统计和规则相结合的中文机构名称识别

张艳丽　黄德根 张丽静 杨元生

（大连理工大学计算机系 116024）

摘 要： 中文机构名称是专名的一种，量大且层出不穷，因而大多不能收入词典，这便给自然语言处理，尤其是机器翻译和机器理解带来很大困扰。本文将统计和规则两种方法结合起来，建立了中文机构名称的识别模型。系统闭式精确率和召回率分别达 92.5%和 92%，开式精确率和召回率分别达 88. 5%和 76. 6%。

关键词： 中文机构名称，单词频度，双词同现频度

Identification of Chinese Organization Names Based on Statistics and Rules

Zhang Yanli　Huang Degen　Zhang Lijing　Yang Yuansheng

(Department of Computer Science and Technology, Dalian University of Technology 116024)

ABSTRACT: Chinese organization names are one kind of proper noun, most of them can't be stored in the dictionary. Identification of Chinese organization names is very important to improve the accuracy of automatic word segmentation and the intelligibility of machine translation. In this paper, we present one model based on statistics and rules, in which we use the conception of the reliability for the word segment and some appropriate rules to identify Chinese organization names. The preliminary experiment shows that the precision and recall rate respectively reach 92.5% and 92% by close test, while the precision and recall rate are 88.5% and 76.6% by open test.

Key Words: Chinese organization names , Uni-gram Frequency , Bi-gram Frequency

1 引言

专名的自动分析和识别在汉语处理中占有很重要的地位。专名就是专有名称，是语义上具有专指性词语的总称，它在语言性质上包括两种语言单位：专有名词（老舍、北京等）和专有词组（大连理工大学、秋林公司等）。

专名是词汇系统中种类繁多，数量庞杂的一类词汇成分。文献[1]将专名划分为如下十类：

◆人名：毛泽东、鲁迅　　◆地名：北京市、大连

◆品名：长城电扇、茅台酒　　◆企业名：黑又亮日用品总厂

◆机关团体名：中国共产党中央委员会　　◆作品名：《红楼梦》、《蒙娜丽莎》

◆种族名：汉人、布衣族、中华民族　　◆时代名：唐朝、贞观、开元

◆节日名：春节、元旦、中秋节　　◆事务名：毛泽东号、淮海战役

学术界近年已经发表了不少关于人名和地名自动识别的论文[2][3][4]，并且已经取得了比较理想的结果，但对于专名中的机构名称识别（统指企业名和机关团体名）的研究则比较少[5]。针对目前的情况，本文对专名中的中文机构名称的自动识别问题作了较深入的研究。

2 基于统计和规则相结合的中文机构名称识别模型

首先，分析一下中文机构名称、中文人名（这里主要是指汉族的人名）及地名之间的差别和联系。从机器识别角度均可认为它们由两部分（本名和类名）组成：人名是由姓（类名）和名（本名）组成的，而机构名称是由机构名称特征词（类名）和机构名称前部词(本名)组成的，地名也是如此。但它们在词法和句法方面还是有很大差别的，人名中的姓和名都有比较好的性质，且中文姓名的字数一般也不超过4个字，使得中文姓名的自动识别比较容易；但机构名称用字和用词的随意性和机构名称长度的不确定性，导致机构名称的左边界很难确定，因而机构名称被认为是它们中最难识别的，而地名的识别难度则介于两者之间。

其次，我们简要介绍一下该识别模型的思想：首先从后向前扫描常规切分后得到的汉语句子，根据机构名称特征词表（*SpecOrg*）、机构名称前部词表（*ForeOrg*）和双词词典（*Bigram*）建立潜在机构名链；其次借助双词接续词表计算每一个潜在机构名的可信度；再次，根据可信度对潜在机构名链进行处理，在潜在机构名的同源对和互斥对之间做出抉择；最后，对分词结果进行调整。

定义**1** 机构名称的构造是P^+S，本文称P为机构名称前部词，而S为机构名称特征词（如：公司、大学等），即机构名称是由一个或一个以上的机构名称前部词加上机构名称特征词组成的。

