汉语机构名的构成模式

雷静
北京大正语言知识处理研究院有限公司，北京 100083
ruiyunxuan@sina.com

摘要:机构名识别是未登录词识别的一个难点。本文探讨了在机构名识别中以机构名通名为激活信息，匹配通名对应的机构名模式来进行机构名识别的方法。提出了五个大类的机构名构成模式，并以此为依据，进行了机构名识别策略的设想。

关键词： 未登录词 机构名识别 机构名构成模式

The Patterns of Organization Names in Chinese

Lei jing
Linguistry Management Institute&Com.Ltd, Dazheng , Beijing, Beijing 100083
ruiyunxuan@sina.com

Abstract: Identification of organization names is the key nodus of Unkown Word Indentification. This article explores the approach of organization names' indentification based on the model, that use the general names of organization to be activation key, with matching organization's names corresponding general. This article also advanced five general sorts of model of construction of organization names, as well as the assumption of organization name's indentification according to the group categories.

Keywords: unknow word; identification of organization name; organization name's indentification

一 引言

机构名和人名、地名并列为三大类常见的专名，在自然语言处理中，专名是未登录词中的一个大类。所谓未登录词，也就是语言处理系统的词表中未收入的词。未登录词的识别是自然语言处理中要解决的一个重要问题。对专名的识别，已有很多研究成果，其中，人名和地名识别的研究成果比较多，已取得了比较好的效果，机构名的识别则相对比较薄弱。本文研究汉语机构名的构成模式，目的就是服务于机构名的识别，希望对改善这个薄弱环节有所裨益。

二 机构名及其识别方法

目前未登录词识别的解决方案主要分为两种：个别解决方案和一揽子解决方案。所谓个别解决方案，就是针对某一个类型的专有名词，专门设置一对一的解决方法；一揽子解决方案则是适用于所有类型的未登录词。

现阶段已有的个别解决方案中，一般都是针对人名或者地名的方法，因此目前的机构名识别方法基本上都是包括在一揽子解决方案中的，已经被提出的一揽子解决方案有：1988 年张普提出的有穷多层例举法；王开涛 1995 年提出的语料库统计的方法；同样是在 1995 年白拴虎提出的综合词性标注的方法；沈达阳和刘挺分别在 1997 年和 1998 年提出的局部统计的方法（主要针对人名）；1999 年陈小荷提出的的两趟分词、在“分词碎片”中计算单字成词概率和未登录词概率的一揽子解决方案；吕雅娟、赵铁军、杨沐昀、于浩、李生等人在 2001 年提出以对未登录词进行整体识别为目标，采用分解处理策略降低处理难度，并使用动态规划方法实现最佳路径的搜索，来解决未登录词之间的冲突问题的方法。

那么，机构名识别方法为什么与人名识别和地名识别不同，需要建立特有的机构明构成模式呢？

据我们粗略统计，汉语中汉族常用姓氏大概 1000 个左右。结合汉族人名构成的规律，和在语境中的出现状况，可以采取通过姓氏和前称谓信息确定前边界，然后借助汉语姓名用字表和另一部分称谓信息和语法知识确定后边界的方法来识别人名。相对于人名来说，地名用字比人名用字的范围更广，因此更有难度。在地名识别中，可以建立地名通名库，借此来确定地名的后边界，同时统计地名用字出现频率提高判断准确度。

机构名主要是机关、团体、社会组织和企事业单位的名称。它的数目庞大，并且随着社会和经济的发展不断扩大并产生新的名词。机构名不像人名有姓氏作为前边界激活信息，但它同地名类似，后面一般有通名，可以作为机构名是别的激活信息和用来辅助后边界。

我们设想如果在 HNC 的体系下进行机构名识别，大致应该按以下流程进行：初步分词——根据激活信息调入规则——进行语义块感知和句类假设——运用句类分析的结果进一步判定未登录词是否为机构名并且确认其前边界。

