自动分词系统中姓氏人名处理策略探讨*

郑家恒　刘开瑛
山西大学计算机科学系

摘要：人名虽然在汉语文本语料中所占比例不大，但是对自动分词系统的准确率却有很大的影响。本文分析了语料中人名姓氏的情况，探讨了处理中国人名姓氏的策略和规则。

一、引言

在汉语语料中，虽然人名出现的比率不足百分之一，但在自动分词过程中，人名的出错率却不低。在对10万字新闻语料的测试中，总体准确率为99.28%，歧义准确率为78.03%，新词（包括人名、地名、机关名）为79.43%，得出分词软件切分正确率为96.54% [1]。因此研究对人名的切分是很有必要的。本文统计和分析了10万语料中中国人名、外国人名和姓氏出现的情况；研究了带有姓氏的400多个句子；提出了处理中国人名的策略。

二、人名姓氏统计信息的分析

我们选取了科普、科技、新闻、法规、政务信息各二万字，组成了十万字的实验语料库。对这十万语料进行了分词和人工标注词类。分词标准是按中华人民共和国国家标准的“信息处理用现代汉语分词规范”[2]，词类分类标准是按“汉语语料库语法标记”[3]。然后，对加工后的语料就有关人名的信息进行了统计。

1. 由于实验语料仅有10万字，人名出现率不超过百分之一，因此把出现过1次的姓氏都统计在内，10万语料的总词次数64660，其中人名（中国人名和外国人名）有390次，占总词次数的0.6%。总词条数为8050，其中人名的词条数为72条，占总词条数的0.9%。姓氏共出现了56个。

2. 由于语料类型的不同，人名在各类语料出现的次数有很大的区别。在390个词次中，新闻占49%，科技占23%，科普占8%，法规占0%，政务信息占30%。在56个姓氏中，新闻类31个，科技类14个，科普类3个，法规类0个，政务信息类27个。由此可见，人名多集中在新闻和政务信息的语料中，法规类没有人名的出现。这一现象确实符合语料的现实性。针对这种情况，我们又选取新闻、政务信息1万字，进行人名和姓氏的分析。在这1万语料中，人名和含有姓氏的词占4%，因此在对这两类语料进行分词时，更应该注意人名姓氏的处理。

某些兼类的姓氏，随语料类型的不同，兼类很有规律。例如“马”，共出现47次，作为名词的有34次，作为姓氏的有13次。作为名词的“马”，34次全集中在科普类语料中。原因是在选取科普类语料时，选了一篇专门介绍马的文章。而作为姓氏的“马”，却都集中在其它三类语料里。

3. 中国姓氏大约有400多个，在10万语料中按姓氏出现的有56个，其中非兼类的有44个，兼类的有12个。下面用到的词类标记取自参考文献[3]，其中：人名：NPF，姓氏：NPFF，名字：NPFS，动词：V，形容词：A，量词：Q，介词：P，名词：NG。

例1　“顾”兼动词

---

* 国家自然科学基金资助项目

记者　顾（NPFF）　筑胜

导致　只　顾（V）　短期　的　经济　利益　的　倾向。

例 2　“黄”兼形容词

黄（NPFF）　祥喜　探讨　了　情报　检索　的　模糊数学　描述。

如果　让　彩色　光带　射到　黄（A）　玻璃　上。

例 3　“周”兼量词

由　周（NPFF）　恩来　任　主任。

地球　每　24　小时　自转　一　周（Q）。

例 4　“曾”共出现了 21 次，其中作为姓氏的仅有 1 次，其余的都是副词。“都”共出现了 52 次，作为姓氏的仅有 1 次，其余的也都是副词。

例 5　“向”、“祝”、“管”，在语料中分别按介词、动词等词类出现的。

4. 语料中人名是以下述四种情况出现的。

a）姓和名字分开的中国人名

如：江（NPFF）泽民（NPFS）　祖（NPFF）冲之（NPFS）

b）姓和名不分开的中国人名

如：李鹏（NPF）　马林（NPF）

c）姓氏与老、小、总等组成的简称

如：老王（NPF）　小李（NPF）　张总（NPF）　李工（NPF）

d）外国人名

三、人名姓氏信息处理策略和规则

1. 策略

根据语料的统计结果和语言现象，我们采用了姓氏分类，姓氏加权和建立常用人名库等策略。

1.1 中国姓氏大约有 400 多个，10 万语料中出现过 62 个，其中作为姓氏出现的 56 个，不作为姓氏出现的有 6 个。我们把这 62 个姓氏分成作为姓氏、不作姓氏、兼类和待定四类。

可作姓氏的有：曾、陈、邓、丁、蒋、康、鲁、公孙等 48 个，其中 45 个在语料中都是按姓氏出现的。“蒋”、“康”、“鲁”这三个姓氏在语料中是按兼类出现的。“蒋”出现 11 次，10 次是姓氏，1 次是名词。名词是在“李宗仁率兵反蒋”这个句子中，被标注为名词的。“康”、“鲁”作为简称各出现 1 次，其余都是姓氏。鉴于这种情况，我们把这三个归入作姓氏一类。

