文章编号:1671-8860(2010)08-0930-06　　**文献标志码**:A

中文文本的地名解析方法研究

唐旭日[1]　陈小荷[1]　张雪英[2]

(1　南京师范大学文学院,南京市宁海路 122 号,210097)
(2　南京师范大学虚拟地理环境教育部重点实验室,南京市文苑路 1 号, 210046)

摘　要:讨论了中文文本的地名解析流程,提出基于条件随机场和篇章地名关系的地名识别方法、基于局部模糊匹配的地名标准化方法以及基于认知显著度的地理编码方法,并构建了地名解析原型系统。实验显示,该系统可以获得较为满意的精确率、召回率和 F-1 值,同时讨论了地名词典的完备性、地名识别精度以及地名语义歧义消除等影响地名解析性能的主要因素。

关键词:地名解析;地名识别;地理编码;地名匹配

中图法分类号:P208

在 21 世纪,地理信息系统将走出科研院所和政府机关,成为全社会人人具备的信息服务工具[1]。通过文本的地名解析(toponym resolution),可以搭建起地理信息系统和自然语言理解之间的桥梁,对于推动地理信息系统的大众化服务,如医疗、公共安全、计算机科学、商业选址、城市规划等领域的应用具有十分重要的作用[2-4],并在地理信息的自然语言查询结构的构建[5]、信息抽取、问答系统等应用研究中具有重要价值。地名解析是对文本中的地名进行识别和语义判断,并将其映射到地理坐标的过程[6]。近几年来,地名解析及应用研究发展迅速,在国内外取得了一些比较有代表性的研究成果[6]。

基于文本形式的地名解析包含两个关键步骤:地名分析和地理编码。地名分析又可分为地名识别和地名标准化两个子任务。地名识别所面临的最大问题是歧义问题,包括词的边界歧义和词语指向歧义。中文边界歧义是汉语书写习惯造成的。汉语文本书写时没有边界,故地名识别需要确定词的边界。词语指向歧义是指同一词串具有不同的语义指向,如"北京"可以指向地名(如"北京 城")或人名(如"李 北京")。地名识别方法可分为三类。其一是机器学习方法,通过大规模语料库获取地名识别的统计模型,如层叠条件随机场模型[7]、层叠隐马尔科夫模型[8]、最大熵模型[9]、最大间隔马尔科夫模型[10]、支持向量机[11]等。其二是规则方法与统计方法的结合[12,13]。其三是利用地名词典和地理编码信息进行地名分析[14]。与机器学习方法相比较,依赖于地名词典和最大匹配的方法在地名识别精度方面稍有逊色,受词典规模和切分歧义的影响较大。

地名标准化是将同一地名的不同拼写形式和不同用字进行规范化处理,从而保证在地理编码过程中能够找到相对应的地理实体[4,15-17]。在西方语言中,地名标准化主要是拼写、格式、字符集等的不一致问题。而中文地名的标准化所面临的主要问题是地名脱落,即地名在使用过程中通名部分可以省略。如"北京市"省略为"北京","西藏藏族自治区"省略为"西藏"。地名标准化需要将各种脱落后的地名映射到地名词典相应的地理实体中去。在西方语言中,地名标准化应用的技术包括词典查找与规则匹配、基于隐马尔可夫的机器学习技术[15]等。

地理编码将地理对象在确定的参考系中按一定的规则赋予惟一和可识别的代码,建立地理对象(地名所指示的地理实体)与坐标系统的映射,从而将地理位置信息转换成可以被用于 GIS 的地理坐标[18]。地名指向歧义是地理编码所面临

收稿日期:2010-06-15。

项目来源:国家 863 计划资助项目(2007AA12Z221);国家自然科学基金资助项目(40971231,60773173);国家社科基金资助项目(07BYY050)。