◆本文识别的机构名称包括：

（1）企业名：就是指从事各种经济活动的大小公私经济实体单位的名称，包括：

①工矿企业名：大庆油田、鸡西煤矿、一休服装厂、龙滨不锈钢容器厂

②运输企业名：北方航空公司、哈尔滨铁路局、天鹅出租汽车公司

③建筑企业名：长城建筑工程公司、北方装饰实业有限公司、龙深建筑装饰联合公司

④金融企业名：中国银行、东洋金店、哈尔滨信托投资公司、华美城市信用合作社

⑤商业企业名：秋林公司、滨广电子产品销售公司、龙江酒类联销公司

⑥服务业企业名：金狮酒楼、三义扒肉馆、爱丽咖啡屋、陈氏接骨院、

（2）机关团体名：是指党政机关、人民团体等由国家经费开支、不进行经济核算的事业单位的名称，包括：

① 党政军机关名：中国共产党中央委员会、中国人民解放军总参谋部、哈尔滨市水产局

② 党派名：中国共产党、九三学社、中国国民党革命委员会

③ 军队名：中国人民解放军、中国人民武装警察部队

④ 科研教学机构名：中国科学院、故宫博物院、大连大学、哈尔滨师范大学附属中学

⑤ 宣传机构名：中央电视台、商务印书馆、中国少儿出版社、北方论丛编辑部

本文对机构名缩略语不进行识别，包括有机构特征词的，如：工程兵科研二所、铁十六局；无机构特征词的：如西安交大；再有，我们在对机构名称识别结果进行统计时，对已登录的机构名不进行统计，如：清华大学、中国共产党等。

◆机构名称的特征词和前部词

（1）机构名称特征词的性质是比较好的，大多数是普通名词，为数有限，比较容易收集。

（2）但机构名称前部词的用字和用词则比较随意，因而增加了机构名称识别的难度和复杂性：

① 前部词的词性比较复杂，而且存在大量的兼类情形。统计结果表明，其中，名词最多，其次是名动兼类词，然后依次是动词、数词、区别词、形容词等

② 前部词的词形也很复杂，有英文字符串（*TCL*集团、白宫*OTC*药物部）、数字（国际*21*世纪教育委员会）、成词单字（民族*和*宗教委员会、*市*药材公司）、不成词单字（香港*利*丰集团、深圳*赛格*集团）及一般的非单字词

③ 前部词中还有大部分未登录人名和地名，如重庆*黔江*嘉德酒业有限责任公司、*李宁*体育用品有限公司

④ 实际统计的结果还显示，机构前词（见定义2）和机构前部词有一部分是重叠的，这导致机构名称的左边界很难确定。

如：意大利教育*和*科研部，

朱荣基总理*和*含山县粮食局局长薄德豪进行了一番意味深长的对话。

⑤ 机构名称长度的不确定性，从三、四个字到几十个字不等；如管道局、山东省聊城市　东昌府区人民检察院起诉科等

定义 **2**　机构名称前词是指在实际文本中，机构名称的前一个词。本文中，我们对前词进行了分类：

①确定的前词：机构名称识别过程中，遇到该词就无需继续向前搜索（“这个、做好、来自、的、了、不久前”等）

②不确定的前词：机构名称识别过程中，遇到该词还需要继续向前搜索（“建成、开、城区、巨人”等）

③半确定的前词：机构名称识别过程中，遇到该词还需继续向前搜索，但它作为机构名称前词的可信度比不确定的前词要大些，需要根据具体情况进行分析（“和、及、与、由”等）

2.1 机构名称识别的统计模型

2.1.1 机构名称特征词的可信度

为了让计算机识别中文机构名称，首先必须构造两个词表，机构名称特征词词表和机构名称前部词词表。为此，我们对1998年《人民日报》近200万字的真实语料进行人工标注，从中抽取出近5000个机构名称；然后对这些机构名称进行处理和统计，建立了机构名称特征词词表和机构名称前部词词表。

然后，根据统计得到的数字，分别对机构名称特征词表和前部词词表中的每一个特征词和前部词计算其可信度。

目前共得到特征词1038个、前部词2528个。

本文对机构名称特征词进行了分析，将机构名称特征词分成以下四类，分别赋予不同的属性值（*Attribute*）

（1）*Attribute*=1：非单字特征词且非兼类词（大学、公司等）

（2）*Attribute*=2：非单字特征词且兼类词（组织、学会等）

（3）*Attribute*=3：单字特征词且非兼类词（店、厂等）

（4）*Attribute*=4：单字特征词且兼类词（局、部等）

定义 **3**　于$\forall S \in SpecOrg$，定义其可信度如下：

$$P_s(S)=\begin{cases} MAXORGSPEC * SPECAWARD1 & Attribute=1 \\ T(S) * SPECAWARD2 & Attribute=2 \\ T(S) * SPECAWARD3 & Attribute=3 \\ T(S) * SPECAWARD4 & Attribute=4 \end{cases}$$

$$T(S)=\frac{\log(N_S+2)}{\sum_{y\in SpecOrg}\log(N_y+2)}$$

其中，N_S 为建立机构名特征词词表时特征词 S 出现的次数，*MAXORGSPEC* 为机构名称特征词词表的可信度的最大值。

2.1.2 机构名称前部词可信度

定义 **4** 于 $\forall F \in ForeOrg$，定义机构名称前部词 F 的可信度如下：

$$P_f(F)=\frac{\log(N_F+2)}{\sum_{y\in ForeOrg}(N_y+2)}$$

其中，N_F 为建立机构名称前部词词表时前部词 F 出现的次数。

上面的定义只是用于得到最初的机构名称前部词的可信度，由于前部词的复杂性，为了平衡开式和闭式结果，也为了平衡潜在机构名称的频度值，识别模型中对前部词的可信度做了如下的调整：

