中文单字国名简称的自动识别

李 斌 方 芳
(南京师范大学文学院,南京 210097)
E-mail: gothere@126.com

摘 要 单字地名简称识别是中文专名识别的重要组成部分,其中单字国名简称又占到了78.43%。但就现有的词性标注系统来看,还不能很好地对其进行识别。文章提出了一个基于规则的识别方法,以分词文本为底本,通过两次扫描,第一次扫描建立基于语篇的临时国名词表,第二次扫描利用上下文特征词等评分机制,从分词碎片中召回单字国名简称。在封闭测试和开放测试中,都取得了较好的实验结果,以人工分词文本为输入底本,调和平均值分别达到了96.33%和94.96%。

关键词 单字地名 单字国名简称 评分机制 临时词表

文章编号 1002-8331(2006)28-0167-03 **文献标识码** A **中图分类号** TP391

Single Chinese Character Country Name Recognition

LI Bin FANG Fang
(Nanjing Normal University, Nanjing 210097)

Abatract: Single Character(shortened form) Country Name(SCCN), is a kind of shorted form of a Country name, which composed of one Chinese Character, such as "英"(ying1, England) and "美"(mei3, America). The SCCN recognition is a part of the task of Chinese Named Entity Recognition(NER). This paper investigates the usage of SCCNs in real text, then builds a modal for recognition based on rules, using temporary wordlist and contextual information as main resources. Using the manual segmented text as input, F-score of our method achieved 96.33% in the close test and 94.96% in the open test.

Keywords: Single Character Location Name, Single Character Country Name, evaluation mechanism, temporary wordlist

中文专名识别是中文信息处理的重要内容。单字简称略语识别作为其中的一个组成部分,有着特殊的作用。如果单字简称略语可以分小类标注出来,对于专名的嵌套处理和提高中外人名识别的精确率[1]等方面都有益处。由于现有的词性标注系统在这方面还缺乏深入的研究,识别的精确率和召回率都不是很高。本文以单字简称略语中最主要部分的单字国名简称识别为突破口,在文本调查的基础上,来发现规则、建立资源、设计算法,主要利用了局部的语篇信息、上下文特征词以及单字国名简称用字本身的信息等启发信息,取得了较好的实验结果。

1 单字国名简称和相关研究

1.1 术语界定

"简称",一般是指专名的缩略形式,是专名的一种,包括人名、地名、机构名等等。如,"美利坚合众国"的简称,是"美"和"美国"。按照是否是单字的标准,"美国"是多字简称,而"美"就是单字简称。本文着重讨论的就是"美"、"俄"之类的单字国名简称(包括"欧"、"亚"、"太"、"非"等经常和国名连用的大的洲名、地区名)的自动识别问题。

相对于简称,"略语"是指专名以外的词或短语的缩略形式。如,"一国两制"、"通胀"等。笔者对人民日报1998年1月语料(北京大学计算语言学研究所人工分词、标注)进行了调查[2]。统计发现,单字国名简称分别占单字地名简称的78.43%和所有简称略语的26.39%。所以,本文首先选择单字国名简称作为识别对象。

1.2 相关研究

在中文地名识别方面,前人做的工作比较多,但是关于单字地名简称识别,目前只有Zhu et al(2003)做过较好的研究[3]。该文认为,由于缺乏深入的研究,单字专名在专名识别中成为一个主要的错误来源。该文使用了规模为1M的测试语料,比较了三个词性标注系统对单字地名简称的识别结果。分别是微软的MSWS系统、北京语言大学的LCWS系统和微软的PBWS系统。测试结果显示,这三个系统的调和平均值都没有超过45%。在此基础上,该文使用了改进的通道信噪模型、最大熵模型和向量空间模型。实验数据表明,改进的通道信噪模型对于单字地名简称有着较好的识别能力,调和平均值达到了81.01%。笔者估计,由于单字国名简称在单字地名简称中所占的比例最高,该模型对于单字国名简称的识别也应该能够达到80%左右的调和平均值。