的主要问题。地名指向歧义的消解一般采用基于地名词典和语言学知识的启发式搜索方法。文献[6]对已使用的各种启发式搜索方法进行了比较和综合,并将其归纳为 3 大类、16 小类。文献[19]侧重于多语言文本的地理编码过程,所采用的启发式方法包括地名人名歧义区分、地理实体重要性、文本中讨论的国家、地理编码停用词、直接上下文语境、最小地理空间距离等。文献[20]在地理编码阶段主要利用了一个连续地名字串中的空间包含关系以及其他的汉语表达空间概念的模式。在启发式搜索中,可利用的信息类型种类很多,缺乏统一的概念,如何综合考虑各种影响因素以及如何设置各种因素的权重是需研究的问题。本文系统讨论了中文文本的地名解析流程,并给出了地名分析与地理编码的实现算法及实验评估结果。

1　地名解析系统流程

图 1 给出了地名解析系统的各个功能模块及流程。系统分 5 个任务模块:地名识别模型构建、地名词典构建、基于篇章的地名识别、局部模糊地理实体匹配和基于认知显著度的文本地理编码。在地名概念决策后,CRFs 地名识别模型通过语料库训练获取,动态地名关系数据库通过地名关系抽取获得,地名词典通过对国家基础地理数据和 GIS 服务商提供的数据进行结构化处理获得。文本输入后,需进行篇章化预处理,将输入中文文本以篇章为单位进行划分。基于篇章的地名识别和局部模糊地名匹配,构成地名分析任务阶段。基于篇章的地名识别以篇章为单位,综合利用 CRFs 地名识别模型、动态地名关系数据库和地

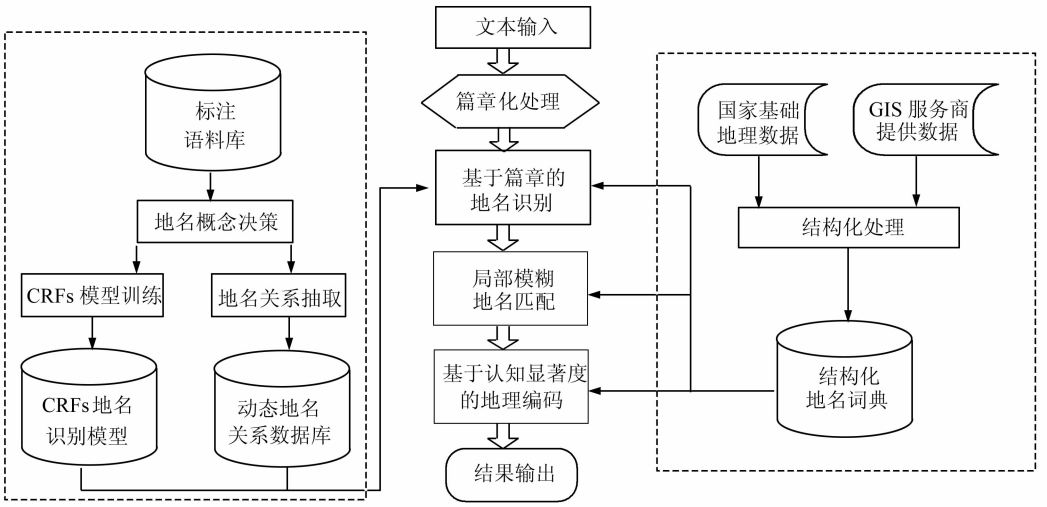

图 1　地名解析系统流程

Fig. 1　Flow Chart of Toponym Resolution

名词典信息。局部模糊地理实体匹配主要完成地名标准化任务。地理编码采用了基于认知显著度的地理编码方法,依赖于结构化的地名词典。

2　地名分析

地名识别的基本思路是首先通过 CRFs 模型识别出篇章中的地名,然后利用篇章中地名之间存在的静态、动态关系对识别结果进行优化处理。为此,需首先明确地名的概念,建立 CRFs 地名识别模型、动态地名关系库和静态地名关系库,然后设计地名识别过程。

2.1　地名概念决策、CRFs 地名分析模型与动态地名关系库

地理信息系统中,不同空间操作对地名的概念定义并不一样。因此,地名识别需首先明确地名的概念,即地名概念决策。语料库中,地名的概念是通过标注实例化建立起来的,通过地名字串与标注符号之间的映射确定地名概念的外延,两者映射的建立需考虑标注类型和标注结构两方面。而在标注结构方面,复杂地名结构包含多个词语,而简单地名则只由一个词语构成。针对不同需求,可以建立不同的映射关系。

