基于构成模式的汉语机构名识别

雷静[1]，张舵[1]，冯霞[2]

(1. 北京大正语言知识处理科技有限公司, 100081; 2. 北京师范大学中文信息处理研究所, 北京, 100875)

摘 要：汉语机构名识别是现代汉语未登录词识别中最难实现的一项，至今未能找到十分令人满意的解决方法。本文提出了一种基于构成模式的汉语机构名识别方法，其基本思想是根据机构名的构成模式，建立模式的转移矩阵，在转移矩阵的基础上用自动机算法实现中文机构名的自动匹配。

关键词：机构名；未登录词识别；构成模式；

Recognition of Chinese Organization Name Based on Constitution Pattern

Lei jing[1], Zhang duo[1], Feng xia[2]

(1. Linguistry Management. Ltd, Dazheng , Beijing 100081; 2. Institute of Chinese Information Processing, Beijing Normal University, Beijing, 100875)

Abstract: Automatic recognition of organization name is modern Chinese has not registered an item which in the word recognition is most difficult to realize, until now has not been able to find the extremely satisfying solution.This paper presents an approach for organization name recognition based on the constitution pattern. Its basic thought is according to the organization constitutionpattern, establishes the pattern the shift matrix, realizes a Chinese organization automatic match in the shift matrix foundation with the automaton algorithm.

Keywords: Organization Name ; Unknown words recognition;Pattern;

汉语机构名具有涵盖范围大、用词广泛、长度极其不固定，含有大量的简称的四大特点。它是现代汉语未登录词识别中最难实现的一项，至今未能找到十分令人满意的解决方法。本文在大规模真实语料的基础上，从机构名的构成模式入手，探讨了一种采用规则导向、基于构成模式的机构名识别方法。

1 机构名构成的特点

通过对未登录机构名的分析，我们发现，机构名虽然数量繁多、复杂多样，但是它同时有很多利于识别的特点：

（1）机构名的激活条件比较明确

所有机构名的全称都是以通名激活的，机构名通名是指同类机构名的通用名称。不同的通名激活不同的机构名类别相对固定。如："公司"只能激活企业类机构名；"大学"只能激活教育机构类机构名。这样我们可以通过建立通名库激活机构名并确定机构名的类型。

（2）机构名的构成模式相对固定

每一类型的机构名都有它相对固定的构成模式。如："公司企业名"的一般构成模式为："地名+名称词+行业+企业通名 （其中地名、行业词可省略）"。我们对收录的 12000 多公司名称进行统计后发现，85%以上的公司名称都遵循这条模式。模式相对固定特性就有利于我们从机构名

内部构成模式入手来进行识别。

（3）机构名角色词的词性范围确定

充当机构名角色词的名词最多，形容词其次，同时还有少量的动词和数词。其它类词性很少出现。这样有利于我们从词性范围限制角色词库。

（4）机构名角色词的类型确定

角色词一般由以下几类构成：名称词，地名，国名，行业学科，数词，字母组合，功能方式，年龄，性别，人名，机构名组成。其中的名称词是指机构名中专造的名称，如："联想集团"中的联想，"清华"大学中的清华。

（5）角色词的组织顺序相对确定

一般规律是：地名，人名，机构名等表从属关系的词在前；名称词，数词，字母组合等表名称的词仅接其后，后面跟行业学科，功能范围等功能词。如：一般可接受"北京大正语言研究院"，但不能接受"大正北京语言研究院"或"语言北京大正研究院"的名称。

（6）大多数机构名有较强的前后边界提示

大多数机构名以地名作为前边界，参考文献[5]统计指出 68.7%的机构名以地名作为开始。同时很多机构名前后跟有前导词和后导词。这有利于前后边界的确认。

机构名的数目虽然庞大，并且随着社会和经济的发展不断产生新的名词，但从总体上来说，都有一定的命名规则，在理论上可以通过对已有机构名的分析列举出现有机构名构成的所有规则。其构成模式相对固定，激活条件比较明确，各角色词间有较强的组织顺序，同时还有地名，前后导词等边界提示。这些特点使得从分析机构名内部构成模式入手进行识别成为一个具可操作性的识别思路。