国内的各词性标注系统,主要使用隐马尔科夫模型,把部分单字简称直接收入静态词表。虽然没有专门针对单字地名简

基金项目:南京师范大学211资助项目(编号:1240702504)

作者简介:李斌(1981-),男,博士生,主要研究方向:计算语言学。方芳(1981-),女,硕士,主要研究方向:计算语言学。

[1] 本文中,精确率、召回率以及调和平均值(β=1),都是按照现在通常公认的计算方法,这里不再赘述,参见黄昌宁等(2003)。

[2] 在统计的过程中,由于和笔者的研究目的无关,没有统计人名简称的数据。

称的识别手段，仍然可以取得较好的召回率。笔者测试了中科院计算所开发的 ICTCLAS 系统，召回了 82.26%的单字国名简称，但是精确率较低，单字简称略语的总精确率只有 53.38%（参见表 7）。导致精确率低的原因，主要是把人名中的单字国名用字和其他单字简称略语用字标注了出来。如，“穆/j 巴/j 拉/j 克/j 总统”。[3]

2 基于多种评分机制的识别方法

2.1 基本思路

本文所做的工作，就是在切词底本的基础上，从分词碎片中召回单字国名简称。全世界的国家实体名称总共只有 200 多个，包括近几十年来政权更迭的新旧国名，如‘苏（苏联）’、‘白（白俄罗斯）’等。而一些国家的简称是相同的，如，“阿曼”、“阿联酋”的简称都是‘阿’。所以，单字国名简称用字仅有 100 多个。利用这一现象，笔者考虑直接锁定文本中的简称单字国名用字，根据它们的不同类型（参见表 1），采用多种机制进行判断评分（加减分运算），最后用一个总的阈值来确定是否为单字国名。

表 1 单字国名简称类型分析

类型		实例	有无边界问题	频次	所占比例
单用		俄外长	无	997	36.65%
连用	普通连用	中日邦交	有	1 561	57.39%
	“、”连用	中、日、韩三国	无	162	5.96%
合计				2 720	100%

2.2 上下文评分机制

单字国名简称的上下文一般会有一些特殊词语，可以作为启发信息来进行识别。如，“中俄总理会晤”中的‘总理’和‘会晤’。

2.2.1 文本调查

笔者人工标注和校对了 1998 年 1 月人民日报语料中所有单字国名简称，使用检索程序，自动提取出单字国名简称的上下文，左右分别为三个词。然后，人工选择和提取出上下文的特征词。如表 2，可以提取出‘友好’、‘合作’、‘关系’三个词。

表 2 自动提取单字国名简称上下文样例

类型	左三	左二	左一	单字国名简称	右一	右二	右三
单用	一贯	重视	对	华	友好	关系	,
连用	俄罗斯	政府	对	俄中	友好	合作	关系

2.2.2 文本调查后建立的评分机制

在文本调查中发现，单字国名简称在单独出现（单用）和连续出现（连用）的时候，有着一些不同的上下文。上表中的‘友好’，在语料里一般只作为连续出现的单字国名简称的下文（如，“中美友好”），或者是出现在‘对 友好’的短语框架中。因此考虑分别提取出单用和连用的上下文词表作为重要的评分依据。另外，上文特征词、下文特征词不一定会同时出现，而短语框架又是一种要求上文和下文同现的特征词，所以分别建立了六个上下文词典（参见表 3）。

在这些词典中，很多词条会有一些交叉。如，“重申”一词，收入了单用上文、单用下文、连用上文、连用下文四个词典中。这些上下文词典在存储和使用的时候，并没有给出具体的频率和位置信息。如上文提到的‘友好’，只是宽泛地作为单用框架和连用下文收入词典。在识别时，只要‘友好’在连用右文的三个词以内出现，就给加分。短语框架的前后两个词也是如此，如“发展 对 美 的 友好 合作 关系”，“美”也可以获得加分。同时，“发展”又是单用上文词，“美”又可以获得一次加分。这样，得到的六个词典总词条数只有 941 条，较好地避免了统计模型的数据存储量大又难以避免数据稀疏的问题。

