基于词性探测的中文姓名识别算法*)

王源媛　何中市
(重庆大学计算机学院　重庆400044)

摘　要　本文提出了一种新的基于统计和规则相结合的中文姓名识别方法，即词性探测算法。该方法的特点是在对文本进行分词和词性标注一体化处理的基础上，通过探测候选中文姓名后的词性和比较单字的相对成词能力，能够对分词碎片中的姓名进行有效识别。

关键词　中文姓名，识别，词性探测，统计语言模型

Algorithm for Chinese Person Names Recognition Based on Part-of-Speech Detecting

WANG Yuan-Yuan　HE Zhong-Shi
(College of Computer Science, Chongqing University, Chongqing400030)

Abstract　This paper presents a new approach to identify Chinese person names based on a statistical model integrating Chinese word segmentation with part-of-speech tagging. The algorithm features in boundary affirmation of person names by detecting candidate characters' part-of-speech and in comparing the relative probability of would-be-word character.

Keywords　Chinese person names, Recognition, POS detecting, Statistical language modeling

1　引言

词是自然语言中有意义的、可以独立运用的最小单位。在诸多现实领域(文本的自动检索、输入法、机器翻译等)，中文自动分词都扮演着重要的角色。我国关于自动分词的研究，早在上个世纪70年代就受到广泛关注，并取得了较好的成果。但长期以来困扰中文自动分词发展的两大难题，就是歧义切分和未登录词的识别。目前，歧义切分问题已经得到了很好的解决[1,2]，而未登录词的识别则成为中文自动分词领域研究的瓶颈。

未登录词是指分词系统的词典中未收录因而机器不能认识的词，如中文姓名、地名、机构名、品牌名等。未登录词种类繁多，且本质上不可穷举，因为新词随着时间的推移还在不断地增加。但是，一个中文信息处理系统如果不具备处理未登录词的能力，它就不能自动处理大规模的语料，而且还会造成分词错误，直接影响到中文自动分词及整个句法分析的正确性[3]。在现在正蓬勃发展的信息检索和数据挖掘研究领域中，未登录词的识别有着更为重要的意义：识别未登录词可以提高分词和抽词的精度，扩充依据词典提取的关键词集合，从而更加准确地对信息资源进行描述、分析与理解。

本文提出的词性探测算法就是一种对未登录词中的中文姓名进行识别的方法。最后，通过实验证明该算法是可行的，并取得了有效的测试结果。

2　中文姓名识别的研究与现状

根据未登录词的性质，可以将目前关于未登录词的研究分成两个大类—非专名和专名。对非专名的研究包括简称、方言、行业用词等，由于非专名具有范围广泛，难以运用规则进行约束等特点，对非专名的识别较后者要困难一些。基本的识别方法有：有穷多层次列举法[4]，窗口移动扩展法[5]等；专名研究的任务主要是对文本中出现的人名、地名、机构名进行识别。基本方法有：语料库统计、局部统计、结合词性标注[6]等。

中文姓名识别属于专名识别的范畴。由于中文姓名本身具有的特性，如：词长趋于稳定，词首用字分布集中等，到目前为止已经出现了很多识别中文姓名的方法，但最终都可以归为以下三类：基于统计的[7]、基于规则的和两者相结合的方法[5,6,8]。

基于统计的方法一般是在对输入文本进行分词的基础上，寻找可能构成中文姓名的字串，计算其组合概率，并应用一定的筛选公式来识别中文姓名。基于规则的方法是根据语言学知识，建立中文姓名的构词规则、上下文特征词库等来辅助识别中文姓名。上述两种方法很少被孤立使用，一般都是将统计与规则结合起来识别未登录中文姓名，即混合法。

