

# 中文地名识别与歧义消除

## ——以中国县级以上行政区划地名为例

杜 萍, 刘 勇

(兰州大学资源环境学院, 甘肃 兰州 730000)

**摘要:**介绍了本体、地理本体和地名本体的基本概念,探讨了在文本工程通用框架 GATE(General Architecture for Text Engineering)下,以自然语言处理为基础,借助地名本体,完成 Web 文本的中文地名识别与歧义消除的关键问题,设计并实现了一个原型系统。通过 geo/non-geo 和 geo/geo 地名歧义的消除,使得识别出来的地名与地球表面具体的地理位置相对应,进而为 Web 文本中的中国行政区划地名赋予地理坐标和地理语义。做了验证实验,并对实验结果进行了分析。

**关键词:**地名识别;歧义消除;地理解析;地理编码

**中图分类号:**P 208 **文献标志码:**A **文章编号:**1004-0323(2011)06-0868-06

**引用格式:**Du Ping, Liu Yong. Recognition and Disambiguation Chinese Toponym from Web Texts——Take the Names of Chinese Administrative Division above County for Example[J]. Remote Sensing Technology and Application, 2011, 26(6): 868-873. [杜萍, 刘勇. 中文地名识别与歧义消除——以中国县级以上行政区划地名为例[J]. 遥感技术与应用, 2011, 26(6): 868-873.]

## 1 引言

Web 文本蕴含丰富的地理信息。这些地理信息大多以自然语言的形式存在,简短并具有很强的描述性,但是不直观、不精确,加上地名歧义的存在、新旧地名的差异等原因,不仅无法满足人们对地理信息日益增长的查询需求,还严重地阻碍了地理信息检索等领域的发展。因此,研究如何从 Web 文本中将地理信息识别出来具有重要的意义。

地理命名实体是 Web 文本中主要的地理信息。地名和机构名是最常见的两种地理命名实体<sup>[1]</sup>,地名的数量往往多于机构名,而且很多机构名都包含地名。因此识别地名,并将它们转化为结构化的 GIS 数据是从 Web 文本中挖掘地理信息的基础和关键。

本文探讨了中文地名识别与消歧的关键问题,设计并实现了一个原型系统。系统在自然语言处理的基础上,借助构建的地名本体,通过地理解析和地理编码,将中国县级以上行政区划地名从 Web 文本中标识出来,并映射到地球表面能够用多边形或点等几何类型表达的某处空间,从而给它们分配地理

坐标及地理语义。该过程把地名这种以自然语言形式表达的间接的空间参照“翻译”成精确的地理坐标,使得识别结果不仅能够作为 GIS 的数据来源,还能应用于基于位置的服务等领域。

## 2 地名本体的构建

本体(Ontology)最早是一个哲学范畴,侧重于反映现实,是对客观存在的系统解释和说明。Studer 等<sup>[2]</sup>认为,本体是共享概念模型明确的形式化的规范说明。1995 年,Egenhofer 等<sup>[3]</sup>在对常识地理学(Naive Geography)进行研究的过程中,将本体论引入了地理信息科学,使其成为地理信息科学的一个新兴研究方向。地理本体作为地理空间领域的共享概念模型,提供地理概念及概念间的关系,并通过概念间的关系来描述概念的语义。地理本体不仅要表达一般的属性特征,还要表达极其重要的空间特征。

地名是人类对地理实体(包括地域)共同约定的名称,是人们在日常生活中进行空间定位的重要方式。根据地名的构成,可以分为两种类型:特称加通

称地名(如“广州市”,“广州”是特称,“市”是通称)、特称地名(如“甘肃”)。地名的特称只用于称呼某一个地点,地名的通称可通用于其他地点。地名的通称标识地名类型,蕴含着丰富的地理语义。许多地名歧义都是由地名的通称脱落以后引起的,如“新乡市”,当通称脱落以后,特称地名“新乡”就成为一个具有“新乡市”和“新乡县”两个候选地理位置的歧义地名了。

由于地名是 Web 文本中最常见的地理信息,因此,可以建立地名本体来完成 Web 文本中地名的识别。地名本体是关于地理位置的本体,其核心概念和主要描述对象是地理实体。地名本体涵盖地名领域的知识,提供对地名领域知识的共同理解,确定地名领域内共同认可的词汇,并给出这些词汇和词汇间相互关系的明确定义,为地名识别与歧义消除提供了依据。