不作姓氏的有：曾、都、向、管、祝、于。“曾”、“都”这两个姓氏在语料中有作姓氏的情况，但作副词的机率远远大于作名词的机率。因此把它们归入不作姓氏一类。“祝”、“向”、“管”、“于”都是 400 个姓氏中的一个，但在 10 万语料中均不作姓氏。

例如：向（P）　在　“夏打……”。

省里　把　权力　放　给　管（V）　工业……

取　信　于（P）　民。

祝（P）　同志们　身体　健康。

兼类的有：黄、马、张、周等。10 万语料的统计如下：

黄：NPFF 3 次，A 3 次，共 6 次。

马：NPFF 13 次，N 34 次，共 47 次。

张：NPFF 2次，V 3次，Q 2次，共7次。

周：NPFF 1次，Q 3次，共4次。

叶：NPFF 1次，N 1次，共2次。

顾：NPFF 3次，V 2次，共5次。

待定的有：任（NPFF，V），房（NPFF，N），方（NPFF，N），这几个姓氏都有兼类，由于语料的数量有限，还需再做进一步的探讨。

1．2 对某些兼类姓氏，可以根据待分词语料的类型，给个权值。例如：在处理科普语料中，若“马”字出现次数很多，就可以把马作为名词的权值加大。而在处理其它类型时，则把作为姓氏的权值加大。

1．3 经常用到的中国人名和外国人名可以建立一个常用人名库。

2．规则

针对以“小李”、“老王”、“张总”等出现的简称和部分兼类的姓氏，作了如下规则。在阐述规则之前，先将使用的符号作一介绍：

当前词：X，前一词：F，后一词：B1，后二词：B2。400个姓氏集为N1，可作姓氏集为N2，数字词集为M。

R1：设 K1＝｛小、老｝

若X∈N1，并且F∈K1，则FX为NPF。

例：小王（NPF），老李（NPF）。

R2：设K2＝｛工、总｝

若X∈N1，并且B1∈K2，则XB1为NPF。

例：张工（NPF）、张总（NPF）。

R3：设K3＝｛张、项、周、章｝

若F∈M，且X∈K3，则X为Q。

例：一 周（Q） 内，

按 第 七 章（Q） 规定。

R4：设K4＝｛多、各｝

若X∈｛方、项、章、段｝，F∈K4，则X不作姓氏，可按词类划分标准再标注X的词类。

例：多 方（N） 筹备…… 各 项（Q）准备 工作……

新的人名在分词过程中，往往被分成单个的字串，在一串单字的字串中正确地区分出人名，是自动分词的一个难点。经过对300多个句子的分析，我们提出了如下处理人名的规则。

R5：若X∈N2，且B1是二字词语，则B1为人名中的名字。

例：罗（NPFF）胜利（NPFS）

R6：若X∈N2，并且B1是一字词语，B2∈｛标点符号、的、了、是｝或B2∈｛二字以上的词语｝或B2的词类是动词，则XB1为NPF。

例：

字串	结果
瑞金医院 的 陈 柯、	瑞金医院 的 陈柯（NPF）、
主治医师 毛 羽 说：	主治医师 毛羽（NPF）说：
常委 乔 石今天上午	常委 乔石（NPF） 今天 上午

R7：若X∈N2，B1是一字词语，B2为一字词语，并且B2∉｛标点符号、的、了、是、

在〕或 B2 不是动词，则 X 为 NPFF，B1B2 为 NPFS。

例：	字串	结果
	记者顾筑胜）	记者　顾（NPFF）筑胜（NPFS））
	由吴俊洲副省长	由 吴（NPFF）俊洲（NPFS）副省长
	老工人李传杰，	老 工人 李（NPFF）传杰（NPFS）
	以贺广祥为首	以 贺（NPFF）广祥（NPFS）为 首
	熊子和率领	熊（NPFF）子和（NPFS）率领

四、实验结果

经过 400 多个句子的测试，准确率可达到 85%，错误主要集中在 1. 姓氏是兼类的情况。2. 当 B2 为动词时，按 R6 规则，XB1 应为人名，但有时应该是 B1B2 为名字，即分成 X B1B2。今后还应该将语料数量增多，覆盖面扩大，把 400 多个姓氏作更细的分类。另外还应增加处理兼类姓氏的规则，对名字的组成再做进一步的研究。

参考文献

[1] 郑家恒 、刘开瑛 、赵军，书面汉语自动分词软件评测方法研究，模式识别与人工智能，合肥 ，1992

[2] 信息处理用现代汉语分词规范，国家标准局，1989. 9

[3] 汉语语料库词类标记 ，山西大学应用研究所，1991

[4] 张俊威等，多语料库作法之中文姓名辨识，中文信息学报，1992. 3

APPROACH of PROCESSING TACTICS on the NAMES and SUNAMES in CHINESE AUTOMATIC SEQMENTING SYSTEM

Zheng Jiaheng Liu Kaiying
Dept. of computer science, Shanxi University
Zip Code 030006 Taiyan Shanxi

ABSTRACT

Names makes up few proportions in chinese coups, but they af—fect deeply the accuracy of automatic seqmenting system. This papers analyses the case of names and surnames in chinese cou—ps, and makes areseach on tacties and rules about processing c—hinese names and surnames.