蒙古文人名自动识别研究*

那顺乌日图　雪艳　淑琴　敖日格乐**

内蒙古大学蒙古学学院，**内蒙古明安途互联网技术开发有限公司 呼和浩特 010021

E-mail:mgnasun@imu.edu.cn; casvn@yahoo.com.cn; iragv@eyou.com; orgil@vip.sina.com

摘要：人名自动识别是语料库深加工及机器翻译等蒙古文信息处理工作中的重要环节之一。我们针对蒙古文人名词语的不同构成特点采取直接标注、词典匹配以及基于上下文的算法等方式进行自动识别。经过初步测试，该算法的识别召回率可达89%，准确率可达86%。

关键词：蒙古文，人名，自动识别，算法，语料库加工

Research on Automatic Recognition of Mongolian Names

Nasun-urtu　Xueyan　Shuqin　Orgil**

(Academy of Mongolian Studies Inner Mongolia University,

** Inner Mongolia Mingat Internet T&D Co.Ltd.)

Abstract: The automatic recognition of names is one of the important parts of Mongolian Information Processing such as corpora processing and machine translation etc. According to the different structures of Mongolian names, we implement the automatic recognition using directly tagging, dictionary matching and the algorithm based on context. The preliminary test shows that the recall rate reaches 89% with a precision rate of 86%.

Key words: Mongolian, names, automatic recognition, algorithm, corpora processing

一、引言

人名自动识别是蒙古文信息处理中的重要环节之一，其重要性尤其体现在语料库深加工

*此项研究得到国家自然科学基金资助（62063001）和国家社科基金项目（01BYY025）支持。

及机器翻译中。由于蒙古民族历史文化背景和蒙古文本身的特点，蒙古文人名识别也有其鲜明的特色，而这种特色也给蒙古文人名识别带来了特有的难点。

蒙古民族在历史长河中历经沧桑，接触过诸多民族，接受过多种文化，蒙古人的姓名也因其所生活的时代和地域的不同，留下了浓厚的时代、地区印记。纵观蒙古人姓名的发展，可以总结出以下几点：13 世纪文献中出现的人名基本上都是当时的蒙古语词语，如 OGEDEI（斡阔台）JUCI（术赤）CAGATAI（察哈太），这些人名在其后的近 8 个世纪中很多一部分没有传承下来。而 16-20 世纪蒙古族由于信奉佛教，梵、藏语人名开始盛行，如 SENGGE（森格），WCIR（敖其尔）等，与此同时，本民族语言人名和汉语人名也占有一定比重。到 20 世纪中期本民族语言人名有明显的上升趋势，而且在构词、意义方面都有了新的变化。这些蒙古语人名可以分为两个类：一部分是由双词根词组成的人名，例如，NASVNVRTV，由 NASV（寿）、VRTV（长）两个词构成，ALTANBAGAN_A，由 ALTA（金）、BAGAN_A（柱）两个词构成，而另一部分 BAYAR（喜）、AGVLA（山）、VLAGAN（红）等则由一个词根构成。除此之外，仍有一些汉语来源的人名，如．BOV GUWe CING（包国庆）、HONG SIYA（红霞）等。所以针对这些不同来源或构成的人名，必须采取不同的、相应处理对策。

本文主要以语料库人名识别、标注为目标，其中兼类人名的处理是本文的重点。因为无论从目前工作的需要，还是从技术含量的要求来讲，这部分工作具有较高的理论和应用价值。

二、蒙古文人名的构成特点、语料库分布及其处理对策

（一）词源特点及其对策

对蒙古文人名进行识别时，我们针对人名词语构成的不同特点采取不同的处理方法。其内容包括：直接标注；词典匹配；利用算法进行标注等。而在策略上由简而繁，先进行直接标记，再进行词典匹配，最后进行自动标注，在自动标注阶段仍由简单算法逐渐过渡到复杂算法。其具体做法为：