2 机构名构成模式分析及相关资源的建立

2.1 汉语机构名的构成模式

机构名的构成模式各异，我们在建立模式库的时候，采取了分门别类的办法，把机构名分为九个大类和若干小类，每一类都有自己的构成模式。每个模式由若干模块加符号再加机构名通名组成。其中所用到的符号定义如下："+"：模块各依次出现；"<>"：该模块可以省略；"/"：其前后的模块只能出现一个；"|"：其前后的模块出现顺序不定；"()"表示通名属性。同时规定：通名前的各模块至少出现一个，分支机构名不能单独出现。

1 企业名

1.1 <地名>＋名称＋<行业>+<数词>＋企业通名

1.2 人名+〈行业〉＋企业通名

1.3 <地名>+字母组合+企业通名

1.4 地名+企业通名(分)

1.5 企业机构名+驻+地名+企业通名(分)

2 教育科研机构

2.1 <地名/表示地域的名词>|<机构名>|<数词>|<人名>|<名称>|<办学方式>|<分科/行业> + 教科通名

2.2 <地名>+<名称>+教科通名(分)

2.3 <学科>+教科通名(分)

2.4 教科机构名+"附属"|<数词>+教科通名

3 体育团队

<地名>+<机构名>+<专造名称词>+<体育项目>+<数词>+<年龄>+<性别> + 体育通名

4 金融机构名

.4.1 <地名>+<名称/行业>+金融通名

4.2 <地名>+<数词>+金融通名(分)

4.3 <地名>+<数词>+金融通名(分)

5 政党

地名+<名称>+政党通名

6 组织，社团

6.1 <地名>|<名称>+<行业>+组团通名

.6.2 政党机构名/教科机构名/组团机构名+<地名>|<名称>+组团通名

7 政府机构

7.1 <地名>+<数词>+<行业>+行政通名

7.2 <地名>+行政通名(分)

7.3 <国名>+"驻"+<国名>+行政通名

7.4 政府机构名/教科机构名/企业机构名/公众机构名/政党机构名/军警机构名+行政通名

8 公众事业（医院、气象台）

8.1 <地名>+<行业>+<名称>+<数词>+公众通名

8.2 机构名+"附属"+<地名>+<数词>+<名称>+公众通名

8.3 <地名>|<行业>+公众通名(分)

8.4 <公众机构名>+<行业>|<数词>+公众.出版通名

.8.5 <公众机构名>+公众.医疗通名

9 军警单位

9.1 <地名>+<数词>|<名称>+<性别>+军警通名

9.2 <地名>+<数词>+军警通名(分)

9.3 军警机构名+"驻"+<地名>+<数词>+<军警通名(分)>+军警通名(分)

2.2 机构名识别涉及的相关资源

我们下载了北京大学标注的《人民日报》1998 年 1 月的语料，对其进行了人工分析和机器统计，在此基础上初步建立了机构名识别的各类资源。

1 机构名前导词词表

前导词是确定机构名前边界的重要信息，目前我们已建立包括 385 条记录的前导词词表。

2 机构名通名表

我们从北大语料中提取出不重复的机构名共 4397 个，经过人工筛选，形成了 1027 个词组成的通名表。每个通名都标注了通名等级和所能激活的机构名类型。通名等级是我们给通名设立的级别，以“1”、“2”、“3”命名，分别表示该通名级别只能为高级；只能为低级和既能为高级又能为低级。如“北京大学”是一个完整的机构名，不能再分出其它，“大学”即为高级通名。又如“北京大学中文系”，“中文系”接在“北京大学”后面时，整个词才能成为机构名，“系”的通名属性就为低级。再如“北京大学教育学院”，“教育学院”接在“北京大学”后面时，“北

京大学教育学院”才能成为机构名，此时，“学院”的通名属性就为低级；但是，同时也存在“北京教育学院”这样完整的机构名，此时“学院”即为高级通名，所以“学院”的通名属性就为既能高级又能低级。