表 3 单字简称上下文词典

上下文词典	类型	实例	词条数/条
单字单用	上文	访（美），留（美）	60
	下文	（俄）外长，（美）军方	408
	框架	留（法）学生，来（华）投资	214
单字连用	上文	远销（欧美），出口（日韩）	21
	下文	（中日）邦交，（中美）贸易	209
	框架	庆祝（中美）建交	29
合计			941

2.3 国名词表评分机制

仅靠上下文特征词，既无法保证召回所有的单字国名简称，也无法保证识别的正确率。举个简单的例子，“加拿大 国会 中也有人 要求”[4]。如果仅依靠上下文特征词，“中也”会被加分。所以，笔者进一步从单字国名简称本身入手来解决问题，主要使用了三种策略。

2.3.1 基于单个语篇的临时词表评分机制

本文的一个重要的方法就是建立基于语篇的临时国名词表。分词文本和待标注文本通常都是由多个语篇所构成的。[5] 而在单个语篇当中，单字国名简称的出现，往往伴随着原称的出现（之前或之后），尤其是一些不经常出现或很少出现的国名，如‘贝（贝宁）’、‘格（格鲁吉亚）’（参见表 4）。在单个语篇的范围内，利用国名全称和非单字简称的词表，在分词底本的非切词碎片中寻找国名，收入临时词表。建立临时词表的目的，主要是为了给中低频简称国名用字进行评分，来提高识别的精确率。最高频的几个简称国名，如，“英”、“美”、“中”等，语篇中原称往往不出现，所以对最常见的简称国名，加减分要控制在较小的幅度内。接着看上文的例子，“加拿大 国会 中 也 有人 要求”。在一个语篇中，如果出现了国名“也门”，则‘也’便进入临时词表，获得一定的加分。在处理到‘也’的时候，作为国名简称识别出来的可能性就大些。相反地，如果没有出现‘也门’，就会被大幅度减分，从而基本可以确定该处不是国名简称。

表 4 单字简称与原称的共现

位置分布		频次	所占比例
与原称共现	原称前	491	18.05%
	原称后	1 845	67.83%
不与原称共现		384	14.12%
合计		2 720	100%

临时词表是一个重要的评分资源，从表 5 大致可以看出其重要作用。

2.3.2 高低频国名的评分机制

根据在观察语料中出现的频次，把所有单字国名简称（130 个）进行分类，对前 30 个高频的和后面 100 个低频的（很多低

[3] 由于该系统可以识别多种单字简称略语，但不分小类，都只标注出‘/j’的简称标记，所以单字国名的精确率不好统计。总精确率是指，该系统对所有单字简称略语的识别的整体精确率。

[4] 加波浪线表示该字是单字国名用字，加下划线表示该词是上下文特征词典中的词。

表 5 [6]封闭测试结果

分词底本	精确率	召回率	调和平均值
每个字符都切开	46.46%	58.42%	51.76%
最大匹配法切词	78.16%	89.82%	83.59%
人工切词(无临时词表机制)	53.77%	96.21%	68.99%
人工切词[7]	97.26%	95.40%	96.33%

频的在观察语料中没有出现)分别处理。高频国名会得到加分,低频国名相应减分。目的是为了补偿没有出现原称或上下文特征词的高频国名,惩罚即使出现了原称的低频国名。如,单独出现的“美”,就会得到一定的加分,而单独出现的“也”(“也”本身是一个单字词)就会被减分。但是,这只是一个平衡性的评分机制,作用有限。如,“中也”连用,一个被加分,一个被减分,还是难以解决。而且,如果在一个语篇中已经出现了国名“也门”,这种策略就成了一种失误,必须和其他机制结合使用。