1. 汉语来源的人名中的姓氏部分或名字的前一个词都是单音节，这在其他来源的人名中是不会出现（或极少出现）的，所以对这部分词我们采取直接标记的方法。在区别汉语人名与汉语地名时采用词典匹配的方法（我们已有一部用拉丁文转写的汉语地名电子词典）。

2. 西方语言来源的人名具有易于识别的标记（在这类词中出现蒙古语固有词中不会出现的，专用记录借词的字符），如 FLADIMIRcOF（弗拉季米尔佐夫）中的 F、c 等，我们以此作为依据，把这部分词分离出来，通过词典匹配的方法与外国地名进行区分（我们已有一部用拉丁文转写的外国地名电子词典）。

3. 梵、藏语来源和由蒙古语双词根构成的蒙古文人名在文本中不会和普通词语混淆，所以还是采取词典匹配的方法（我们做了一部专用于人名名词的词典）。

4. 以单词根词形式出现的人名。如 BATU、AGVLA、CECEG、MINGGAN 等，它们易与普通词

语混淆，而且词典匹配的方法无法区分这些词的不同用法，所以只有通过语法条件和算法来进行识别、标注。

（二）词类特点及其对策

蒙古文人名中由单词根构成的词没有明显的区别性特征可作为识别标记，而且还有跨词类或兼类的特点。如，BAGATVR 一词可作为人名（巴特尔），又可作为普通名词（如 OREGDEGSEN BAGATVR“烈士”），也可作为普通形容词（如 TAL_A NVTVG-VN BAGATVR BICIHAN EGECI DEGUU“草原英雄小姐妹”）。虽然这些词在文本中作为人名出现的上下文环境大致一样，但由于不同词类的词具有不同的语法属性和语义属性，所以在进行识别时必须对不同词类的词分别进行处理。为了便于归纳出每个类的识别算法，我们对这些词进行如下归类：

1. 名词。我们根据其语义进行分类，并在“可做人名的普通词语词表（以下称“set1”）”中通过设置属性字段标记了其不同归属，如 CILAGV（意为“石头”，标记 N1，自然形成物）、ARSLAN（意为“狮子”，标记 N2，动物）、CECEG（意为“花”，标记 N3，植物）JIRGAL（意为“幸福”，标记 N4，抽象名词）、DALAI（意为“海”，标记 N5，自然现象）、SUHE（意为“斧头”，标记 N6，工具）、DABSILTA（意为“进步”，标记 N7，派生词）、ALTAI（意为“阿尔泰”，标记 N8，地名）。这种分类对区别普通名词和专有名词（人名）提供可作依据的上下文信息。

2. 形容词。以形容词的构成及功能分成两个分类，即“性质形容词”和“关系形容词”，并在“set1”中以属性字段形式进行描述。如 VLAGAN（意为“红”，标记 A1，性质形容词）、JORIGTV（意为“勇敢”，标记 A2，关系形容词）。因为就形容词而言，其构成及功能差异能够成为普通形容词和专有名词的区别特征。

3. 动词。以动词的变化形式分成三类，即“形动词形式”、“陈述式形式”和“词根形式”。由于他们作为人名和作为普通动词时的情况有很大差异，并且不同动词形式之间也有差别，所以在“set1”中进行描述是必要的。如，MANDVHV（意为“兴旺”，标记 V1，形动词形式），SVRVN_A（意为“学习”，标记 V2，陈述式形式），DVWALA（意为“闪闪发光”，标记 V3，词根形式）。

4. 数词。以数词构成的人名，发生歧义的主要是基数词（以复合形式出现的人名由词典匹配来处理），所以数词内部未进行分类。

5. 副词。蒙古语副词内部根据其意义、功能可分几个类，单可作为人名的副词为数极少，所以也无需进行再分类。

（三）在语料库中的分布情况

据统计，在 100 万词的蒙古文语料库中，带有人名标记的词语出现于 8435 个句子中，共出现 17096 次，人名在文本中占 1.7%（这个比例比中文人名出现比例高出近一倍）[2]。其中，只做人名的名词（不包括外族人名及重复出现的人名）占所有人名的 73.2%，而具有兼类性质的人名占 26.8%。这说明以词典匹配的方式可以识别出蒙古文文本中的大部分人名。而剩下的将近 30%左右的人名，或与名词，或与形容词，或与动词（形动词形式、陈述式形式或词根形式）兼类。这是蒙古族人名识别中的难点。