常用的中文姓名识别指标有两个[3]：

$$\text{查准率}=\frac{\text{正确确识别出的姓名}}{\text{系统统判定为姓名总数}}\times 100\%;$$

$$\text{查全率}=\frac{\text{正确确识别出的姓名}}{\text{语料中的姓名总数}}\times 100\%;$$

现有的中文姓名识别方法普遍存在查全率与查准率难以同时保证较高的问题。通常是以牺牲部分查全率来换取较好的查准率。本文提出的词性探测法在保证较高查全率的同时，尽可能地结合统计、规则、词性等各种手段，最大限度提高中文姓名识别的查准率。

3　词性探测算法

词性探测算法，顾名思义是建立在对文本进行词性标注基础之上的。在整个算法实施过程中，我们用到了单字词的一元模型、分词和词性标注一体化模型。模型所需的全部参数由《人民日报》1998年1月份标注语料库(该语料库从北京大学计

*)基金项目：国家自然科学基金项目(60173060)。王源媛　硕士研究生，主要研究方向为自然语言处理。何中市　教授，博导，主要从事自然语言处理、数据挖掘技术和计算机网络可靠性等方面的研究。

算机语言学研究所主页上免费下载)训练获得。

3.1 统计语言模型

3.1.1 单字词的一元模型 这里用到的统计语言模型是 N 元模型 (N-gram)[10,12],即是用变量 W 代表一个文本中顺序排列的 n 个词:$W=w_1w_2\cdots w_n$,并假设任意一个词 w_i 的出现概率只与它前面的 $N-1$ 个词有关:

$$P(w_i|w_1w_2\cdots w_{i-1}) = P(w_i|w_{i-N+1}\cdots w_{i-1})$$

因此:

$$P(W)=P(w_1w_2\cdots w_n)=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_i|w_{i-N+1}\cdots w_{i-1})$$

$$=\prod_{i=1}^{n}P(w_i|w_{i-N+1}\cdots w_{i-1})$$

分别称 N 取1、2时相应的语言模型为一元模型和二元模型。

由上述 N 元模型的定义可知,一元模型假设任意一个词 w_i 的出现概率与其前面的词无关,即 $P(w_i|w_1w_2\cdots w_{i-1})=P(w_i)$。

在汉语中,一个单字可能成词(指该单字构成一个词)、也可能不成词。也就是说,单字成词能力有强弱之分。我们用单字相对成词概率来表示一个单字成为词的能力的大小,并用 $P_{CW}(c)$ 表示,其值可以通过训练语料计算:

$$P_{CW}(c)=\frac{\text{单字 } c \text{ 作为单字词出现的次数}}{\text{单字 } c \text{ 出现的总次数}}\times 100\%$$

3.1.2 基于统计的分词和词性标注一体化模型 词性探测算法,顾名思义是建立在对文本进行词性标注基础之上的。这里我们引入分词和词性标注一体化分析的思想[9]。过去进行词法分析时,人们通常将分词、词性标注分开单独处理。实际上,二者有着密切联系。比如,利用语法知识可以消解90%以上的分词歧义。因此,将分词过程和词性标注过程融为一体将有利于消解歧义,同时减少了系统开销。

中文分词和词性标注一体化问题可描述为:输入含有 n 个汉字的待处理字串 $C=c_1c_2\cdots c_n$,在所有可能的切分序列 W 和相应词性标记序列 T 中,选取具有最大可能(最高评分)的切分词串 $W^*=w_1w_2\cdots w_m$ 和相应词性标记串 $T^*=t_1t_2\cdots t_m$ 作为输出结果。因此,分词和词性标注一体化处理的一般模型为:

$$W^*T^*=\arg\max_{W,T}P(W,T|C) \quad (1)$$

文[9]提出了对分词和词性标注一体化处理的概率评分模型。该模型的任务是在所有候选二元组 (W,T) 中,寻找一个二元组 (W^*,T^*),使得 Score(W^*,T^*) 最大。在一般模型基础上,概率评分模型引入了词形的概念:将输入汉字串 $c_1c_2\cdots c_n$ 中任意一个字 c_i 看作待切分点,以该字为起点且能够在系统词典里找到匹配项的所有词构成 c_i 的词形 β_i,用 F 表示 $\bigcup_{j=1}^{n}\beta_j$。显然,候选切分句 $W=w_1w_2\cdots w_m$ 是由从 c_1 到 c_n 间首尾相连的词形串构成的,与 W 对应的词形串我们记为 $f_1f_2\cdots f_m$。这样,F 成为词串 W 的条件概率 $P(W|F)$,就在一定程度上,从统计学的角度反映了汉字构词规律,为分词和词性标注一体化处理提供了有利信息。因此,概率评分模型可表示如下:

$$Score(W,T)=P(T,W,F|C)\cong P(T)P(W|T)P(W|F)P(F|C) \quad (2)$$

现引入三个假设:

(1)词类二元模型:任意词类标记 t_i 的出现概率只同它紧邻的前一个词类标记 t_{i-1} 有关。有:

$$P(T)=P(t_1t_2\cdots t_n)=\prod_{i=1}^{n}P(t_i|t_{i-1})$$

其中,$P(t_i|t_{i-1})$ 是词类标记的转移概率,也叫做词性标注的二元模型,其计算方式如下:

$$P(t_i|t_{i-1})=\frac{\text{训练语料中 } t_i \text{ 出现在 } t_{i-1} \text{ 之后的次数}}{\text{训练语料中 } t_{i-1} \text{ 出现的总次数}}\times 100\%$$

(2)独立性假设:w_i 分别依赖于 t_i,f_i;

(3)假设对一定的汉字串,$P(F|C)=P(F)$,即 f_i 仅依赖于自身。于是公式(2)可简化为:

$$Score(W,T)=\prod_{i=1}^{m}P(t_i|t_{i-1})P(w_i|t_i)P(w_i|f_i)P(f_i) \quad (3)$$

这就是分词和词性标注一体化处理的概率评分模型。

3.2 词性探测算法原理

中文姓名一般分为"姓+单名"、"姓+双名"两种,这里我们不妨把单名或双名简称为名。通过对人民日报1998年1月份的标注语料库(含44个词类标记)进行统计分析,我们发现在该语料库的15177个中文姓名中,"名"由连词(conjunction)、介词(preposition)、助词(auxiliary word)的单字或连续单字形式(以下简称这三类词性为"cpu")构成的概率很小,仅为0.0132,而名词(如"雪、玲、涛")、形容词(如"红、亮、富")、动词(如"飞、行")等其他各类词性的词出现频繁。为了进一步验证这一思想,我们收集了重庆市某医院2002年9月到2003年10月,共计163520个病人姓名进行统计。发现"名"由词性为"cpu"的单字或连续单字形式构成的概率为0.0276。基于以上事实,我们可以在中文姓名的识别过程中,考虑将候选姓名后词性为"cpu"的词作为姓名识别的边界。

具体作法是,在切分且已标注词性的文本中,以姓氏为起点,依次探测该姓氏后第三个、第四个单字的词性。一旦发现该单字的词性为"cpu",探测立刻停止。将探测终止前的单字合并为名,连同姓氏一起作为识别出的中文姓名。经过上述初步探测,可以将部分人名识别出来。而姓氏后第三个、第四个单字均不为"cpu"的情况我们将借助单字相对成词能力来辅助识别,将成词能力较强的作为姓名的边界词。例如:

陈/nr 雨/v 飞/v 说/v 的/u 故事/n

我/r 听/v 过/v 。/w

显然"飞"和"说"词性均不为"cpu",比较二者的单字成词概率:"飞"为0.16637,"说"为0.77981,前者远远小于后者。因此,我们将"雨"、"飞"合并,成功识别出了"陈雨飞"这一人名,而不是将"陈雨"错误地识别出来。(注:"/"为分词系统处理后的切分标志,"/"后的符号是词性标注结果,标注规则和符号集含义参见文[11])。