2.3.3 特殊单字简称国名评分机制

“也”、“不”、“越”、“所”等低频国名用字,是高频的单字词,经常伴随特征上下文出现。如,“正如 基本法 所保证 的 一样”,“不是 越 小 越 经济”。一般情况下,这些字在临时词表的作用下会被惩罚掉。但是,一旦语篇中有相应的非单字国名出现,又会得到加分。所以综合起来考虑,给这几个词比较多的减分。只有在上下文特征词出现得比较多,总得分较高,通过了最后的阈值,才确定为简称。至此,“中也”的问题,基本得到了解决。

另外,“俄”、“柬”、“缅”,作为单字词,基本上就是国名简称,应该给予加分。

“以”、“新”、“日”、“美”、“德”、“中”等常用单字词,暂时还没有更好的解决办法。顾及精确率的需要,采取了小幅减分的策略。

2.4 特殊现象的挖掘利用

2.4.1 “中”、“华”的互补分布

在文本调查中发现,“中”、“华”同样是“中国”的单字简称,却有着截然不同的分布特征。“中”是观察语料中出现频率最高的单字简称国名。由于它还是方位词,经常出现在切词碎片里。“中”都是在连用时出现,如“中美”等。“华”恰恰相反,都使在单用时出现,如“访华”。两个词正好形成了互补分布。利用这一点,就可以避免误识出文本中大量出现的“中”单独使用的情况,如,“国家队 的 出访 中”。

2.4.2 “欧美”、“亚太”和“、”的特殊处理和利用

这两个词出现的比较多,而且基本上已经词化了,所以考虑把这两个词作为特殊模式,暂时处理为简称连用。

由表 1 可以看出,“英、美、法、德等国”中的“、”是非常值得利用的标志,也应该给予加分。

2.5 评分分值和阈值的选取

经过各个评分策略以后,文本中的单个国名和连用国名块都会得到一个总分。设定一个阈值作为过滤器,超过该阈值即确定为单字国名简称,带上特殊标记输出。阈值的调整和各评分机制的选值密切相关,是一个动态调整的过程,表 3 和表 5 给出了简化的加减分数据作为示例。阈值选定的依据是最优的封闭测试结果。

3 算法设计

根据上文的各种识别策略,下面给出基本算法和流程。

(1)读入切词底本。

(2.1)寻找单个语篇的边界,如果左右边界重合,则表示到达文本末端,转(4)。

(2.2)进行第一遍扫描,在单个语篇内使用“国家非单字简称词表”搜索建立单字国名简称临时词表。

(3)在该语篇内进行第二遍扫描,使用“单字国名简称词表”锁定候选国名。

(3.1)进行“亚太”、“欧美”、“中”、“华”的预处理,“亚太”、“欧美”直接输出。

(3.2)根据单用和连用情况,分别进行运算。

(3.2.1)根据上下文特征词典,进行评分。

(3.2.2)根据临时词表,进行评分。

(3.2.3)根据高频、低频、常用国名、干扰词等词表分别进行评分。

(3.2.4)核算总得分,通过阈值,带标记输出。否则,直接按原文输出。转(2.1)。

(4)结束,退出。

4 实验结果及分析

4.1 封闭测试

4.1.1 测试结果

笔者在 Microsoft Visual C++6.0 的平台上,使用控制台程序,实现了上述算法。并设计了专门的文本比对程序来测试精确率、召回率以及调和平均值。测试语料为 1998 年 1 月的人民日报语料,3 147 个语篇,约183 万字,共有单字国名简称 2 720 条。

表 6 开放测试结果

分词底本	自动标注	精确率	召回率	调和平均值
ICTCLAS 切分	ICTCLAS	53.38%*	82.26%	*
ICTCLAS 切分	笔者的系统	79.82%	84.54%	82.12%
人工切词[7]	笔者的系统	96.17%	93.76%	94.96%