地名识别可以转化为序列标识问题。条件随机场在处理序列标记问题方面表现优秀[7,12],提供了识别歧义消除的良好机器学习模型。理论上,条件随机场模型是在给定输入观察值条件下计算输出标记的条件概率的无向图模型 $G=(V,E)$,其中,V 为节点集合,E 为边集合[22]。给定输入串 X,可以获得标识串 Y。

本文从北大语料1～5月份语料中抽取包含简单地名(ns)标记的句子组成训练语料,以CRF^{++}(0.49)作为模型训练平台,采用L-BFGS参数评估算法[9],六位标注集(S,B,B1,B2,I,E)以字为单位进行标示。考虑到机构名(nt)与地名之间的关联,在训练时,同时考虑了简单机构名(nt)的标注。运用CRF^{++}平台进行训练后,即获得CRFs地名识别模型。

动态地名关系库记录地名在同一篇章中的同现关系。动态地理关系体现为地名对,如(北京,天安门)、(南京,江苏)等。地名之间存在的固有语义关系和人们在不同空间的活动是不同地名在同一篇章中共同出现的潜在原因,构成了语篇中地名之间的相互关联。本文同样以北大语料1～5月份语料为基础训练语料,以篇章为单位,抽取篇章同现地名对,作为动态地名关系库数据。实验中共抽取322 012个动态地名关系对组成动态地名关系库。

2.2 地名词典

地名词典的构建包括国家基础地理数据的收集和整理、GIS服务商数据的获取以及数据的结构化。在国家基础地理数据收集方面,采用《中华人民共和国地名大辞典》,经纬度信息则从Google Earth中下载。结构化处理主要体现为地理实体的结构化表示方法。地理实体具有四级行政区划属性(如表1所示),各个属性对应一个地名实体,由此形成地理实体之间的层级关系。这种层级关系是结构化地名词典的主要特征,在地理编码中具有非常重要的应用。数据经过结构化处理后,构成地名词典,其中共包含地理实体162 344个,分为64个子类,包括一、二、三、四级行政区划、山脉、河流、峡谷、盆地、车站、农场、学校等不同的地理要素类型。地名分析成功的地名可以通过地理坐标信息在GIS系统获得空间定位。

表1 结构化地理词典

Tab.1 Structured Gazetteer

ID	名称	一级行政区划(Adm1)	二级行政区划(Adm2)	三级行政区划(Adm3)	四级行政区划(Adm4)
1730	城关镇	北京市	北京	通县	城关镇
3951	城关镇	天津市	天津	宝坻县	城关镇

2.3 基于篇章的地名分析

本文以篇章为单位进行地名识别。输入文本不经过分词预处理,而直接利用CRFs地名识别模型,同时进行地名识别和分词,获得识别地名集合S,然后利用动态地名关系数据库中的动态地名关系和地名词典中确定的静态地名关系(即地名的结构化关联)进行后处理。后处理对CRFs地名识别结果进行扫描,对于任一未标注为地名的字串w,如果下列条件成立,则将其补充标注为地名:① w具有地名的结构特征;② w与S中的某一个地名l存在动态地名关系;③ w与S中的某一个地名l存在静态地名关系。通过分析w的内部结构,可以判断其是否具有地名的结构特征;查询动态地名关系库,可以确定w和l之间是否存在基于篇章的动态地名关系;通过对地名词典的查询,可以确定w和l之间是否存在静态地名关系。静态地名关系可以是地理实体之间的行政隶属关系,也可以是同时隶属于某一行政区划的关系。

2.4 局部模糊地名匹配

局部模糊地名匹配是地名标准化处理过程的关键环节。地理实体在汉语的使用中存在多种不同的地名表示形式,因此,需要将多种表示形式映射到特定地理实体。目前还未见完善的地理实体的使用名称知识库,因此无法通过检索直接获取地名实体的表示形式。然而汉语中地名的使用具有较强的规律性,一般而言,地理实体的名称一般由专名+通名构成,在使用中,通名部分往往可以脱落。据此,本文提出一种局部模糊匹配方法,通过对地名进行形式变换来获得地名的表示形式。