3.3 词性探测算法要点

(1)对输入文本进行分词和词性标注一体化预处理。我们采用分词和词性标注一体化方法对输入文本进行预处理。实践证明,一体化处理后,可以较好地消除歧义,为后面进行中文姓名的识别提供良好的上下文环境。

(2)提取分词碎片。分词碎片是指经过预处理后,文本中包含的一个或多个连续单字[6],例如:

A:阿/m 毛/q 是/v 个/q 好/a 孩子/n

B:张/q 清/t 楚楚动人/v 地/u 站/v 在/p 我们/r 的/u 面前/f

A 句有分词碎片"阿/m 毛/q 是/v 个/q 好/a ";B 句有分词碎片"张/q 清/t"、"地/u 站/v 在/p"等。且 A、B 各含一个未登录人名"阿毛"、"张清"。我们把候选中文姓名锁定在分词碎片之内,是因为大多数人名在经过预处理后都被切分成了连续单字的形式。

(3)定义规则并加以应用。根据中文姓名的构词特征,我们定义了几条规则对分词碎片进行过滤,在此过程中可以提

取小部分人名。

Ⅰ.切分标志词规则:收集一定数量的单字切分词,如"的"、"是"、"吧""啊"等,和常用标点符号集一起构成单字切分词词表,将与之匹配的单字、标点从分词碎片中抽取出来。

Ⅱ.前缀词规则:收集经常出现在姓氏前的单字词构成前缀词词表,如"老"、"小"、"阿"、"大"等。如果发现这类词后的单字在常用姓氏表中,则立刻将"前缀词+单字"作为识别出的一个人名,从分词碎片中提取出来。

Ⅲ.后缀词规则:单字后缀词有"总"、"老"、"伯"、"叔"等。如果发现这类词前的单字出现在常用姓氏表中,则立刻将"单字+后缀词"作为识别出的一个人名,从分词碎片中提取出来。

(4)常用中文姓氏表。我们搜集整理了670余个中文常用姓氏及其使用频度,构成常用姓氏表。在实验过程中我们发现,由于本文提出的算法是以姓氏为起点进行探测,很多使用频度非常低的姓氏,一般都具有一些特殊性(通常都不能单独成词,如"过"、"都"等),常常被当作是姓名的起点而被错误识别出来。在很大程度上影响了系统识别中文姓名的精确性。因此,我们设置一阈值λ(单位:次),在算法实施过程中不予考虑频度小于λ的姓氏。

3.4 词性探测算法描述

(1)对输入文本进行预处理;

(2)获取分词碎片,过滤切分标志词;

(3)应用规则提取一部分姓名;

(4)查找碎片中是否有与常用姓氏匹配的单字,有则转(5),否则转(7);

(5)探测该姓氏后第三、第四个单字的词性,为"cpu"时终止,否则转(6);

(6)比较该姓氏后第三、第四个单字的相对成词概率;

(7)分词碎片遍历是否完毕,是则算法结束,否则转(4)。

4 实验结果

我们在不同题材的测试语料库中抽取新闻、文艺作品各一篇,作为测试集。实验结果如下表所示。

题材	语料大小(kB)	λ的取值(次)	查全率(%)	查准率(%)
新闻	120	10	91.71	73.44
		31	89.12	86.12
文艺	6	10	93.42	83.87
		31	92.11	90.91

上述数据显示,当阈值λ取10时,用词性探测算法进行中文姓名识别获得了较高的查全率,但查准率偏低。这是由于词性探测算法以姓氏驱动所致。尽管我们在常用姓氏表中只保留了频度>10的姓氏,其中仍然有很多姓氏在真实文本中经常作为一般用字出现,而不是作为姓氏出现,干扰了中文姓名的正确识别。为了平衡查全率和查精率这一对矛盾,我们进一步缩小了收入常用姓氏表的姓氏范围,减少低频姓氏作为探测起点对正确识别中文姓名所造成的影响。经我们反复实验,当λ取31时,查全率虽然略有下降,但是查准率却得到了大幅度提升。

