基于统计和规则的人名识别方法*

吴雪军　朱靖波　陈学耀　卓红霞

（东北大学信息学院计算机软件研究所中文信息处理实验室　沈阳　110004）

E-mail {wuxj1977@163.com , zhujingbo@yahoo.com }

摘要　本文在对大规模语料统计的基础上，建立了一个包含人名姓氏及其统计概率、人名用字及其统计概率、人名前后缀和人名引导词等资源的人名识别的知识库，提出了一种统计和规则相结合的人名识别方法。本文把人名分为三类，一类是中文姓名，一类是不带姓氏的人名，还有一类是外国人名。利用我们构造的知识库，采用人名姓氏和人名用字组成概率与前后缀引导词规则相结合的方法分别对这三类人名进识别。该方法兼顾了准确率和召回率，获得了较好的识别效果。经封闭测试，召回率达到了 91.35%，准确率达到了 92.23%。

关键词　人名识别　姓氏使用概率　人名用字使用概率

作者简介：吴雪军（1977－）男，浙江龙游人，硕士研究生，研究方向：语言分析和信息检索；朱靖波（1973－），男，浙江永康人，博士，副教授，研究方向：计算语言学理论

Person Names Recognition Method Based on Statistic and Rules

Wu XueJun　Zhu JingBo　Chen XueYao　Zhuo HongXia

Institute of Computer Software & Theory, Northeastern University, Shenyang 110004

E-mail {wuxj1977@163.com , zhujingbo@yahoo.com }

Abstract This paper builds a knowledge base of person names recognition based on large-scale corpus and presents the method of person name recognition with statistics and rules . Our knowledge base of person names recognition include Chinese surname and its surname, Chinese name word and its usage probability, prefix and suffix leader word of Chinese name .This paper divides personal name into three partition: one is Chinese personal name, one is personal name without family name and the other is foreign name. Making use of knowledge database we constructed, we use the probability of personal name makeup and the method of combination of prefix and suffix leader word rules to individually identify these three kinds of personal name. Precision rate and recall rate are all considered in this method, which gains better recognizable effect. After test, the recall rate and precision rate are respectively 91.35% and 92.23% .

Keywords Person name recognition　Usage probability of Chinese surname　Usage probability of Chinese name word