局部模糊匹配方法试图将待考查地名与地名词典中地理实体名称的可能变化形式进行匹配。给定篇章中出现的地名字符串W_1^n和地理实体I的名称字符串K_1^m,如果满足下列条件之一,则认为W_1^n是I的可能地名表示形式:

1) $W_1^n=K_1^m$(即W_1^n与K_1^m为相同字符串);

2) $W_1^n=K_1^n$,且K_{n+1}^m为地名通名(即W_1^n与K_1^m的前n个字符串相同,而K_1^m中后$m-n$个字符串为地名通名);

3) $W_1^{n-i}=K_1^{n-i}$,$W_{n-i+1}^n=K_{m-i+1}^m$,W_{n-i+1}^n为通名或空字符,K_{m-i+1}^m中包含"族"或者"自治"子串(即W_1^n与K_1^m的前$n-i$个字符串相同,W_1^n的后$n-i+1$个字符为空或构成地名通名,K_1^m的后$m-i+1$中包含"族"或者"自治"子串。)

3 基于认知显著度的地理编码

给定篇章中的一个地名,通过地名匹配可以获得该地名的可能指向地理实体集合。地理编码的目标是利用上下文信息进行歧义消除,将地名与所获取的地理实体集合中某一地理实体关联起

来,从而确定该地名在地理空间中的地理坐标。在实际操作中,地理实体集合所包含的元素数目可能会很大,从而导致地理编码需要搜索的空间也非常庞大。

本文采用基于认知显著度的地理编码算法。认知显著度是指心理体验中某一概念被激活的容易程度或可能性[23]。在语言使用中,一个词语可能与多个概念对应,但是这些概念被激活的可能性是不一样的,受概念自身的使用频率、词语与概念之间的关联紧密程度、词语使用的上下文语境等因素的影响。地名所指向的不同地理实体可以理解为地名的不同对应概念。可以直观地认为,语言使用中,地名与地理实体的关联程度是由该地理实体在具体语言环境中相对于该地名的认知显著度决定的,具有最大显著度的地理实体就是该地名在该上下文语境中所指向的地理实体。因此,地名 w 指向的地理实体 I' 的计算方法为:

$$I' = \arg\max_{I_i \in S_w} \mathrm{SL}(I_i, S_{\mathrm{context}(w)}) \tag{1}$$

式(2)表示从地名 w 可能指向的地理实体集合 S_w 中选择具有最大认知显著度的地理实体作为 I'。其中,context(w)为在地名 w 上下文中出现的所有其他地名的集合;$S_{\mathrm{context}(w)}$ 为对 context(w)中所有地名进行地名匹配后获取的地理实体集合;$\mathrm{SL}(I_i, S_{\mathrm{context}(w)})$ 为地理实体 I_i 的认知显著度。

地理实体的认知显著度与地名和地理实体名称之间的匹配程度、地理实体自身的类型以及该地名使用的上下文环境相关。本文对文献[24]提出的显著度计算方法进行了修改,考虑了匹配程度、类型显著度和上下文显著度三个因素:

$$\mathrm{SL}(I_i, S_{\mathrm{context}(w)}) = M(I_i, w) \times C(I_i) \times X(I_i, S_{\mathrm{context}(w)}) \tag{2}$$

其中,$M(I_i, w)$ 是匹配程度,由地理实体 I_i 的名称与地名 w 之间的匹配程度确定,其数值设定如下:

$$M(I_i, w) = \begin{cases} 1, & \text{如果 } I_i \text{ 的名称与 } w \text{ 之间完全匹配} \\ 0.8, & \text{如果 } I_i \text{ 的名称与 } w \text{ 之间不完全匹配} \end{cases} \tag{3}$$

$X(I_i, S_{\mathrm{context}(w)})$ 为根据地名 w 的上下文语境来确定地理实体 I_i 的显著度;$C(I_i)$ 为依据地名要素 I_i 的地理实体类型而确定的认知显著度,不同行政区划和其他地理要素类型的认知显著度不尽相同,本文设置见表 2。