有了这个预期的流程，就需要建立运行流程需要的资源。资源主要有两个，一个是机构名通名库，另一个就是机构名构成模式库。机构名通名是指同类机构名的通用名称。如联想集团公司、北大方正集团公司，集团公司就是这两个机构名的通名。在机构名识别中，通名可以作为识别的激活信息。而机构名的构成模式则要找出无穷尽的机构名构词的共同特点。显而易见，机构名通名库的建立是比较容易的，本文讨论的就是建立的时候相对而言比较困难的模式库。因此我们这里主要探讨机构名模式库的建立。

三 机构名构成模式

为了研究机构名的构成规律，我们在互联网上搜索公司、企业、机构名录列表，找到了上万个机构名，在这个基础上对机构名的构成作研究。机构名的种类繁多，构成模式各异，因此我们在建立模式库的时候，也采取了分门别类的办法。机构名大致可分为五个大类和若干小类，每一类都有自己的构成模式，在机构名识别中，如果未登录词的构成和模式相匹配，那么则可以根据模式判定机构名。下面我们逐一对各类构成模式进行说明，其中（）内的内容表示可以没有。

1 团体、社团、民间组织名称的模式

这类机构一般属于民间自发性质，所属领域各异，从事的活动也包罗万象。这个类别大概有会馆、公会、公会联合会、联合会、同乡会、善社、华联会、崇正会、致公总堂、总商会、总会、协进会、工商会、同学会、研究社、积善堂、联谊会、堂、同宗会、从善堂、德善堂、敦善堂、归善堂、集善堂、良善堂、喜善堂、宗亲会、同乡总会、福利会、协会、分会、研究会、研究中心、等通名。

◆ 单个或多个地名+机构名通名

比如马来西亚安顺广东会馆，马来西亚、安顺、广东都是地名，会馆是机构名通名。

再如：马来西亚巴南广东公会、毛里求斯华联会、美国北加州苏浙同乡会

◆ 地名+与经济、文化、法律、科技、教育、卫生领域的专业活动有关的名词+机构名通名

比如成都市房地产业协会协会，成都是地名，房地产业属于经济领域的专业活动，协会是机构名通名。

再如：四平市城市住宅研究会、大连市房地产经济学会

2 与竞技、竞争、技艺等有联系的团队

这个类别的机构名通名主要就是队、团、代表队、代表团等等。

◆ 机构名+文化、科技等领域的专业活动有关的名词*+机构名通名

比如清华大学跳水队，清华大学是机构名，跳水属于文化中的体育，代表队是机构名通名。

再如：北京大学物理代表队、清华大学跳水队、北京联合大学辩论队

◆ 地名+文化、科技等领域的专业活动有关的名词+机构名通名

比如中国羽毛球队，中国是地名，羽毛球属于文化中的体育，队是机构名通名。

再如：北京射击队、英国芭蕾舞团、法国歌剧团

◆ 地名+ 名词+（文化、科技等领域的专业活动有关的名词）+机构名通名

比如芝加哥公牛队，芝加哥是地名，公牛是一个名词，队是机构名通名。再看山东鲁能泰山足球队，山东是地名，鲁能泰山就是一个一般的名词，足球属于文化中的体育，队是机构名通名。

其他的例子还有：北京国安队、北京奥林匹克数学队、西雅图超音速队

3 教育科研机构

在这个模式中，通名分为两类。一类是学院、大学、学校、中学、小学、幼儿园、托儿所等等，另一类在行政级别上下属于第一类，比如分校、系、学院、所、研究所、研究院等。值得注意的一点是，在两类通名中，有个别词是重复的，这时的判定就要从总体来把握了。

