中文基本地名识别

钱小飞 侯敏

中国传媒大学 国家语言资源监测与研究中心 有声媒体语言分中心 北京 100024

Email:qierflying@163.com, houminxx@263.net

摘要： 本文探讨了地名的概念、构成等问题，并针对其分布特点，使用常见地名匹配、碎片分析和组合扩展相结合的方法初步识别了中文基本地名，包括中国地名和汉译地名。开放测试取得了 88.16%的正确率和 87.32%的召回率。

关键词： 中文基本地名；命名实体；识别

Recognition of Chinese Basic Location

Qian Xiaofei, Hou Min

National Language Resources Monitoring and Research Center (Broadcast Media Language Branch), Communication University of China, Beijing 100024

Email:qierflying@163.com, houminxx@263.net

Abstract: This paper discusses the concept of location as well as its composing, and it combines some methods such as common location matching, segment fragment analysis and words extension to recognize Chinese Basic Location, including the Locations of China and the foreign Locations in Chinese. The experiment achieves about 88.16% in precision and 87.32% in recall.

Keywords: Chinese Basic Location; Name Entity; Recognition

1 引言

命名实体识别是汉语自动分词的难题。地名作为一种常见的命名实体，在文本中有着较为广泛的分布。它的识别可以有效地提高分词精度，同时在信息检索和问答系统等领域也有着重要的应用。

人脑识别地名通常有三种手段：匹配已知地名，基于内部构成猜测未知地名，基于上下文知识猜测未知地名。在以往的研究中，这三种手段常常以不同的组合方式集成到统计和规则系统中。主要的研究有：沈达阳等（1995）利用属性矩阵和频级进行筛选识别地名，刘开瑛（2000）根据地名词典和语料库估计地名首中尾字的出现概率，并通过各位置用字的概率限制和上下文规则识别地名；在此基础上，谭红叶等（2001，2002）使用转换及基本地名匹配的方法有效地提高了精确率。黄德根等（2003）基于地名词表定义了地名的构词可信度，并进一步利用上下文信息定义了接续可信度，取得了较好的效果。

在识别策略上，如果将已知地名及其内部构成、在语料库中的上下文标志看作是前景信息，那么非地名构成成分和普通词则是地名识别的背景信息；以往的研究主要着力于前景特征的提取，而较少从识别背景信息考虑。而据陈小荷（1999）的研究，大量的未登录词存在于分词碎片中，现有的自动分词系统将词表中查不到的单字都权且当作一个词，是不能发现和识别未登录词的根本原因，因而通过单字概率和单字词概率等背景信息估计未登录词概率是一种有效的策略。

从评价手段看，以往的研究也不完全一致。刘开瑛（2000）将“湖北京山县”作为两个地名，

而黄德根（2003）对相邻的地名作了合并，将“江西省九江市江矶村”识别为一个地名。简单的相邻地名组合可能会带来一定的风险，首先，可能将系统本来识别错误的地名合并成一个正确的地名；其次，可能将不相干的或处于并列关系的地名合并为一个地名，如“北京中国书店”。这一方面可能是研究目的的不同，另一方面也反映出我们关于地名的认识还不够明确。

本文在前人研究的基础上，进一步探讨了中文地名的概念，并基于以上三种手段的框架，给出了一种综合前景信息和背景信息的基本地名识别方法。

2 中文地名构成及相关问题

2.1 中文地名的构成

本文所说的中文地名是指由汉字表示的中国地名及外国地名。从信息处理的角度出发，我们把中文地名定义为基本地名和复合地名构成的二级体系。

基本地名是地名的最小成词单位，对应于人脑中存储地名的最小单位；它是人们对具有特定方位、地域范围的地理实体赋予的专有名称[1]。作为地名的原子类型，基本地名满足指称性、非类指性（专门性）、词汇性、开放性等命名实体特征并具有指位性的功能特征。

典型的基本地名由“命名成分+通名”构成，命名成分是所指的标志符，不可缺省，如“江苏省”的“江苏”，“佛罗里达州”的“佛罗里达”；通名标识了所指单位的大小级别或类别，当命名成分已另有所指或为单字时常不可缺省，如“江苏路”中的“路”，“蓟县”的“县”。