表 2　不同要素类型的认知显著度

Tab. 2　Cognitive Salience by Geographic Feature Categories

地理实体类型	显著度
一级行政区,河流、岛屿、海湾、湖泊、火山、草原、山脉、山谷、峡谷	1
二级行政区	1/2
三级行政区	1/3
四级行政区	1/4

从语言使用单位的角度考虑,地名 w 的语境包括三种不同类型的单位:句子、段落和篇章,不同类型的语境单位在确定显著度时的权重不同,其中句子单位的近距离语境权重较大,段落语境权重次之,篇章语境权重最小。基于这一思想,对于地名 w,可以获取与其在同一句中同现的其他地名集合 L_s,与其在同一段落同现中的其他地名集合 L_p,以及在同一语篇中同现的其他地名集合 L_d,并通过地名匹配分别获取 L_s、L_p 以及 L_d 可能指向的地理实体集合 $S_{L_s} = \bigcup_{w_i \in L_s} S_{w_i}$,$S_{L_p} = \bigcup_{w_i \in L_p} S_{w_i}$ 和 $S_{L_d} = \bigcup_{w_i \in L_d} S_{w_i}$。$I_i$ 的语境显著度可通过下式进行计算:

$$X(I_i, S_{\mathrm{context}(w)}) = \alpha X'(I_i, S_{L_s}) \times \beta X'(I_i, S_{L_p}) \times \gamma X'(I_i, S_{L_d}) \tag{4}$$

其中,$X'(I_i, S_{s(w)})$、$X'(I_i, S_{p(w)})$、$X'(I_i, S_{d(w)})$ 分别表示以句子单位语境、段落单位语境和篇章单位语境计算地理实体 I_i 的语境显著度;α、β、γ 分别为不同语境单位的权重。在本文实验中,$\alpha=1$,$\beta=1/2$,$\gamma=1/3$。

对于某一语境单位 U(例化为 L_s、L_p 和 L_d),地理实体 I_i 的语境显著度 $X'(I_i, S_U)$ 是通过 I_i 的属性(即 I_i 的所属上级行政区划)在地理实体集合 S_U 中出现的频率确定的。S_U 中的地理实体可依据其类型划分为四个子集 S_U^{Adm1}、S_U^{Adm2}、S_U^{Adm3} 和 S_U^{Adm4}。而依据表 1,地理实体 I_i 具有 Adm1、Adm2、Adm3、Adm4 四个属性,如表 1 中序号为 3951 的地理实体的属性 Adm1、Adm2、Adm3 和 Adm4 的取值分别为天津市、天津、宝坻县和城关镇。对于地理实体 I_i 的属性 l(例化为 Adm1、Adm2、Adm3、Adm4),记为 $a_{I_i}^l$,可以获取其在 S_U 中对应子集 S_U^l 出现的频率 $P(a_{I_i}^l)$,这一频率即代表属性 l 在该语境单位 U 中的显著度。进而通过综合考察 I_i 的四个属性,获得 I_i 在该地名单位上下文中的认知显著度:

$$X'(I_i, S_U) = \prod_{l \in \{\mathrm{Adm1}, \mathrm{Adm2}, \mathrm{Adm3}, \mathrm{Adm4}\}} P(a_{I_i}^l) =$$

$$\prod_{l\in\{\mathrm{Adm1},\mathrm{Adm2},\mathrm{Adm3},\mathrm{Adm4}\}} F(a_{I_i}^l)/\sum_{I_j\in S_U^l}(I_j) \quad (5)$$