另外,我们从中文姓名识别的相关文献中,搜集整理了若干个用常用方法无法识别或识别错误的实例,用词性探测法进行测试,效果较为理想:

题材	语料大小(kB)	λ的取值(次)	查全率(%)	查准率(%)
难识别姓名集	1	10	85.71	66.67
		31	80.95	77.28

其中,识别正确的有(下划线表示系统识别出的中文姓名):"赵 明生于1960年","会员王 占为此苦苦思索","阿 杜唱的那首歌很好听,陈 阿毛小时候很调皮"等。用现有方法容易识别错误的,如"决赛今日于蓉城揭晓","三叶枫影影绰绰",用词性探测法进行识别收到了较好的效果,没有将"于蓉城"、"叶枫"当作姓名识别出来。

错误及没有识别出来的情况如"梁山 伯","高 海洋","刘 光亮","田 边熬","齐 伯"等。引起上述错误的原因主要有几个:1)本算法未对跨越分词碎片的中文姓名作出相应处理;2)由于一些常用姓氏同时又是普通用字,以该字为起点探测姓名时导致错误;3)词性探测法须借助于词性标注信息,必然在一定程度上受到词性标注结果的影响。

结论 本文提出了一种将统计与规则、词性巧妙结合起来的中文姓名识别算法,并同时获得了较高的查全率和查准率。特别是对于一些用传统方法难以识别或识别错误的情况,词性探测算法都可以将其避免。针对该算法的特点,我们今后将在以下几个方面继续展开研究并加以改进,力求在系统性能上有更大突破:

(1) 在分词和词性标注一体化过程中提高分词和词性标注的精度,这是提高词性探测算法查全率、查准率的关键。

(2) 搜集大量中文姓名样本,扩充现有常用中文姓氏表,增强各姓氏使用频度的权威性。

(3) 积累中文姓名的构词规则,充实系统的规则库,实践证明,定义适当的规则对于大幅度提高准确率很有帮助。

(4) 在保证较高查准率的同时,尝试对跨越分词碎片的中文姓名做适当处理。

(5) 将词性探测算法的思想应用到其它种类的未登录词识别系统中。

参 考 文 献

1 周昌乐.心脑计算举要.北京:清华大学出版社,2002.27~30

2 Li Liangyan,He Zhongshi,Yi Yong. Principles and algorithms of semantic analysis. In:2003 Int. Conf. on Machine Learning and Cybernetics (ICMLC03),Xi'an China,Nov.2003.1613~1618

3 吕雅鹏,赵铁军,杨沐昀,于浩,李生.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,2000,15(1):440~446

4 张普,张尧汉.现代汉语"有穷多层列举"自动分词方法的讨论.语言与计算机,1986(3)

5 聂颂,何丕廉,孙越恒.统计与规则结合的一种新词识别方法.微型机与应用,2003(10):58~60

6 陈小荷.自动分词中未登录词问题的一揽子解决方案.语言文字应用,1999(3):103~109

7 黄德根,杨元生,王省,张艳丽,钟万勰.基于统计方法的中文姓名识别.中文信息学报,2001,15(2):31~44

8 刘秉伟,黄萱菁,郭以昆,吴立德.基于统计方法的中文姓名识别.中文信息学报,1999,14(3):16~24

9 付国宏,王平,王晓龙.汉语分词和词性标注一体化分析的方法研究.计算机应用研究,2001(7):24~26

10 黄昌宁.统计语言模型能做什么?.语言文字应用,2002(1):77~84

11 俞士汶,段慧明,朱学锋,孙斌.北京大学现代汉语语料库基本加工规范.中文信息学报,2002,16(5):49~65

12 Rosenfeld R. Two Decades Of Statistical Language Modeling: Where Do We Go From Here?:[School of Computer Science Carnegie Mellon University Pittsburgh. PA 15213 USA]. http://www-2.cs.cmu.edu/~roni/papers