3 模块对应词表

第三部分构成模式的模块中除了“名称”类有很大的随意性之外，基本上都是有规律可循的，因此我们建立模块对应词表，设立词形和对应模块字段，尽量为绝大多数模块举出与之相对应的词。这些模块包括：行业、学科、办学方式、性别、年龄、体育项目、国名、分科等，共3926个词条。

3 识别方法

识别算法我们采用了自动机的思想，先根据每一条规则的状态图建立起一个转移矩阵。以机构名通名作为激活条件。根据通名的类型和等级组合对应规则的转移矩阵，建立起一个组合转移矩阵，然后进入自动机进行识别。

3.1 算法描述

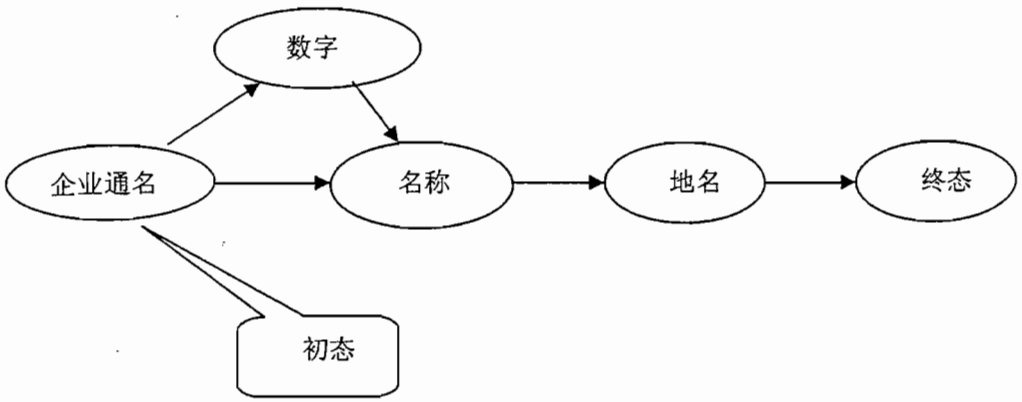
（1）初始化，加载所有单个规则的转移矩阵。（2）对输入文本按常规切分进行分词。（3）从后向前扫描切分序列，当匹配上机构名通名时激活。（4）根据通名的类型和等级组合对应规则的转移矩阵。对于既可以作一级通名又可以作二级通名的词优先按一级通名匹配。（5）根据组合的转移矩阵进入自动机进行状态的自动切换。（6）当自动机转换到每一个可能作终态的角色就把它作为潜在机构名，并根据评价系统给出该机构名的可信度值。遇到只能作终态的角色退出该自动机。（7）根据可信度值在机构名的同源对和竞争对间作出选择，确定最终认定的机构名序列。

3.2 算法示例

3.2.1 单个规则转移矩阵的建立

如：对于规则[1.1] <地名>＋名称＋<行业>+<数词>＋企业通名

它对应的状态转化图如下：



根据该图建立的转移矩阵如下：

	通名	数字	行业	名称	地名	初态	终态
通名	0	1	1	1	0	1	0

数字	0	0	1	1	0	0	0
行业	0	0	0	1	0	0	0
名称	0	0	0	0	1	0	0
地名	0	0	0	0	0	0	1

表 1 单个规则的转移矩阵

上表描述了每条规则的状态转换情况，若从 A 到 B 有路径通过，则 A 行 B 列处填 1，否则填 0。

3.2.2 激活通名对转移矩阵的组合

激活通名对转移矩阵的组合即根据通名的类型和等级组合对应规则的转移矩阵。

如：激活通名“有限公司”，从属性中得出它的类型为：企业通名。它的通名级别为：只能作高级通名。于是它可以激活所有的高级企业通名对应的模式。即以下模式：[1.1] <地名>＋名称＋<行业><数词>＋企业通名；[1.2] 人名+〈行业〉＋企业通名；[1.3] <地名>+字母组合+企业通名