其中，$F(a_{I_i}^l)$为地理实体I_i的l属性$a_{I_i}^l$在上下文中出现的频次；$F(I_j)$为所有l类型地理实体出现的频次。

4 实验评估

地名解析可以看作是分类问题，即将一地名归属分为某一个地理实体在文本使用中的具体实例。因此，地名解析也能够采用一般的评价标准进行系统性能评测[6]，建立标注语料库，使用精确率、召回率和F-1值方法进行评测。地名解析的结果一般有三种类型：① 解析结果正确(T_c)，即系统标注的地理实体与评测语料一致；② 解析结果错误(T_e)，即系统标注的地理实体与标注语料不一致，导致错误的原因可能是地名分析错误，也可能是地理编码错误；③ 未解释(T_u)，即由于该地名或者没有识别，或者地名词典中缺乏该地名的对应信息，使得系统没能对该地名进行解析。系统的精确度P、召回率R和F-1值计算如下：

$$P = T_c/(T_c + T_e), R = T_c/T$$

$$F\text{-}1 = (2\times P\times R)/(P+R)$$

其中，T为测试语料中地名出现的词次。

为考察系统的实际使用性能，对系统进行开放测试。测试语料为从人民网和新浪网随机选择的包含中国地名的网页89个，在网页正文抽取后进行了人工校对，然后进行地名标注和地理实体标注，建立起系统评测语料。该评测语料约93 000字，其中地名出现2 364词次。经过地理实体标注后，地名词典中未收录地名426个，其他的1 938个地名共分别标注为428种地理实体。

表3给出了地名识别和地名解析的实验结果。如前所述，本文的地名概念仅局限于北大语料中标注为ns的简单地名。在地名识别方面，实验F-1值达到了95.70%。与本文在语料测试规模上相近是文献[9]所进行的地名识别实验。该文献采用最大熵识别模型，在开放测试最佳识别结果中，精确率为91.30%，召回率为93.13%，F-1值为92.19%。值得注意的是，本文中，地名概念仅局限于简单地名，而在文献[9]中，地名概念包括简单地名和复杂地名等类型，且不同文献所使用的训练语料和测试语料也不相同，因而本文的实验结果无法与文献[9]进行直接比较，也不能与文献[7-12]直接比较。即便如此，在开放测试中，F-1值能达到95.7%，说明基于篇章的地名识别效果是令人满意的。

表3 地名分析和地名解析实验结果

Tab.3 Experiment Results

	精确率/%	召回率/%	F-1/%	正确标识个数/T_c	错误标识个数/T_e	未标识个数/T_u
地名识别	96.31	95.11	95.70	2 248	91	116
地名解析	88.54	78.93	83.46	1 530	198	834

与地名识别相比较，地名解析的精度下降幅度较大，F-1值仅为83.46%。造成地名识别与地名解析之间差距的原因很多，具体而言，包含以下几个方面：

1) 地名词典的不健全。本文所建立的地名词典共包含162 344个不同的地理实体，但是在测试语料所出现的2 364个地名中，有17%的地名未能在库中找到对应的地理实体。在实验中，这些地名都属于未解释结果。因此，地名词典的不健全是影响地名解析的非常重要的因素。然而构建大规模地名词典并非易事，行政区域划分的变更等都会使地名词典的建立更为困难。例如，在本文建立的要素库中，由于行政区划的变更，“荆州”可能与“湖北省荆州地区”和“湖北省荆沙市荆州区”两个地理实体相关联。

2) 地名分析的错误。地名分析是地名解析的基础，高精度的地名分析对于地名解析至关重要。如果地名字串被错误识别，如“双石铺乡”、“克孜勒苏柯尔克孜”等被识别成“双/o 石铺乡/ns”，“克孜勒苏/ns 柯尔克孜/ns”，则无法触发正确的地名匹配，无法在地名中找得相对应的地理实体，因而无法获得正确的地名解析结果。

3) 歧义消解错误。一方面，不同地理实体类型之间存在歧义。由于历史文化的原因，汉语中行政区划名称与自然地理实体名称之间紧密关联，许多行政区划名称都源于自然地理实体，从而造成了两者区分的困难。如“西山”可以用来指一个行政区划，如“云南省昆明市西山区”，也可以用来指位于多个行政区划中的自然地理实体“西山”。另一方面，同一地理实体类型中不同地理实体存在歧义。如前文所提及的“鼓楼”、“城关”等都属于这种类型的歧义。本文所采取的基于显著度计算的歧义消解方法对这两类错误进行了一定程度的区分。