表格说明:*ICTCLAS 切分并标注的精确率和调和平均值难以统计,参见脚注 3。

4.1.2 错误分析

通过对人工分词底本进行标注产生的错误进行分析。归纳出以下三类典型错误:

(1)上文中提到的“以”、“美”、“新”、“日”等字造成的误标。如:

“巴方仍将以通过谈判解决巴以争端”,“俄法都以相当严峻的态度”,“景美游客多”,“以尽量减少政府的财政赤字”,“以英爱新协定”,“问题日受重视”。

这一类很难解决,从上下文 3 个词的范围来看,很多词串本身就是词汇意义造成的歧义短语。

(2)人名中含有的国名简称用字造成的误标。需要结合人名识别来进行处理。如:

“中国的韩晶娜今天以 0 2 负于韩国的李宙泫”,“华大使

(下转 176 页)

[5] 待处理文本在形式上,往往带有明显的或者是隐性的语篇分割信息和标志,可以用来建立基于语篇的临时国名词表。即使没有形式上的分割信息,也可以大致地按照上下文的长度做一些技术处理。

[6] 表 4 中的数据,是在同样的评分条件下得到的,如果对各项评分值进行调整,前几个底本也可能会得到更好的结果,这里只是为了分析分词底本和临时词表的作用。

[7] 人工切词底本是指分词过的文本。笔者把词性标注过的熟语料,通过简单的程序,转换得到分词文本。

[8] “韩晶娜”的误标,是由于人工分词底本中,姓和名一律切开了。如果合并起来,识别的效果会更好。

90%以上，其余的也在 88%以上，这说明我们选择的特征合适，总体方法是有效的。

5 结论与展望

本文中，我们分析了足球视频的语义结构，为每个镜头赋予了语义，使之成为“语义镜头”，一定的语义镜头的序列表达了一定的“语义事件”。这种结构便于用户根据语义对足球视频进行浏览和查询。我们应用 HMMs 来分析足球视频的语义结构，实验证明，选取合适的“合成特征”可以达到相当好的效果。在未来，我们要确定更丰富、更高层次的镜头语义，如裁判员镜头等；还要细化“语义事件”，如进球、任意球等。另外，我们还将把 HMM 方法应用到其他视频的语义结构分析当中。这将要求我们选择更多的合成特征，并引入“多模态”的方法，进行视频中的音频分析和视频中的文字识别。（收稿日期：2006 年 5 月）

参考文献

1.A Ekin, A M Tekalp.Generic event detection in sports video using cinematic features[C].In: Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video(EVENT 2003), Madison, Wisconsin, USA, 2003: 17~24

2.X Yu, C Xu, H Leong et al.Trajectory-Based Ball Detection and Tracking with Applications to Semantic Analysis of Broadcast Soccer Video[C].In: Proc of ACM Multimedia, 2003: 11~20

3.K Wan, J Lim, C Xu et al.Real-time camera field-view tracking in soccer video[C].In: Proc of International Conference on Acoustics Speech and Signal Processing, 2003; 3: 185~188

4.L Xie, P Xu, Shih-Fu Chang et al.Structure analysis of soccer video with domain knowledge and hidden Markov models[J].PRL, 2004; 25(7): 767~775

5.Liu X M, Chen T.Shot Boundary Detection Using Temporal Statistics Modeling[C].In: IEEE Intl Conf on Acoustics, Speech and Signal Processing, ICASSP2002, Orlando, FL, U S, 2002-05

6.Ricardo Lenardi, Pierangelo Migliorati, Maria Prandini.Semantic Indexing of Soccer Audio-Visual Sequence: A multimodal approach based on controlled Markov chains[J].IEEE Trans on Circuits & System for Video Technology, 2004; 5: 634~643