◆ 单个或多个地名+机构名通名

比如北京大学，北京是地名，大学是机构名通名。

再如：哥伦比亚大学、剑桥大学

◆ 表示地域的名词+表示专业活动的名词+机构名通名

比如北方工业大学，北方是一个地域，工业属于经济领域的活动，大学是机构名通名。

再如：华北科技学院、北京理工大学

◆ 地名+表示专业活动的名词*+机构名通名

比如北京师范大学，北京是地名，师范属于教育类别，大学是机构名通名。

再如：北京工商大学、北京农学院

◆ 地名+“立”+机构名通名

比如佛罗里达州立大学，佛罗里达州是地名，大学是机构名通名。

再如：佛罗里达州立大学、衣阿华州立大学、北卡罗来纳州立大学

在这里，有一个需要说明的问题。还是以佛罗里达州立大学为例，从语义上分析，这个词的结构应该是佛罗里达+“州立”+大学，但是由于在现实中，学校有省立、国立、县立等等，如果按照这个模式，系统就很难做出准确判断，因此，为了日后程序处理上的简化，统一把模式改为了标题中的式样。

◆ 人名+机构名通名

比如约翰．霍普金斯大学为例，约翰．霍普金斯是人名，大学是机构名通名。

再如：陈经伦中学、劳伦斯大学

◆ 地名+基数+机构名通名

比如北京 101 中学，北京是地名，101 是基数，中学是机构名通名。

再如：北京 25 中学、北京 185 中学

◆ 地名+序数+机构名通名

比如北京市第五十四中学，北京是地名，第五十四是序数，学校是机构名通名。

再如：北京市第二十九中学、北京第 80 中学

◆ 地名+词表收录词或未成词单字串+机构名通名

比如北京培英小学，北京是地名，培英是动词，小学是机构名通名。

再如：北京汇文中学、北京育英学校、北京实验小学

◆ 已判定的学校类机构名+表示学科的名词或词表收录词或未成词单字串+机构名通名

比如北京师范大学中文系，如果判定出北京师范大学是学校类机构名，中文是表示学科的名词，系是机构名通名。

再如：北京师范大学附属中学、清华大学物理系、伊利诺伊大学芝加哥分校

4 金融机构名称

在这个类别里，我们先搜集了中外银行名录，收入词表，然后从中找出了银行名的构成模式，大致有以下几类。

◆ 银行名称+地名+（序数）+支行

如：交通银行深圳红荔支行、交通银行深圳南山支行、中国银行呼和浩特分行、建设银行呼和浩特第二支行

◆ 银行名称+地名+分行+（地名）+（序数）+（办事处）

如：中国银行集宁市分行、中国银行二连浩特市分行、建设银行包头分行民东路办事处、交通银行石家庄分行第四办事处

◆ 银行名称+（地名）+（“支行”）+地名+办事处

如：建设银行巴盟二营解放街办事处、工商银行朔州市支行城建办事处

◆ 银行名称+地名+分行+房地产信贷部

如：工商银行邢台分行房地产信贷部、工商银行邯郸分行房地产信贷部

◆ 银行名称+地名+分行+信贷业务部

如：交通银行唐山分行信贷业务部、交通银行秦皇岛分行信贷业务部

◆ 银行名称+地名+分行营业部

如：工商银行张家口分行营业部、交通银行太原分行营业部

◆ 银行名称+地名+支行营业部

如：建设银行北京东四支行营业部、建设银行北京朝阳支行营业部

◆ 银行名称+地名+“分理处”

如：建设银行北京劲松南分理处、建设银行北京潘家园分理处、农业银行北京交道口分理处

5 其他企事业单位名

在这个类别中，由于把企事业单位和在一起通名杂，主要有党校、局、公司、有限公司、有限责任公司、集团、集团公司、总公司、分公司、厂、有限责任公司、责任公司、加工厂、制造厂、经销处、医院、职工医院、人民医院、门诊部、人民政府、部、局、饭店、酒店、大酒店、大饭店、宾馆、山庄、国际酒店、商务酒店、饭店、中心、假日饭店、国际交流中心、公寓、酒店式公寓、假日大酒店、俱乐部、度假村、度假饭店、休闲会馆