基本地名通过合理组合形成复合地名。这里“合理”的意思是组合后形成的新地名有且只有一个所指，如“江苏省南京市”。复合地名是一个意义单位，相邻基本地名是否存在单向的领属关系是能否组合为一个复合地名的关键。因此，让计算机正确地识别、分析和理解复合地名有赖于基本地名的识别和基本地名之间关系的识别。

2.2 相关问题

中文地名识别的相关问题主要有两个，一是基本地名等命名实体的重名问题；二是地名与其它命名实体，特别是机构名的区分问题。

以非类指性代替专门性可以更明确地对基本地名等命名实体的重名现象作出解释。给实体命名的目的是将特定实体与其他实体区分开来，通常需要给定一个唯一的名字。然而，对应于多个所指的基本地名符号在词汇层面通常仍然被感知为所指唯一的命名实体。我们认为，命名实体是人类认知中的非类化的概念，与之对立的普通名词则是范畴化的产物。人们并没有从上海的“中山路”和南京的“中山路”抽象出“中山路”的共同特征，却从不同的树中抽象出“树”所共有的特征；从命名过程来看，“中山路”是各自直接命名，而“树”则是先进行范畴化的过程，再命名。所以，命名实体每次只能指称一个元素，而普通名词则可以指称一个集合。但是，人们一旦从基本地名中找到范畴化的动因，基本地名也可以像普通名词转化，如“唐人街”正处于这样的转化过程中。

地名和机构名的区分是地名识别需要解决的一个理论问题。通常地名和机构名都可以有相应的地理实体，因此地名和机构名都具有一定的指位功能。这使得人们在辨识这两种命名实体时容

[1] 这里借用了地名学关于地名的定义。

易产生混淆。1998 年《人民日报》标注语料（以下简称 RM9801）便存在着这样的标注错误和不一致问题。标注错误的例子，如将所有的“饭店”实体都标注为地名：

（1）本报/r 讯/Ng 1 月/t 1 2 日/t ，/w 数十/m 位/q 首都/n 记者/n 兴致勃勃/i 地/u 参加/v 了/u 由/p [中国/ns 大/a 饭店/n]ns //w [国贸/nz 饭店/n]ns 组织/v 的/u 以/p 体育/n 锻炼/vn 为/v 特色/n 的/u “/w ’ /w 9 8 /m 新闻界/n 新年/t 联谊会/n ”/w 。/w

（2）生于/v 1 9 0 0 年/t 的/u 董/nr 竹君/nr 是/v 洋车夫/n 的/u 女儿/n ，/w 曾/d 入/v 青楼/n 卖唱/v ，/w 曾/d 为/v 督军/n 夫人/n ，/w 也/d 曾/d 是/v [上海/ns 锦江/nz 饭店/n]ns 第一/m 个/q 老板/n ……/w

由地理实体组织联谊会或者充当地理实体的老板显然是不合理的。标注错误往往也与标注不一致共存，如“剧院”：

（3）音乐会/n 由/p [世纪/n 剧院/n]ns 、/w [北京/ns 音乐厅/n]ns 主办/v ，/w [世纪/n 演出/vn 公司/n]nt 承办/v ，/w [太平洋/ns 保险/n 公司/n 北京/ns 分公司/n]nt 协办/v 。/w

（4）第六/m 天/q ，/w 参加/v [杭州/ns 胜利/vn 剧院/n]nt 千/m 人/n 报告会/n ；/w

（5）1 月/t 1 7 日/t ，/w 这/r 台/q 音乐会/n 将/d 移/v 至/v [北京/ns 世纪/n 剧院/n]ns 再次/d 演出/v 。/w

一方面，（3）中主办方、承办方与协办方的实体类型不一致；另一方面，（3）、（4）、（5）关于“剧院”实体类型的标注也不一致。类似问题还存在于“博物馆”、“酒店”等实体类型的标注中。这种标注错误及其与标注不一致共存的现象反映出地名和机构名缺乏一个明确的区分尺度。