例如：HUDER MORIN-ACA-BAN SALAB BAGVL_A. （胡德日利落地跳下马。）在这个句子中HUDER是人名，而它又与形容词和名词兼类，作为形容词它具有“强悍的、强壮的、健壮的”等意思，作为名词它具有“矿石、矿物”的意思。因此，在这个蒙文句子结构中，HUDER完全可以充当定语而起到修饰“马”的作用。这样句子结构就产生了歧义。

蒙古文人名识别与中文人名识别也有很大差别，从有利的一面来讲，蒙古文人名基本上不存在切分问题，因为蒙古文中词与词之间是有空格的，所以两个空格之间的字符串可以认为是一个词（极少数人名有特殊情况外）。不利的一面是，蒙古文人名几乎没有像中文人名的姓氏字那样易于识别的标记。

由于在蒙古文中能够做人名的普通词（非专做人名词）数量并不太多，我们通过语料库查询和词典查询，共筛选出278个词做成“兼类人名表（set1）”，并从已有的“蒙古语语法信息词典”总库抽出其有关的语法信息，作为自动识别的知识资源。

三、蒙古文人名的识别算法及其实现

目前，我们拥有100万词级用拉丁文转写的粗加工语料库，库中人名已做了人工标注。为了能够较准确地识别具有兼类性质的人名，我们除了对100万词蒙古语文语料库中所有含有此类人名的句子进行了分析外，还分析了近500个含有与人名兼类但不以人名意义出现的词语的句子，从而获取了一些识别规则，并将这些规则做成算法，进行了测试。

规则算法中的符号说明：

C：当前词，F0：位于当前词前面的第一个词，F1：位于当前词前面的第二个词，B0：位于当前词之后的第一个词，B1：位于当前词之后的第二个词，B2：位于当前词之后的第三个词。

[set0]：只做人名集；[set1]：兼类人名集；[set1.a]：与形容词兼类的人名；[set1.v]：与动词兼类的人名；[set2]：称谓、职位名称集；[set3]：代词（部分）集；[set4]：人称领属语气词（部分）集；[set5]：搭配动词集；[set6]：基数词（从2至10）集；[set7]：集合数词集；[set8]：十位数词（后接动态词根）集；[set9]：岁数词集；[set10]：连接词和附加成分集；[set11]：地名集。

除以上符号外，规则算法中还会出现如BICIHAN,GAGCAGAR等以拉丁文转写的蒙古语词，这些词语由于经常只与表示人的词语（如人名等）同时出现，因此可以作为人名识别标志。

以下是我们归纳出的识别规则中的5个例子，这5个例子的识别标志各不相同：

R1：这个规则以有可能出现在人名前后的称谓及职位名称作为判断条件。

若C∈[set1]且F0∈[set2]或B0∈[set2]或F0=BICIHAN,则C为人名。

例：BATV SUJI AJIL-ACA-BAN AMARABA. （巴图书记退休了。）

R2：地名加上格附加成分这一格式所修饰的兼类词在蒙古文中为人名的机率很大。

若 C∈[set1]且 F0∈[set10]且 F1∈[set11],则 C 为人名。

例：ORDOS-VN MONGHE-YI TA TANIHV VV ?(您认识来自鄂尔多斯的孟和吗？)

R3：人称领属语气词是蒙古语的一个特点，可以用来判别与动词或形容兼类的人名。

若 C∈[set1.a]或 C∈[set1.v]且 B0∈[set4],则 C 为人名.

例：HOGJIHU MINI SAYIN SVR/V/LCA!(我的呼格吉夫你要努力学习！)

R4：人称代词和反身代词以及表示“独自”的词 GAGCAGAR 也可作为人名识别标志。

若 C∈[set1]且 B0∈[set3]或 B0=GAGCAGAR,则 C 为人名.