（1）如果该词为英文字符串（“TCL,OTC”），赋予机构名前部词可信度的平均值

（2）如果该词为数字，若其后一个词为量词，赋予 0，否则赋予机构名前部词可信度的最小值

（3）判断该词是否为已登录地名，若是，同时依据位置信息在 *ForeOrg* 中查找后取出相应的频度值，再进行相应的奖励（乘以可调整参数 Pspname）；否则继续

（4）在 *ForeOrg* 中查找，若确切的查到（在确切的位置），设置标志位 Flag=1；若非确切的查到，Flag=2

（5）若该词是单字词，则要根据 Flag 的值、其本身的词性，及前后紧接的两个词和其是否在双词接续词典中出现赋予不同的值

（6）若该词非单字词，则要根据 Flag 的值、其本身的词性和是否在双词接续词典中出现赋予不同的值

注：（5）、（6）的具体实现，是通过采用不同的调整参数实现的。

2.1.3 构词可信度和接续可信度

定义 5 不考虑识别机构名称的分词方式为常规切分，考虑识别机构名称的分词方式为按机构名切分。

为了评价分词效果，本文也采用了构词可信度与接续可信度的概念。从 98 年《人民日报》上抽取 200 万字的语料，作为基础语料库，通过统计语料库中的各词与各对相邻词的出现频率，

建立单词词典（*Unigram*）与双词词典（*Bigram*）。

单词词典和双词词典的建立与论文《基于统计方法的中文姓名识别》中的词典建立方式是类似的，在这里就不赘述了，详细算法可见参考文献[2]。

如前分析，机构名称和人名的有一定的相似性，但也存在很大的差别，因而，本文在姓名识别模型的基础上，经过对中文机构名称特性的分析，总结出中文机构名称的构词可信度算法公式如下：

定义 6 在按机构名切分中，对 $\forall org = F^+S, S \in SpecOrg$ ， $F^+ = F_1F_2...F_n$ ，$n \leq \text{MaxSeg}$ ，$F_i \in ForeOrg$ （$i = 1,2.....n$），（n 代表机构名称前部词的个数）定义 *org* 的构词可信度：

$$P_w(org) = \omega \times [P_w(W_{k-1})P_w^{'}(W_k)P_w(W_{k+1})]^{1/3} + (1-\omega) \times [P_b(<W_{k-1}, ORG>)P_b(<ORG, W_{k+1}>)]^{1/2} \quad (1)$$

$$P^{'}(W_k) = C_n \times \sqrt{n} \times (\sum_{i=1}^{n} P_f(F_i) + P_s(S))/(n+1) \quad (2)$$

其中，$W_k = org$ ，W_{k-1} 与 W_{k+1} 分别为 *org* 左边与右边的词，当 *org* 在句首（句尾）时，W_{k-1}（W_{k+1}）按标点符号对待。$P_w(W_{k-1}), P_w(W_{k+1})$ 分别为单词 W_{k-1} 、W_{k+1} 的单词频度。$P_b(<W_{k-1}, ORG>), P_b(<ORG, W_{k+1},>)$ 分别为 $<W_{k-1}, ORG>, <ORG, W_{k+1}>$ 的双词接续频度。ω 为平衡潜在机构名称可信度与单词频度和双词频度可比性系数。$P_w^{'}(W_k)$ 是指不考虑上下文时，只依赖于机构名特征词词表和机构名前部词词字表计算出来 *org* 的可信度。*MaxSeg* 是指中文机构名称前部词所可能的最大词段个数。由于机构名称词典尺度与单词词典尺度不一样，为使它们之间可信度可比，所以采用 C_n 作为调整系数。

在上面的定义中，$F_i \in ForeOrg$ （$i = 1,2.....n$）是不确切的，因为机构名称用字用词的随意性，导致不可能像姓名识别中对名用字那样统计的那么完备，即无法建立绝对完备的前部词词表，所以在开式测试中，会经常遇到未训练过的前部词；因而，确定机构名称的左边界要借助于双词词典及机构名称本身的特性进行综合分析，需采用规则的方法来辅助实现。

