中国地名的自动辨识*

沈达阳 孙茂松 黄昌宁
清华大学计算机科学与技术系

摘要: 中国地名辨识对汉语自动分词研究具有一定意义。本文提出了一种在中文文本中自动辨识中国地名的算法。我们从新华通讯社新闻语料库中随机抽取了350个含中国地名的句子作为测试样本。实验表明，精确率达到81.1%，召回率达到95.0%。

Identifying Chinese Place Names in Unrestricted Text

Shen Dayang, Sun Maosong, Huang Changning
Dept. of Computer Science and Technology
Tsinghua Univ., Beijing 100084, P.R.C.

ABSTRACT: The processing of Chinese place names is significant to the approach of Chinese word segmentation. This paper presents an algorithm for automatically identifying this sort of proper nouns in Chinese texts. The testing sample, involoving 350 sentences each of which contains at least one Chinese place name, is extracted at random from the Xinhua News Corpus. The preliminary experiment shows that the precision and recall of this algorithm reaches 81.1% and 95.0% respectiveiy.

1. 引言

在研究汉语自动分词系统的过程中，我们发现，如果文本中存在未被辨识的中国地名，将会导致相当严重的分词错误。已有的大多数汉语自动分词系统对中国地名的处理，往往是通过在词典中穷举地名来实现的。这对面向真实文本的分词系统来说，显然存在问题：首先，地名，或都市州府，或市井乡村，或名川大泽，或穷山僻壤，理论上可以穷举，实际上不可能，而且，同一地名，在真实文本中可能以不同的形式出现（如："新康等乡"，"大濠诸岛"），更增加了变化的可能性；其次，即使能够穷举出来，地名数量必然极其庞大，都放入分词词典中，一则，加重系统的资源负担，降低运行效率，二则，对切分精度会有影响（如句子"由于山区生活困难"，本来没有切分歧义。但"于山"是一冷僻地名，若词典收录了，将会使"由于山区"变成链长为2的交集歧义）。因此，有必要研究中国地名的特点与规律，在真实文本中对中国地名实现自动辨识。

2. 中国地名辨识当用资源

为了系统研究中国地名的特点与规律，我们建立了中国地名库ＣＰＣ。ＣＰＣ含１７６３７个地名，包括省、自治区、直辖市，各省（区）的市、县、区，重要集镇及山脉，河流、湖泊、峡谷、海湾、港湾、岛屿、半岛及岬角，地形区，关隘、山口，交通，水利，矿区，革命纪念地和名胜古迹等。本文统计数据均系根据ＣＰＣ得到。

2.1. 中国地名的特点

*国家自然科学基金重点项目资助，合同号: 69433010

（1）关于地名用字

一方面，中国地名用字比较自由、分散，共用汉字2595个。地名用字的分布略图见图1:

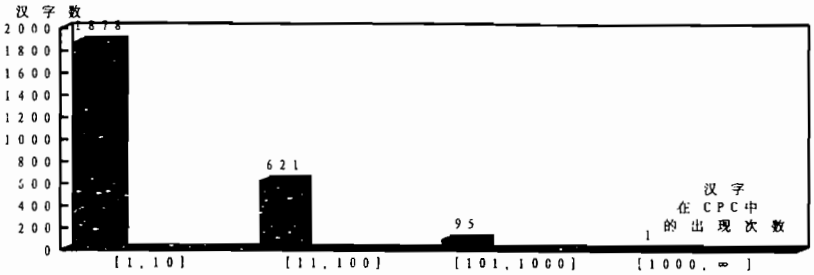

图1. 地名用字频率分布图

可见，出现次数在[1，10]之间的地名用字占绝大多数。

另一方面，地名用字又具有相对集中的覆盖能力，见图2:

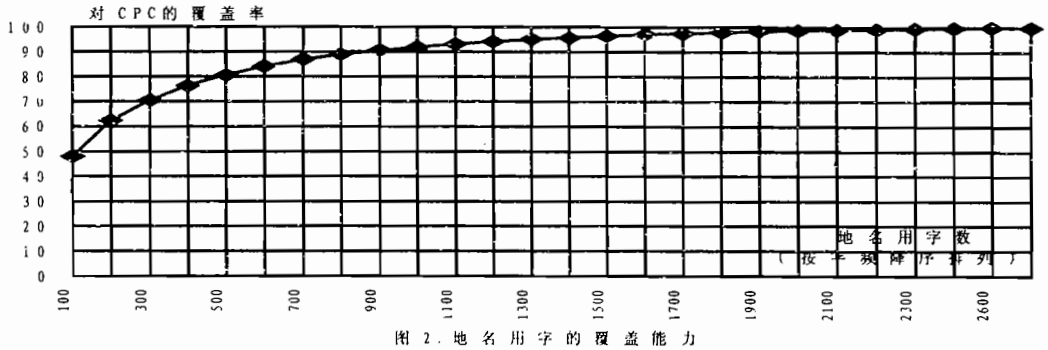

图2. 地名用字的覆盖能力

图2显示，字频最高的前100个字对CPC的覆盖率约50%，前900个字即达90%左右，换言之，剩下的1695个字的覆盖率不及10%。

（2）地名长度没有一定的限制，最短的如“津，京”，最长的如“双江拉祜族佤族布朗族傣族自治县”。

（3）可作单字词的汉字在地名中经常出现，如“西直门”，“马家塔”中的每个单字均为高频单字词。

（4）地名中可含有多字词，如“龙王/洞/山”，“黄/果树/瀑布”，“兵书/宝剑/峡”，“红领巾/路”等。

（5）部分常用后缀对地名有一定的提示作用（如“自治县”，“水库”）。

（6）地名周围缺乏丰富、有效的启发信息（较人名要弱，人名周围经常出现称谓及“说，认为，告诉”之类动词）。

其中，（1）对辨识地名既有不利的一面（对应图1），也有有利的一面（对应图2）。

（2），（3），（4），（6）均增加了辨识的难度。（5）则是我们所乐见的。

2.2. 中国地名的规律

（1）地名首、中、尾部用字规律

从ＣＰＣ可得到中国地名用字表ＰＣＬ（｜ＰＣＬ｜＝２５９５）。则对于任一汉字ｃ∈ＰＣＬ，均存在一个属性向量:

$$F_c=\begin{bmatrix} f_s(c) \\ f_m(c) \\ f_e(c) \end{bmatrix}$$

其中$f_s(c),f_m(c),f_e(c)$表示ｃ出现在ＣＰＣ中地名首，中，尾部的频率（≥0）。

进一步，我们有:

ＰＣＬｓ＝｛ｃ｜ｃ∈ＰＣＬ且$f_s(c)\neq 0$｝

ＰＣＬｍ＝｛ｃ｜ｃ∈ＰＣＬ且$f_m(c)\neq 0$｝

ＰＣＬｅ＝｛ｃ｜ｃ∈ＰＣＬ且$f_e(c)\neq 0$｝

ＰＣＬｓ、ＰＣＬｍ、ＰＣＬｅ对ＣＰＣ地名首中尾部的覆盖率曲线Ｐｓ，Ｐｍ，Ｐｅ如图3所示:

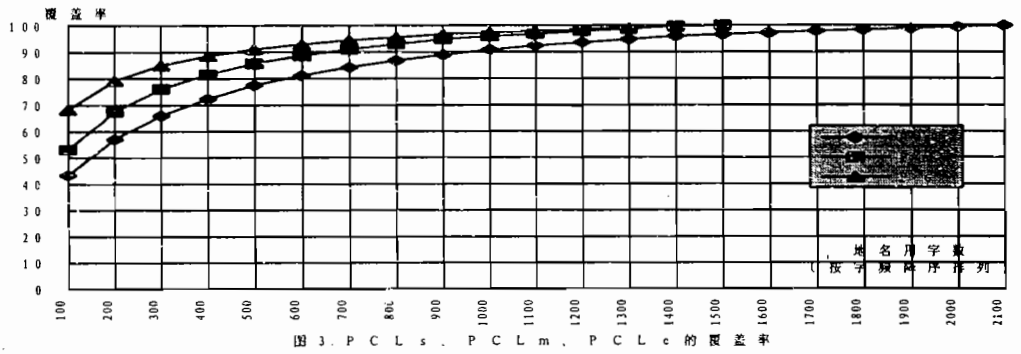

图３．ＰＣＬｓ、ＰＣＬｍ、ＰＣＬｅ的覆盖率

显然，尾字的覆盖能力最强，中字次之，首字最弱。

（２）地名首、中、尾部用词规律

前文已述，地名中可能包含多字词。此类多字词对地名辨识有特殊影响，有必要将它们和词典中其它词区分开来。从ＣＰＣ中可得到中国地名用词表ＰＷＬ（｜ＰＷＬ｜＝916）。同样地，对于任一词ｗ属于ＰＷＬ，对应有属性向量:

$$F_w=\begin{bmatrix} f_s(w) \\ f_m(w) \\ f_e(w) \end{bmatrix}$$

且从ＰＷＬ可产生ＰＷＬｓ，ＰＷＬｍ及ＰＷＬｅ。

（３）地名的构词规律

某些中国地名遵循一定的构词方式，如:

单字中国人名姓氏＋“家”＋后缀	丁家山 于家房 万家坝 习家口 王家场 王家庄
数词＋“里”＋后缀	五里坪 五里店 六里河 十一里铺

（４）地名上下文规律

初步的研究发现，地名的上下文启发因素不多，具备一定指示意义的，上文限于某些表处所的介词（如“在，于，到”），下文限于方位词(如“左边”、“东部”）及机构名词（如“县委”，“省委”）。

3. 中国地名的辨识算法

3.1. 最大匹配分词

首先采用自左向右最大匹配算法，把句子分割成为单字，单字词及多字词。

例句1	新康乡赵江村的周建国谈起冬修就眉飞色舞
切分结果	新/康/乡/赵/江/村/的/周/建国/谈/起/冬/修/就/眉飞色舞/
例句2	记者到枝柳铁路上的八江乡采访
切分结果	记者/到/枝/柳/铁路/上/的/八/江/乡/采访/

3.2. 寻找地名激活字段

地名激活字段α就是触发地名辨识过程的字段，应该满足条件:

（a）任给词w∈α

如果w占据α的起始位置，则有w∈PWLs；

如果w占据α的终止位置，则有w∈PWLe；

否则，有w∈PWLm。

（b）任给单字或单字词c∈α，则有c∈PCL。

在研究中，我们发现PWLs和PWLm中的词，并没有严格的区别，而PWLe中的词规律性较强，因此条件（a）减弱为:

任给词w∈α，

如果w占据α的终止位置，则有w∈PWLe

否则，有w∈PWLs或w∈PWLm;

例句1的α	新康乡赵江村的周,谈起冬修就
例句2的α	到枝柳铁路,上的八江乡（∵“铁路”∈PWLe）

3.3. 利用属性矩阵进行筛选

此时，对地名激活字段α中单字、单字词及多字词不予区别，将α视作连续汉字串$c1..ci..cn$，则得到α的属性矩阵$F\alpha = F_{c1}...F_{ci}...F_{cn}$。我们要继续在α中寻找子串β，$\beta = F_{cp}...F_{ci}...F_{cq}$ (p≥1, p<q≤n)，使得β的属性矩阵$F\beta$满足:

(a) $f_s(cp)>0$

(b) $f_m(ci)>0$ (p<i<q)

(c) $f_e(cq)>0$

α可能裂变为多个β。

为减少数据不足或地名变体等的影响，在实施此种处理之前，对属性向量Fci (1≤i≤n)进行了调整:

如果其中任两个分量>0,且第三个分量=0，则令该分量=1。

例句1	新	康	乡	赵	江	村	的	周	建	国	谈	起	冬	修	就	眉	飞	色	舞
α的属性矩阵	185	24	2	29	74	3	0	42			0	3	1	5	0				
	29	2	3	3	7i	19	0	2			0	3	2	0	0				
	25	16	28	5	296	206	1	7			1	3	0	1	1				

例句2	记	者	到	枝	柳	铁	路	上	的	八	江	乡	采	访
α的属性矩阵			0	4	24	34	11	85	0	73	74	2		
			0	5	15	156	27	3	0	14	71	3		
			0	5	11	1	178	21	1	6	296	28		

结果为:

例句1的β	新康乡赵江村的，起冬修就 (∵f_m(“的”)=0, f_s(“谈”)≠0)
例句2的β	枝柳铁路,上的，八江乡 (∵f_m(“的”)=0)

3.4. 利用频级进行筛选

β中可能包含一些不属于地名的常用单字词，应尽可能将它们滤掉。我们把ＰＣＬ地名用字频率和单字词词频各划分为K级，并据之在β中寻找一子串γ = $u...c_i...c_v$。γ满足:

（a）如果Cu是单字词，则有level(f_s(Cu)) ≥ level(f_w(Cu))

（b）如果Cv是单字词，则有level(f_e(Cv)) ≥ level(f_w(Cv))

γ的搜索过程实际上是以这两个条件为限制，由β两侧向中心逐层逼近。

β	新	康	乡	赵	江	村	的
单字词频	10671	0	645	635	1025	973	80000
单字词频级数	3	0	1	1	2	2	3
地名用字频率级数	3	2	1	2	3	1	0
	2	1	1	1	3	2	0
	2	2	2	1	3	3	1
搜索过程	(1) level(fs(新)) = level(fw(新)),左逼近结束 (2) level(fe(的))<level(fw(的)),右移 (3) level(fe(村))>level(fw(村)),右逼近结束 ∴γ = “新康乡赵江村”						

对例句1及例句2，此过程结束后，有

例句1的γ	新康乡赵江村，冬修
例句2的γ	枝柳铁路，八江乡

3.5. 利用二元语法进行筛选

观察两类γ:

类1. 类2.

γ	冬	修	γ	九	县
单字词频	452	649	单字词频	5441	2455
单字词频级数	1	1	单字词频级数	2	2
地名用字频率级数	1	1	地名用字频率级数	3	1
	1	1		1	1
	1	1		0	3

显然，类1、类2无法靠步骤3.4正确处理。我们进一步地采用地名用字在ＣＰＣ中的同现频率（即二元语法），以解决此种错误。即增加限制

如果Ci,Ci+1均为词，且:

（a）level(f_w(Ci)) ≠ 单字词频最低级 且

level(f_w(Ci+1)) ≠单字词频最低级

或（b）level(f_x(Ci))=地名用字频率最低级 且

level(f_{x+1}(Ci+1))=地名用字频率最低级

（根据Ci，Ci+1在γ中位置x, x+1相应地取s, m或e）

则CiCi+1在CPC中的同现频率必须满足: $f_{cici+1} \geqslant \theta$（θ为阈值），才认为CiCi+1属于地名中的一部分。

条件（a）（b）的约束，使得地名的二元语法（即bigram）的大小不会膨胀。

于是，例句1，例句2中的地名为:

例句1的地名	新康乡赵江村
例句2的地名	枝柳铁路, 八江乡

3.6. 对连续地名的处理

以上各步最终给出的结果可能包含多个连续地名（如“新康乡赵江村”）。最好能将它们分割成更加基本的单位，而不是作为一个整体“囫囵吞枣”地混为一谈。为此，采用了两种方法:

（1） 规则判别法

在新闻语料中，以“某某省某某市”、“某某县某某乡某某村”等形式出现的地名很多，根据“省市区县镇乡村”之类的关键字，容易用规则予以分割。

（2） 权值和判别:

对于地名 $P = c_1 \ldots c_i \ldots c_n$，定义其权值和:

$$FSUM(c_1..c_n) = f_{s(c1)} + \sum_{i=2}^{n-1} f_{m(ci)} + f_{e(cn)}$$

则有利用权值和之判别:

如果P 中，存在一个或多个i（1<i<n-1），使得:

$$FSUM(c_1..c_i) + FSUM(c_{i+1}..c_n) > FSUM(c_1 c_2..c_n),$$

则 $j = \arg\max_i (FSUM(c_1..c_i) + FSUM(c_{i+1}..c_n))$ 为P的分割点。

例如:

P属性矩阵	正 出 现 一 个 深 圳 沙 头 角 似 的 商 业 贸 易 区 12 1 87 17 5 1 2 65 56 19 1 2 30 118 24
结果	FSUM(“深圳沙头角”) = 12 + 2 + 65 + 56 + 24 = 159 FSUM(“深圳”)+FSUM(“沙头角”)=(12+2)+(87+56+24)=181 FSUM(“深圳沙”)+FSUM(“头角”)=(12+2 + 30)+(17+24)=85 ∴ “深圳沙头角” -> “深圳”、“沙头角”

3.7. 其它

地名的构词规律及上下文规律以规则的形式嵌入系统，限于篇幅，不再展开讨论。

4. 实验结果

该中国地名辨识实验系统在ＰＣ４８６/３３上实现，用C++编程。

除了词典之外，系统用于地名辨识的数据，包括:

地名用字表ＰＣＬ及频率信息	１９ｋ
地名用词表ＰＷＬ及频率信息	４ｋ
地名的二元语法	２ｋ
Σ	２５ｋ

这比原来１２７ｋ的ＣＰＣ要小得多。

我们随机抽取了部分１９９１年的新华社新闻语料，以检验我们的系统:

实验语料	15.5k的新闻语料(未训练过,共350个句子)
中国地名	420
报出地名	492
边界完全正确的地名	368
边界不完全正确地名	31
完全没有发现的地名	21
召回率	(368+31)/420 = 95.0%
精确率	(368+31)/492 = 81.1%

部分实验结果:

神府东胜煤田马家塔露天煤矿正式建成投产	神府、东胜、 马家塔
邵阳县委领导到红石乡梅溪村搞农村社会主义教育试点	邵阳、红石乡、梅溪村
访问了榆次市郭家堡乡几个村的个体户	榆次市、郭家堡乡
但地处大别山腹地的安徽省金寨县却春意盎然	大别山、安徽省、金寨县
赞皇县许亭乡治理西沟小流域	赞皇县、许亭乡、西沟
最近在楚雄州武定县首次发现了世界上最古老的太阳历法	楚雄州、武定县
帮助广东东莞、惠州等市县栽立水泥电杆	广东、东莞、惠州
大连金州区广筹教育资金见成效	大连、金州区
这是武清县南蔡村农民写的	武清县、南蔡村
地处伏牛山区的灵宝县寺河山乡	伏牛,灵宝县,寺河

从实验结果看,系统的召回率和精确率基本上满足处理真实文本的需要。说明本文所提出的方法有一定的合理性。与穷举地名的策略相比，我们的方法在时间，空间上都比较节省，尤为重要的是，提供了对任意中国地名的猜测能力。我们认为，随着训练集的扩大和地名辨识算法的调整，以及把地名辨识同其它未登录词辨识，继而同歧义切分处理结合起来，地名辨识的召回率和精确率还可以进一步提高。

参考文献

〔1〕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

〔2〕孙茂松，黄昌宁，高海燕，方捷，“中文姓名的自动辨识”，《中文信息学报》，第9卷，第2期，1995

〔3〕《中国地名词典》，上海辞书出版社，1990
语言学院出版社，1993