例：HUCUTU GAGCAGAR-IYAN EGURGE-BEN BEYELE/GUL/BE.(呼群图独自完成了任务。)

R5：在蒙古语中，表示年龄的数词后通常带有附加成分“TAI”，我们把这类词作成一个集合用作判别人名的依据。

若 C∈[set1]且 F0∈[set9]且 F1∈[set8],则 C 为人名.

例：ARBAN HOYARTAI NACIN ONCAGAI SERGULENG.(十二岁的那琴特别聪明。)

在分析文本序列中的人名时，以句子为分析单元。首先确定含当前分析词的整句左右边界，以空格、分词标示、特殊标点符号来切分词，并把切分词对应地压入线性索引列表中，同时建立参照表来标记分词、特殊标点符号的索引，以便分析时避免无效的计算，节省系统资源，优化计算速度。

按线性索引列表当前的索引来确定当前词（C）、前一个（F0）及后一个词（B0），甚至根据算法需要可取更多的词。通过不同算法的计算，返回的结果如果是人名可以在临时表中记录，并在文本序列中标注，与此同时向下移动线性索引列表当前的索引，移动的步长可根据参照表和此次计算的结果来调整，进入下一个词的分析计算，直到线性索引列表末尾。

清空线性索引列表进入下一个分析单元。这种算法对以后整句的语义和语法分析提供更多的实践。

我们在 20 万词语料库中做了初步测试。这 20 万词语料中，有人名标记的词语共出现 3518 次，其中汉语人名和外国人名占 49.8%（它们不包含在我们此次处理的范围内）；而我们力图识别的人名出现 1765 次，占 50.2%。通过系统初步运行，召回率达 89%，准确率达 86%。其中主要错误集中在没有任何识别标志的兼类人名上，而要达到更为满意的结果，除了应增加处理兼类人名的规则外，还必须依托更为复杂的句法知识和语义知识。

参考文献：

[1]刘开瑛：中文文本自动分词和标注 商务印书馆，2000 年，北京

[2]郑家恒、刘开瑛：自动分词系统中姓氏人名处理策略探讨，陈力为主编：计算语言学研究与应用，北京语言学院出版社，1993

[3]孙茂松、张维杰：英语姓名译名的自动辨识，陈力为主编：计算语言学研究与应用， 北京语言学院出版社，1993

[4]宋柔、朱宏、潘维桂、尹振海：基于语料库和规则库的人名识别法，陈力为主编：计算语言学研究与应

用．北京语言学院出版社，1993，150—154 页

[5]张跃、姚天顺：基于结合性自动识别中文姓名，小型微型计算机系统，第 18 卷，第 10 期

[6]方晓珊、盛焕烨：中国人名识别中规则抽取的一种基于实例的方法，第一届学生计算语言学研讨会论文集，2002

[7]Jimin Liu, Jing Xiao and Tat-Seng Chua: Finding Names in Chinese Text using a Hybrid Rule Induction Model, 第一届学生计算语言学研讨会论文集，2002

[8]吴雪军、朱靖波、陈学耀、卓红霞：基于统计和规则的人名识别方法（摘要），第一届学生计算语言学研讨会论文集．2002

[9]Shuqin: Helen-u MatEriyal-vn Homorge Dehi Monggol Homon-u Ner_e-yi Autocilan Siidburilehu Tvhai, obor Monggol-vn Yehe Svrgagvli-yin Erdem Sinjilegen-u Sedhul, 2002on-v 4 Duger Hvgvcag_a

[10]那顺乌日图：蒙古语语法信息词典框架设计，内蒙古大学博士学位论文，2000 年 6 月

[11]孙茂松、黄昌宁、高海燕、方捷：中文姓名的自动辩识，中文信息学报，第 9 卷第 2 期

[12]季姮、罗振声：基于统计和规则的中文姓名自动辩识，语言文字应用，2001 年，第 1 期

[15]李建华、王晓龙：中文人名自动识别的一种有效方法，高技术通讯，2000 年 2 月