从基本地名的定义看来，地名是对地理实体的命名，显然机构名是对机构实体的命名，因此区分地理实体和机构实体以及命名与实体的关系是区分地名和机构名的关键。一部分机构可能不存在对应的地理实体，如“中国共产党”，所对应的机构名认知上不会和地名发生混淆；一部分机构有对应的地理实体，如“阿迪达斯公司”，“中国农业银行”，但“公司”，“银行”并不直接指示地理实体的存在，而被认知为机构，因此它们显然是对机构而非相应地理实体的命名；一部分机构名同时也用于相应地理实体的命名，并且其名称即明确指示了地理实体的存在，如“古南都饭店”，既可以指称“古南都”饭店的建筑范围，也可以指称“古南都饭店有限公司”，诸如“商场”，“酒店”，“书店”，“剧院”，“图书馆”等都属于这一类型。我们认为，对地名机构名兼类的消歧应该在语料中动态进行，地名仅仅具有指位性，而机构名同时具有指位功能和充当行为主体的功能，因此判别的依据是在句中是否或隐含地充当了行为主体，如上例，“剧院”在（3）中充当了行为主体，作机构名，而在（4）（5）仅仅作为事件发生的地点，作地名。

3 统计和规则相结合的基本地名识别

本文的目标是在分词文本上标注出中文基本地名。需要考虑两种不同的情形：1. 基本地名处于分词碎片中；2. 基本地名处于词语组合中。表 1 给出了两者在小规模抽样语料中的用例分布情况：

表 1 基本地名构成类型的用例分布

情形 / 指标	1	2
频次（例）	1299	121
比例（%）	91.48	8.52

根据上文设计的框架和基本地名的两种不同情形，基本地名识别主要包括三个过程（见图1）：常见地名匹配，分词碎分析，词语组合扩展。表 1 统计数据表明，处于分词碎片的基本地名占据了绝大部分比例，分词碎片分析是识别的重点。组合规则用来识别基本地名分布的第 2 种情形。整个识别流程如下：

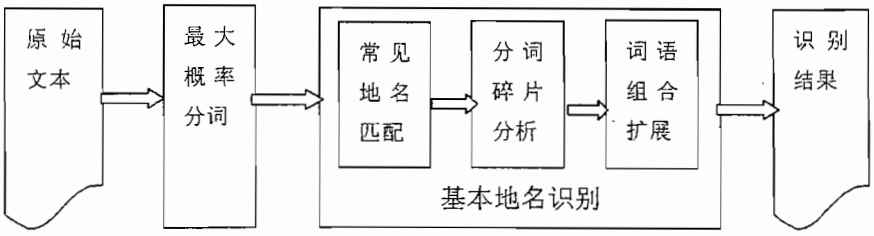

图 1 中文基本地名识别流程图

3.1 常见地名匹配

常见地名的匹配是人脑识别地名的一个重要方式。可以假设大脑存储了常见的基本地名以及与人的自身经历相关的某些地名。对计算机而言，常见意味着在语料中高频出现。为了获得更大的普适性，本文搜集了 2005 年全国县级以上的行政区划单位（澳门只有省级单位），共 2448 个基本地名及其删略通名的形式，以及 393 个常见国名、国都名组成常见地名表。

常见地名匹配在分词文本上进行，使用最大长度匹配的方法，为防止普通词被匹配为地名，我们对常见地名表与普通词表进行了比对排歧，删除了其中与普通词同形的词条。

3.2 分词碎片分析

陈小荷（1999）提出了从分词碎片中识别未登录词的一揽子解决方案；主要思想是从可知的单字词概率估计未登录词的概率，并使用最大概率分词法完成碎片的第二遍分词；其估计未登录词概率的方式是以非词单字串的概率乘以串频。

在此基础上，我们有两点考虑：1. 单字在分词碎片中是否成词基于一元语法很难确定，但是如果考虑二元的组合形式，很多可以构成合理的单字词句法组合，这种合法的二元组合在一定程度上可以降解分词碎片的分析歧义；2. 地名作为一种具体的未登录词类型，可以提出有效的概率估值，使其与单字词概率在同一个概率空间中竞争。

3.2.1 更小分词碎片的获取

将分词碎片中任意邻接的两个单字组合在一起，有四种关系类型：单字词的句法组合；单字词的非法组合；命名实体内部的单字组合；单字词与命名实体用字的组合。其中第一和第三种关系类型受到某种语言规律的制约，可以用于命名实体识别的排歧。

单字词的句法组合形式是一个有限的集合。一个单字词的二元句法组合可以缩减分词碎片的长度或将一个碎片切分为两个更小的碎片。RM9801 中包含了 32672 种单字词组合，其中绝大