2.2 机构名称识别模型中的规则

因为机构名称用字和用词的随意性，单独采用统计模型来识别中文机构名称的效果并不是很

理想。

机构名称识别的主要难度在于确定其左边界，因而，本文根据对机构名称本身和实际语料的分析，总结出以下几条规则，辅助确定识别机构名称的左边界：

（1）若该词为机构特征词，该词可为前词；并且在满足一定条件下，该词可为确定的前词

（2）若当前词段为标点符号，该词段可为确定的前词

（3）若该词可以作为机构名称前词且该词的后一个词为地名时，该词为确定的前词

（4）若当前词可为介词、代词，该词可为前词；并且若它们非单字词，可认为它们是确定的前词

（5）若该词后一个词为已登录地名，则该词可为前词

注：如果一个词是确定的前词是指不用继续向前搜索了，但这不表明该词就是正确的前词，我们还要对以当前特征词结尾的所有潜在机构名的可信度进行比较之后才可以确定到底哪一个才是正确的机构名；而非确定的前词就是指还需向前搜索，如果一直没遇到确定的前词，则会认为确定的前词是句首。

3 算法描述

（1）初始化。

（2）得到输入文本的常规切分序列。

（3）从后向前扫描常规切分序列，根据 *SpecOrg*、*ForeOrg* 和 *Bigram*3 个词表建立潜在机构名链（由于前界和后界的不同，这里可能会有很多交叉的潜在机构名），并根据上面的 3 个词表，计算每一个潜在机构名的可信度 $P_w^{'}(org)$。

（4）扫描潜在机构名链，当 $P_w^{'}(org)$ <FILERVALUE 时，删除该潜在机构名。（FILERVALUE 为潜在机构名的阈值）

（5）依据 *Unigram* 和 *Bigram* 中的频度，计算 $P_w(org)$，把潜在机构名链中的机构名按可信度降序排列，依次取出潜在机构名，删除排在它后面并和它交叉的潜在机构名称。

（6）删除太短的潜在机构名。

（7）根据保留在潜在机构名链中的机构名称，建立起按机构名切分的序列。

4 实验结果

本文所讨论的识别模型的正确性和有效性，首先依赖于前面所述的两个词表和两个词典的完备性与正确性；其次，系统的参数选择也很重要。上述这两点几乎是所有统计模型的共性[2]。

本文从 1998 年《人民日报》光盘版分别随机抽取 400、300 句进行闭式和开式测试，测试结果如下：

开式		闭式	
识别机构名称精确率	识别机构名称召回率	识别机构名称精确率	识别机构名称召回率
88.5%	76.6%	92.5%	92.0%

以下给出部分测试结果，共参考。~~符号之间表示机构名的识别结果，斜体表示正确机构名称：

①识别正确的部分结果：

该厂|与|外商|合资|兴建|了|~*加滨药业有限公司*~|。|
全国|最大|的|国有|破产|企业|~*山西纺织印染厂*~|在|破产|后|，|
马上|与|~*新乡市中青媒体传播中心*~|取得|联系|，|
当|~*索尼制片公司*~|于|１９９７|年|１１|月|宣布|拍摄|一|部|有关|詹姆斯|·|邦|德|的|电影|时|，|
继|一九九六|年|~*米其林沈阳轮胎公司*~|建成|投产|，|
一定|要|到|王兆兰|的|~*聚福隆茶园*~|去|看|一|看|。|
吸引|了|一|批|有|较|强|实力|的|企业|如|~*广夏实业股份有限公司*~|、|~*黄河垦殖公司*~|等|参与投资|荒漠化|土地|治理|，|
三轮|农|用|车|的|行业|骄子|~*巨力集团*~|认为|，|

②识别错误的部分结果：

组成|了|~有*吉斯肉类有限公司*~|、|~硕丰饲料加工厂~|、|~育民粮库种猪场~|等|７|个|紧密|层|企业|的|集团|，|
马丁|内斯|带|我们|去|参观|~韩国三星集团~|在|蒂|华纳|的|~客户工厂~|。|
~双马集团~|的|~核心企业*赤峰糖厂*~|建成|以来|连续|１３|年|盈利|，|
~工程项目学校~|————|~湖北省大悟县吕王镇汝青小学~|。|
在|第|二|届|*中国|环境|与*~*发展国际合作委员会*~|第|二|次|会议|开幕|之际|，|
随后|同|美国|的|~布洛克巴斯特影视公司~|和|*时代*|——|~*沃纳公司*~|合资|建立|了|几个|~广播公司~|，|

从目前进行的测试结果来看，该模型是有效的，对提高汉语自动分词的精确率有明显好处。随着训练集的扩大、学习机制和奖惩机制的建立，开式测试的精确率和召回率还有望进一步提高。

参考文献

[1] 徐国庆．现代汉语词汇学系统论．北京大学出版社，1999.4

[2] 黄德根　杨元生　王省　张艳丽．基于统计方法的中文姓名识别．中文信息学报，2001[2]

[3] 孙茂松　黄昌宁等．中文姓名的自动辨识．中文信息学报，1995,9（2）

[4] 沈达阳　孙茂松　黄昌宁．中文地名的自动识别．计算语言学进展与应用，清华大学出版社，1995

[5] 张小衡　王玲玲．中文机构名称的识别与分析．中文信息学报，1997,11（4）