5 结 语

基于中文文本的地名解析研究，无论是在地

理信息系统研究领域,还是在中文信息处理研究领域,都尚处于初始阶段,然而其应用前景广阔。实验证明,本文所提出的中文地名解析方法具有较好的地名识别精度,能够较好地获取地名的地理实体指向,消除地理编码中的歧义问题。同时也发现,为进一步提高中文文本地名解析的精度,需要进一步完善地名词典,进一步提高地名识别精度,综合利用多种方法处理地理编码中的歧义问题。下一步的研究将集中在地名词典的维护、完善地名语义模糊性处理方法以及地名解析与地图服务的集成应用等。

参 考 文 献

[1] 李德仁. 论 21 世纪遥感与 GIS 的发展[J]. 武汉大学学报·信息科学版, 2003, 28(2): 127-132

[2] Hill L L. Georeferencing: the Geographic Associations of Information[M]. Cambridge: MIT Press, 2009

[3] 江洲,李琦. 地理编码(Geocoding) 的应用研究[J]. 地理与地理信息科学,2003, 19(3):22-25

[4] Goldberg D W, Wilson J P, Knoblock C A. From Text to Geographic Coordinates: The Current State of Geocoding[J]. URISA Journal, 2007, 19(1): 33-46

[5] 马林兵,龚健雅. 空间信息自然语言查询接口的研究与应用[J]. 武汉大学学报·信息科学版, 2003, 28(3):301-305

[6] Leidner J L. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names[D]. Edinburgh: University of Edinburgh, 2007

[7] Pouliquen B, Kimler M, Steinberger R, et al. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualization[C]. The 5th International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy, 2006

[8] 乐小虬. 非结构化网络空间信息智能搜索与服务研究[D]. 北京:中国科学院遥感应用研究所, 2006

[9] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报,2006,34(5):804-809

[10] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔科夫模型的中文命名实体识别[J]. 通信学报,2006,27(2): 87-94

[11] 钱晶,张杰,张涛. 基于最大熵的汉语人名地名识别方法研究[J]. 小型微型计算机系统,2006, 27(9): 1 761-1 765

[12] Li L, Ding Z, Huang D. Recognizing Location Names from Chinese Texts Based on Max-Margin Network[C]. International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2008

[13] 李丽双,黄德根,陈春荣,等. 基于支持向量机的中文文本中地名识别[J]. 大连理工大学学报,2007,47(3): 433-438

[14] 李丽双,黄德根,陈春荣,等. SVM 与规则相结合的中文地名自动识别[J]. 中文信息学报,2006,20(5): 51-57

[15] 向晓雯,史晓东,曾华琳. 一个统计与规则相结合的中文命名实体识别系统[J]. 计算机应用, 2005,25(10): 2 404-2 406

[16] 乐小虬,杨崇俊,刘冬林. 空间命名实体的识别[J]. 计算机工程,2005,31(20):49-53

[17] Christen P, Churches T, Willmore A. A Probabilistic Geocoding System Based on a National Address File, in Data Mining Theory, Methodology, Techniques, and Applications[M]. Berlin: Springer-Verlag,2004: 130-145

[18] Davis C A, Fonseca F T. Assessing the Certainty of Locations Produced by an Address Geocoding System[J]. GeoInformatica, 2007, 11(1): 103-129

[19] Zandbergen P A. A Comparison of Address Point, Parcel and Street Geocoding Techniques[J]. Computers, Environment and Urban Systems, 2008, 32(3): 214-232

[20] 王凌云,李琦,江洲. 国内地理编码数据库系统开发与研究[J]. 计算机工程与应用,2004,40(21): 167-168, 211

[21] 冯元勇,孙乐,张大鲲,等. 基于小规模尾字特征的中文命名实体识别研究[J]. 电子学报,2008,36(9): 1 833-1 838

[22] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. International Conference on Machine Learning, 2001

[23] Zhuo J S. Dramatized Discourse: The Mandarin Chinese Ba-Construction[M]. Philadelphia: John Benjamins Publishing Company, 2005

[24] Tang Xuri. Toponym Resolution in Discourse[C]. International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2008

第一作者简介:唐旭日,副教授,博士生,主要研究方向为中文信息处理和地理信息系统。

E-mail:xrtang@126.com

(下转第 982 页)