 的人名。例如：

ئالىم بولساڭ ئالەم سىنىڭكى

这个说明系统把“ئالىم”这个词根据词典识别为人名可根据边界模板排除这个词。

规则 2 如果右边界词表示单数并且左边界的位置减去右边界的位置大于 2，则排除大于 2 的部分。例如：

كەنىت باشلىقى مامۇت تۇردى تەرىسلەنگەن ھالدا مۇنداق دېدى

这个说明系统默认把“مامۇت تۇردى تەرىسلەنگەن ھالدا”识别为人名，可“مامۇت تۇردى”是人名，因此按规则 2 排除其他部分。

6 实验结果与分析

我们从天山网、新疆作家协会等网站随即抽出了 1542 篇文章的数据，分别进行了两次测试，结果如表 4 所示。

其中正确率 P 的计算方法是：

$$P=\frac{\text{正确识别的人名个数}}{\text{识别出的所有人名个数}}\times 100$$

召回率 R 的计算方法是：

$$R=\frac{\text{正确识别的人名个数}}{\text{测试语料中人名总个数}}\times 100$$

综合指标 F 值的计算方法是：

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

表 4 测试结果

指标 / 语料	识别人名	正确识别	正确率	召回率	F 值
测试语料 1	441	387	78%	87%	0.82
测试语料 2	652	595	88%	90%	0.89

测试语料 1 以篇章为单位，因此包含的人名较少，产生干扰的人名比例较大，因而准确率较低。语料 2 以句子为单位（从 1542 篇文本中提取包含人名的 400 个句子，包含 686 个人名）因此精确率较高。

7 结束语

本文主要介绍在基于规则与统计的维吾尔族人名识别方面所做的初步研究，其中提出了以规则库方法作为基本框条同时结合基于词典与边界模板的优化处理实现维吾尔族人名的自动识别。目前的实验结果比较令人满意。

该方法表明在统计学方法的基础上适当用一些语言特征有关的规则将统计与规则相结合解决了维吾尔族人名识别所遇到的一些困难。

参 考 文 献

[1] 季恒，罗振声．基于统计与规则的中文姓名自动辨识[J]．语言文字应用，2001, 31(1): 14-18.

[2] 窦嵘，加羊吉，黄伟，等. 统计与规则相结合的藏文人名自动识别研究[J]. 长春工程学院学报，2010, 11(2): 113-115.

[3] 黄德根，杨元生，王省等．基于统计方法的中文姓名识别．中文信息学报．2001, 15(2): 31-37.

[4] Bikel, D M. Bikel, Schwartz, et a1. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1-3): 211-231.

[5] Borhwick, A. A maximum entropy approach to named entity recognition[J]. Ph. D. Thesis, New York University. 1999(5): 9-10.

[6] 贾宁，张全．基于最大熵模型的中文姓名识别[J]．计算机工程，2007, 33(9): 31-33.

[7] 李丽双，黄德根，毛婷婷，等．基于支持向量机的中国人名的自动识别[J]．计算机工程, 2006, 32(19): 188-190.

[8] 张秀玲．汉维人名文化异同之比较[J]．新疆大学学报，2009, 7(6): 136-139.

[9] 杜绍源．新疆维吾尔族人名初探[J]．中央民族大学学报，1983, 3: 68-78.

[10] 李中国，刘颖．边界模板和局部统计相结合的中国人名识别[J]．中文信息报, 2006, 20(5): 44-50.