将这些模式对应的转移矩阵进行组合即可。

值得注意的是，我们所说的组合不能是矩阵间的简单相加，那样做可能会添加出新的实际不存在的路径出现。如：路径 1：a 1+ b +c +d.；路径 2：a2 +b + c + e。若矩阵简单相加就会产生实际不存在的路径 3：a 1+ b +c +e，由此我们在组合矩阵的对应项填激活的模式号序列。

3.3 对细节问题的处理

3.3.1 对名称词的处理

由于名称词是一个开放集，多数名称词为未登录词，因此名称词无法事先收录，必须动态的进行识别。我们观察发现，名称词的用词具有以下特点：（1）词性上绝大多数是名词，形容词，同时含有少量的动词。几乎没有其他词性。（2）表示积极向上和褒义成分的词比较多。（3）大多数是分词碎片的组合，真正的成词不多。

基于这些特点，我们建立了名称词常用词库、禁用字库、禁用词库、禁用词性库，在此基础上对名称词进行动态的识别。

3.3.2 对竞争对和同源对的处理

若同时识别出来的一对机构名，激活条件相同，但前边界不同，我们把这一对机构名称为同源对。若同时识别出来的一对机构名，字符序列有交叉部分，我们把这一对机构名称为竞争对，如对字符串序列：ABCDEF 从右向左扫描，若同时识别出 “BCDE”和“CDE”则它们构成同源对。若同时识别出“ABC”和“BCD”则它们构成竞争对。

我们对竞争对和同源对所采用的处理方法是：对每一个潜在机构名，根据评价系统给出的它的可信度值，然后根据可信度值在竞争对和同源对间做出抉择。

该评价系统根据以下特点加权进行建立：（1）该模式在真实文本中的出现概率。（2）模式中是否跟有地名。（3）机构名前后是否跟有前后导词。（4）模式中含有的名称词的个数。

4 实验结果与分析

我们利用初步建立起来的模型进行分析实验，实验语料为北大标注的《人民日报》。语料实验结果如下：

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)

人民日报	11445	10863	8911	82.03	77.86	79.89

表 2 测试试验数据一

注：(1)TOTAL：语料中所有的机构名数；FOUND：系统识别出的机构名数；RIGHT：系统识别正确的机构名数

(2)P：机构名识别的正确率=RIGHT/FOUND×100%；R：召回率=RIGHT/TOTAL×100%；

F：综合指标=2×P×R/（P+R）×100%

在实验中我们发现一些影响面较大的错误类型，经过分析，发现这些错误类型其实可以通过若干的附加规则即可有效解决。主要集中在以下 3 个方面：

1、没有特指类概念的名称词误作为机构名来收录。如：①李鹏就曾多次到<nt>发电总厂</nt>检查指导工作。②全国绝大部分省市区的出版结构基本上是一个模式：人民社、文艺社、少儿社、教育社、科技社、美术社、古籍社等。③特区政府新闻发言人表示，去年年底前，所有政府部门均已在互联网上设立网页，互联网使用者通过“<nt>政府资讯中心</nt>”便可进入各政府部门的网页，浏览到各政府机构的服务、工作及组织架构等资料。例①中的“发电总厂”、例②中的“人民社、文艺社、少儿社”及例③中的“政府资讯中心”均误作为机构名收录了。

所谓机构名指的是特指的一个机构，而上述词都是作为泛指的词，不属于机构名，但程序很容易把它们误作为机构名来处理。因为能够充当机构名的特指概念的只有地名，人名，名称词等。所以我们建立附加规则 1)：地名，人名，机构团体，名称词等表示特指概念的词必须在机构名中出现一个。

2、一些与地域相关的角色词没有召回。如：“华中”，“东华”，“中央”，“中南”，“东北”等词。由此我们建立附加规则 2)：将地名的范围扩大，一切表示：地方，国家，范围，方位的词都都属于地名。

3、模式的误召回。通过对测试结果的分析可以看出，“名称+通名”或“地名+通名”的模式误召回的词很多。

这是由于名称词是需要动态识别的，其边界本身就不好确定；再加上通名可能会出现的兼类情况。如：“中，台，所，社，队，军，局，部，组织，学会”等词，这些词作为通名和不作为通名的可能性都很大，一般称之为通名兼类词。这样一来，名称词或地名直接跟一个通名兼类词误召回的可能性就很大。