部分是非法组合。本文从中选取了 798 对句法组合，剔除其中与地名词典中地名子串重合的条目等，组成二元句法组合库，用于获取更小的分词碎片。如在下面的例子中：

（1）山 高 沟 深 ， 土地 瘠薄

（2)法 国 总理 若斯潘 去年 年底 在 写 给 上 普 罗 旺 斯 阿 尔 卑 斯 省 社会党 议员 比昂古 的 信 中 说

例（1）在组合库中匹配上“山高”和“沟深”，从而避免了将“山高沟深”作为单字串传入基于概率的地名识别模块中；例（2）在组合库中匹配了“写给”和“信中”，只将“法国”，“上普罗旺斯阿尔卑斯省”传给下一个模块，从而减少了歧义。

3.2.2 基于概率竞争的基本地名识别

将地名识别问题简单地看作单字词与地名的概率竞争，即可将基本地名识别的问题还原为最大概率分词问题。这样处理的不足是忽略了单字序列中与地名竞争的可能还有其他的专有名词，如人名；好处可以同时利用地名的概率信息和单字词的概率信息。关键的问题在于如何估计基本地名的概率，使它与单字词的概率在同一个概率空间中形成竞争，其计算公式如下：

$$P(addr) = P_h(hz) * \prod P_m(mz) * P_t(tz) * P(ns) * SQRT(Min(freq, len))$$

其中，$P(addr)$ 表示某一个基本地名的概率估值，$P_h(hz)$ 表示该地名首字的出现概率，$P_m(mz)$ 表示该地名中间用字的出现概率，$P_t(tz)$ 表示该地名尾字的出现概率，$P_h(hz)$，$P_m(mz)$，$P_t(tz)$ 的计算考虑了地名在语料中的实际使用频次；P(ns)表示基本地名在语料中的出现概率，在 RM9801 中，P(ns)的值约为 0.025；freq 表示候选地名串的出现频次，len 表示候选地名串的长度，Min(freq,len)表示取其中的较低值，SQRT 表示取平方根，Min 函数是为了保证尽可能匹配较长的，并且频次较高的地名。

考虑到上下文信息，我们使用上下文词与 addr 的近似共现概率对 P(addr)进行修正：

$$P(addr) = P(addr) * (1 + \lambda_1 P(w_l \mid ns) + \lambda_2 P(w_r \mid ns))$$

其中，$P(w_l \mid ns)$ 表示已知地名出现的情况下，其左部出现某词的概率；$P(w_r \mid ns)$ 表示已知地名出现的情况下，其右部出现某词的概率。

单字词的概率 $P_w(z)$ 使用该单字作为词的出现频次除以语料总词次来计算，如果 $P_w(z)$ 和 Max（$P_h(hz)$，$P_m(mz)$，$P_t(tz)$）都很高，使用它的转移概率：

$$P_w(z_i) = \lambda_3 P_w(z_i) + \lambda_4 P_w(z_{i+1} \mid z_i)$$

其中，$\lambda_3 + \lambda_4 = 1$。基于概率竞争的基本地名识别将分词碎片中的每个串都作为候选地名，将每个单字都看作候选单字词，对分词碎片进行最大概率分词，从中选出一个概率最大的切分串，其中长度大于 1 的切分单位就是识别出来的基本地名。算法如下：

1. 获取分词碎片
2. 获取碎片中所有的候选地名和候选单字词
   a) 对每一个候选地名，计算 P(addr)，存入词表
   b) 对每一个候选单字词，计算 $P_w(z_i)$，存入词表
3. 用动态规划法寻找最优的分词路径
4. 将最优分词路径中长度大于 2 的分词单位标注为地名
5. 将分词路径中“ns+单字通名”的形式归并为地名
6. 反馈分词结果至所在句子，清空词表

3.3 基于词语组合的基本地名扩展

一部分基本地名存在于词语组合中，受“命名成分+通名”构成模式的限制，这部分基本地名在组合时也表现出一定的规律。

1. 当地名和通名邻接出现时，地名往往是命名成分，如“朝鲜（半岛）”；
2. 当地名和通名近距离顺序共现但不相邻时，地名常常只是限定成分而非命名成分，如“北京（炎黄艺术馆）”；
3. 出现通名时，一些特征词，如“在”、“的”等往往指示了基本地名的左边界。
4. 命名成分、通名可以是一个词，如“上海（动物园）”、“（中央）大街”，也可以是一个短语，如“航空知识（展览馆）”、“（首都）国际机场”等。