7.M Barnard, J-M Odobez, S Bengio.Multi-modal audio-visual event recognition for football analysis[C].In: IEEE Workshop on Neural Networks for Signal Processing, NNSP, 2003: 469~478

8.Baillie M, Jose J M.Audio-based Event Detection for Sports Video[C].In: CIVR2003, 2003-07: 300~310

9.L R Rabiner.A tutorial on Hidden Markov Models and selected applications in speech recognition[C].In: Proc of the IEEE, 1989; 77(2): 257~286

（上接169 页）

夫妇乐得合不拢嘴”。

（3）专名中含有的国名简称用字，但由于启发信息不足而造成的漏标。需要结合其他专名识别来处理。如：

“澳大利亚澳纽银行”、“坦赞铁路线”。

前两类所占的比例高达 90%左右。另外还有一类，是由特殊的格式造成的。这一类是偶发的，但也难以处理。如：

“泰人助泰”、“东为印、西为巴”。

4.2 开放测试

测试语料为 1994 年人民日报的部分语料，内容为国内、国际政治新闻，规模为 19 万词次，31.5 万字次，168 个语篇，共有单字国名简称 931 条。

作为对照，笔者采用了在 SIGHAN 的比赛中有上佳表现的 ICTCLAS 系统。其既可以提供分词文本，又可以自动标注词性，而且对人民日报新闻语料的处理能力也相对较强。在 ICTCLAS 分词标注的结果中，单字简称略语的总精确率只有 53.38%。所以，在相同的分词底本上，笔者的标注结果还是比较好的。如果采用人工校对过的分词底本，则接近于封闭测试的成绩。开放测试中的错误类型与封闭测试中的基本相同，不再展开。

5 结语

本文的识别工作可以看作是一种基于规则的识别方法，运用了统计的基本思想和语言学知识。在文本调查的基础上，发现可以利用的规则，进而建立资源、设计算法。在识别过程中利用局部的语篇信息、上下文特征词和单字简称国名用字本身的信息，结合考察在文本调查中发现的“中”和“华”的互补分布现象等等。这些信息和特殊现象的发掘利用，对于单字地名简称的识别工作起到了非常重要的作用。由此看来，在分词和词性标注的过程中，对不同的具体问题可以采取不同的处理手段，以具体问题的特殊为出发点，发挥各种方法的长处。正如孙茂松（1995）所言，“鉴于任何单一手段都不能包打天下，算法研究的重点应转移到多种分词知识与分析策略的集成上来”。本文所做的只是专名识别的一个小模块，笔者将进一步使用此方法，对文本中单字简称的其他类型进行识别（如，省市县等地名）。一方面验证这种方法在相似的识别工作中是否可行，一方面进一步考虑如何与分词过程相结合。

致谢：本文在写作过程中，得到了黄昌宁教授、陈小荷导师的悉心指导和师兄曲维光、许超的热心帮助，在此一并表示衷心地感谢。（收稿日期：2005 年 12 月）

参考文献

1.Xiaodan Zhu, Mu Li, Jianfeng Gao et al.Single Character Chinese Named Entity Recognition[C].In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, ACL, 2003

2.孙茂松，邹嘉彦.汉语自动分词研究中的若干理论问题[J].语言文字应用，1995;（4）：40~46

3.陈小荷.自动分词未登录词问题的一揽子解决方案[J].语言文字应用，1999;（3）：103~109

4.陈小荷.现代汉语自动分析——Visual C++实现[M].北京：北京语言文化大学出版社，2000

5.黄昌宁，高剑锋，李沐.对自动分词的反思[C].见：孙茂松，陈群秀 ed.语言计算与基于内容的文本处理，北京：清华大学出版社，2003：26~37

6.张华平，刘群.基于 N-最短路径方法的中文词语粗分模型[J].中文信息学报，2002;16(5)：1~7

7.张小衡，王玲玲.中文机构名称的识别与分析[J].中文信息学报，1997;11(4)：21~32