我们的解决办法是：首先对于所有通名我们计算出在标注语料中该词作为一般词和通名的次数，取其比值作为考察兼类性的概率。因此给出附加规则 3)：当兼类通名遇到“名称+通名”或“地名+通名”的模式时，必须跟有前导词或后导词，该模式才能成立。同时我们也发现：没有只含有一个字的名称词，名称词的字个数几乎都小于 7。因此建立附加规则 4)：名称词的字个数须大于 1 小于 7。另外，参考文献[2]的统计数据表明：一般组成机构名的用词都小于 8 个词。就此建立附加规则 5)：机构名的用词个数小于 8。

在增加 5 项附加规则的基础上，我们对机构名做重新识别，用北大标注的《人民日报》语料进行测试，结果如下：

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报	11445	11242	9814	87.30	85.75	86.52

表 3 测试试验数据二

由表 3 可以看出，新规则的识别能力有了很大的提高。我们认为如果对模式进行进一步分析，会取得更好的识别效果。

另外我们也发现，基于构成模式的汉语机构名识别还存在以下三大问题：

一是由于基于构成模式识别算法本身的特点，该算法效果的好坏很大程度上取决于词库的质量。由于行业词，通名词等词收录不全，直接造成模式无法正确匹配。这些基础资源缺乏是影响识别效果的最主要原因。我们认为，虽然行业词，学科，通名等在理论上是可以穷举的封闭集合，但仅靠手工来把它们很好的收集。下一步可考虑辅助通过机器学习的方法来扩充模块对应的词库。

二是分词效果影响识别效果。由于机构名的识别是在分词的基础上做的，歧义切分的部分肯定无法正确识别出它的模式。如："美国人/口/研究所/""老挝人/民革/命/党/""香港大学/生"等。

三是由于目前我们所做的对机构名的模式分析是针对一般构成模式的分析，所以对于复杂构成的模式还无法识别出来。如："中国人民/争取/和平/与/裁军/委员会/""/老挝/和平/与/团结/委员会/"等。对于这类词的识别方案，我们认为可以先根据这些模式的特点对模式库进行扩充，但模式库的扩充并不是越多越好，因为过多的模式容易造成更多的误召回。所以对于那些特别复杂的模式，如"联合国/销毁/伊拉克/化学/"就不能进行收录。对于无法进行收录的模式可考虑先对它进行模糊匹配，待确定好边界以后再分析它的构成模式。另外也可考虑通过对标注语料的训练进一步发现更多的模式，从而对模式库进行扩充。

这些问题将是我们下一步的工作重点。

5 结束语

目前的测试结果可以证明，采用规则导向的基于构成模式的机构名识别方法是有效可行的，通过对模式的细化和库资源的扩充，效果上也还有很大的提升空间。另外，对于前边界的竞争问题等可考虑在模式的基础上引入统计模型来实现，可以将理性主义和经验主义很好的结合起来。

本文的实验结果也证实了大多数机构名的构成模式都有章可寻，机构名和专有名词的识别可以通过分析其内部构成规律达到较好效果。诚然，在机构名识别中存在复杂模式构成和简称等模式不确定的因素，但正是这些局部上的不确定性说明了机构名和专有名词的识别很值得进一步的研究，需要"绣花般精雕细刻的耐心"。

参考文献

[1] 俞静. 汉语机构名的构成模式. 语言计算与基于内容的文本处理,清华大学出版社.2003.p85-p90

[2] 张小衡 王玉玲. 中文机构名称的识别与分析. 中文信息学报.1997(4)

[3] 王宁 苑春法等. 中文金融新闻中公司名的识别. 中文信息学报.2002(2)

[4] 张艳丽 黄德根等. 统计和规则相结合的中文机构名称识别. 自然语言理解与机器翻译,清华大学出版社.2001.p233-p239

[5] 吴雪军等. Co-Traing 机器学习法在中文机构名识别中的应用. 语言计算与基于内容的文本处理,清华大学出版社.2003.p85-p90