针对这些特点，制定识别算法如下：

1. 顺序处理文本中每一个单词 W_i 或相邻的两个单词 $W_{i-1}W_i$
   1) 匹配通名表，如果匹配失败，i++，转 1
   2) 判断 W_{i-1} 是不是基本地名，如果是，j=i-1，转 4）
   3) 在 W_{i-1} 向前两个词的范围内搜索
      a) 如果是禁用词，转 1
      b) 如果是基本地名 W_{j-1}，转 4）
      c) 如果是外边界特征词 W_{j-1}，转 4）
   4) 将 $W_j...W_i$ 标识为基本地名，转 1
2. 输出结果

4 实验结果及分析

本文的实验以生语料为基础，经过一个最大概率分词模块，使得核心模块在分词文本上进行，分词模块对人名作了前端处理。处于碎片中的基本地名是本文识别的重点；相应地，实验分为两个部分，实验 1 屏蔽了常见地名匹配的模块和基本地名扩展模块，对统计模块的识别效果进行测试，测试对象是碎片中的基本地名；实验 2 测试整个基本地名识别系统性能。

实验 1 以 RM9801 作为训练语料，并分别在 RM9801 和 RM9802 上进行了封闭测试和开放测试。当 λ_1=0.7，λ_2=0.3，λ_3=0.5，λ_4=0.5 时，实验结果如下：

表 2 统计模块实验结果

字段 测试类型	识别数（例）	正确数（例）	实有数（例）	正确率（%）	召回率（%）	调和平均值（%）
封闭测试	27778	24550	27814	88.38	88.26	88.32
开放测试	27577	24024	27742	87.11	86.59	86.85

实验 2 从 RM9801 和 RM9802 中分别抽取了 2000 个句子进行测试。实验结果如下：

表 3 中文基本地名抽样实验结果

字段 测试类型	识别数（例）	正确数（例）	实有数（例）	正确率（%）	召回率（%）	调和平均值（%）
封闭测试	1418	1277	1420	90.06	89.93	89.99
开放测试	1360	1199	1373	88.16	87.32	87.74

总的看来，实验所出现的错误主要来自三个方面。一是某些高频单字词在概率竞争中仍然容易胜出，如实验 1 中“大同”常常分析为“大 同”，这也造成不同的上下文中同一个地名的识别结果可能有所不同；二是由于数据稀疏问题造成漏识和误识，如“迪戈加西亚岛”；三是在词语组合扩展模块中，未出现左边界特征词的基本地名常难以召回。实验 2 中常见地名匹配模块和基于单字词二元句法组合的碎片分析方法提供了一些易维护的识别方法及消歧策略，召回了第一类错误中部分误识的地名，并提前过滤了第二类错误中部分可能误识的情况。

5 结语

本文探讨了基本地名的概念及相关问题，并针对其分布特点，使用常见地名匹配、概率竞争和组合扩展的方法初步识别了中文基本地名，包括中国地名和中译外国地名。其中基于概率竞争的方法较多地考虑地名内部构成特征，以及背景单字词的出现概率，较好地解决了地名与单字词的关系问题。但是初步实验也表明，部分上下文环境中，高频单字词用字出现在地名中时仍然可能被识别为单字词，也就是说，在具备一定的构成概率的前提下，基本地名相对于单字词背景又似乎具有识别的优先性。这种高频单字词的干扰问题是这一方法需要进一步解决的主要问题。本文进一步改进的方向有：

1．使用线性插值的办法将基于语料库的和基于词表的地名概率估计结合起来。

2．从识别结果中提取精确的地名，进行二次匹配及纠错。

3．探讨高频单字词干扰问题的解决办法。

参 考 文 献

[1] Tan Hongye, Zheng Jiaheng. Research on Method of Automatic Recognition of Chinese Place Name Based on Transformation. Journal of Software,2001.12(11): 1608~1613.

[2] 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,1999.31(3):103~109.

[3] 黄德根,岳广玲,等..基于统计的中文地名识别.中文信息学报,2003.17(2):36~41.

[4] 刘开瑛.中文文本自动分词和标注.北京:商务印书馆,2000.

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

[6] 谭红叶,郑家恒,等.中国地名自动识别系统的设计与实现.计算机工程,2002.28(8):128~129.

[7] 王际桐..地名学概论.北京:中国社会出版社,1993.