◆ 地名+机构名通名+（地名）+（机构名通名）

比如安国市中医院北关门诊部，安国市是地名，中医院是机构名通名，北关是地名。

再如：安福县人民医院、安国市中医院北关门诊部

◆ 地名+（表示专业活动的名词）+（词表收录词）+机构名通名

比如上海建设党校，上海是地名，建设是表示专业活动的名词，党校是机构名通名。

再如：中国国际工程咨询公司，中国是地名，国际是名词，工程是表示专业活动的名词，咨询是动词，公司是机构名通名。

其他的例子还有：包头东河区房产管理局、上海建委党校、重庆市人民政府

◆ 地名+（序数）+表示专业活动的名词+（序数）+机构名通名

比如中国航空工业第一集团公司，中国是地名，航空工业是表示专业活动的名词，第一是序号，集团公司是机构名通名。

再如：中国航空工业第二集团公司、中国第一汽车集团公司

◆ （地名）+（词表收录词或未成词单字串）+（表示现代人造物或宏观基本物的名词）+机构名通名

比如春兰集团，春兰可以被认为是未成词单字，集团是机构名通名。

再如：深圳莱英达（集团）有限公司，深圳是地名，莱英达是未成词单字，有限公司是机构名通名。

其他的例子还有：长虹电器有限公司、长春市长城防水材料厂

◆ （地名）+词表收录词+（表示现代人造物或宏观基本物的名词）+机构名通名

比如北京联想计算机集团公司，北京是地名，联想是词表收录词，计算机是表示现代人造物的名词，集团公司是机构名通名。

再如：熊猫电子集团公司、北京昆仑饭店

◆ （地名）+英文单词或字母组合+（表示现代人造物或宏观基本物的名词）+机构名通名

比如 TCL 公司是字母组合，集团公司是机构名通名。

再如：日本 NEC 公司、美国 IBM 公司

◆ 单个或多个地名+（未成词单字串或词表收录词）机构名通名

比如，北京大饭店,北京、中国都是地名,大酒店是机构名通名。

再如：澳门京都酒店、北京诺富特和平宾馆

四 机构名模式库的建立

上面简要分析了机构名的构成模式，其中机构名的构成模式被分成了 5 大类，这是因为我们在研究机构名模式的时候发现他们在词语构成上具有区别于其他类的共性。可以发现在五个大类的子类中，有一些构成模式是类似甚至相同的，这是兼顾了五个大类整体的语义而进行的划分。

在建立模式库的时候，可以采取模式库与机构名通名库对应的方法。我们认为由于机构名的种类繁多，决定了机构名通名库的建设不像地名通名库那样单一，因此在机构名通名库中，应该建立一个索引，用以标注该通名适用于上面所表述的模式库中的哪一类，并以此来调用对应的模式，缩小匹配范围。因为有些子类模式有重合，还应该出现一个通名对应多个模式的情况。

五 结语

本文是关于机构名识别的理论推想，还没有落实到具体实现。因此在这里，只能提出自己的设想，还不能运用到大规模的语料中去检验，也没有召回率和正确率的数据。今后的工作是，进一步完善我们的设想，使之与软件相结合，争取在模式库中穷尽所有的机构名模式，并进行实例标注测试，建立实用的机构名识别体系，相信在测试中将会取得很好的效果。

此外，关于机构名识别有一个很重要的问题还没有解决，那就是在上下文中的省略问题。比如北京师范大学，在上下文中可能省略为北师大；北京大学可能省略为北大；如何识别省略了的机构名，是以后需要研究的一个课题。

参考文献

[1] 苗传江，HNC 句类知识研究，中国科学院声学研究所博士学位论文， 2001 年 8 月

[2] 黄曾阳，《HNC（概念层次网络）理论——计算机理解语言研究的新思路》， 清华大学出版社 1998 年 11 月第 1 版

[3] 沈达洋 孙茂松 黄昌宁，局部统计在汉语未登录词辨识中应用和实现方法，《语言工程》，清华大学出版社，1997 年

[4] 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 年第三期

[5] 吕雅娟 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》，2001 年第一期