中国县级以上行政区划(包括县级)分为省级、地级和县级三大类<sup>[4]</sup>。其中,省级行政区划分为:省、直辖市、自治区和特别行政区。地级行政区划分为:自治州、地区、盟和地级市。县级行政区划分为:县级市、县、自治县、旗、自治旗、特区、林区和市辖区。目前,我国县级以上行政区划有两种不同的类型:普通县级行政区划和省直管县级行政区划。地名本体中对行政区划的分类以此为基础,其构建借助部分一整体学(Mereology)、定位理论(Location Theory)和拓扑学(Topology)三大理论<sup>[5]</sup>及本体建模工具 Protege 3. 4<sup>[6]</sup>来完成。部分一整体关系符合人们对地理空间的认知过程,直接反映了人类组织地理空间知识的分层思想,在地名歧义消除中发挥着重要的作用。因此,在地名本体中,通过定义 isPartOf 和 isWholeOf 属性显式地表达了行政区划之间存在的部分一整体关系。地名本体中还定义了 touch、contain 和 within 3 种拓扑关系及 isChildOf、isParentOf、isSiblingOf、hasCapital、hasSpatialDescription 和 hasLocation 等关系。其中,hasSpatialDescription 用于表达行政区划的空间几何模型,hasLocation 用于表达行政区划的地理坐标,touch、contain 和 within 及 isChildOf、isParentOf、isSiblingOf、hasCapital 等则用于地名歧义的消除。

### 3 系统结构及功能描述

中文地名识别与歧义消除原型系统主要包括自然语言处理、概念关系库生成、地理解析和地理编码 4 个模块,其体系结构如图 1 所示。

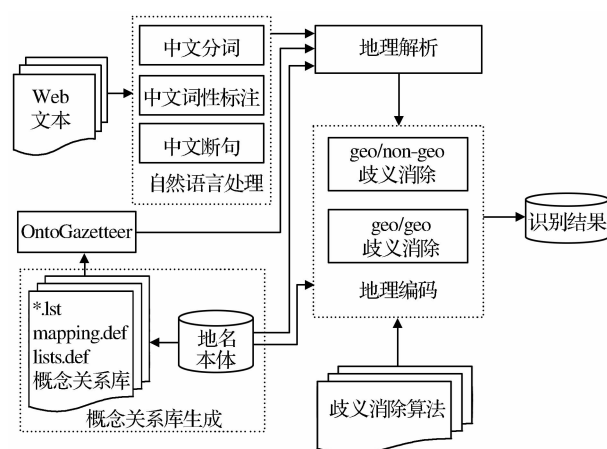


图1 中文地名识别与歧义消除原型系统的体系结构

Fig. 1 Architecture of Chinese toponym recognition and disambiguation prototype system

对于输入系统的 Web 中文文本,首先进行中文分词等自然语言处理,随后借助概念关系库及地名本体完成地理解析,最后根据地名本体和歧义消除算法,完成地理编码。地理解析和地理编码是整个过程中最重要的两个步骤。地理解析是指在自然语言理解的基础上,从 Web 文本中识别地名的过程;地理编码是指通过建立地名与地理位置之间的映射,为识别出来的地名指定地理坐标的过程。地名歧义消除是地理编码过程中最重要的环节。中文地名广泛存在着两种类型的歧义:geo/non-geo 歧义和 geo/geo 歧义。当一个地名有非地理含义的时候,就产生了 geo/non-geo 歧义。通俗地说,geo/non-geo 歧义是由于地名与普通名词(如人名、普通事物名称等)相同引起的。geo/geo 歧义主要是由多个地理位置使用同一个地名,即同名异地产生的。据统计,Web 页面中大约 37% 的地名存在 geo/geo 歧义<sup>[7]</sup>。

### 4 系统实现及地名歧义消除算法

中文地名识别与歧义消除充分利用了文本工程通用框架 GATE (General Architecture for Text Engineering)提供的各种资源及自然语言处理技巧。GATE 是英国舍费尔德大学自然语言处理研究小组开发的免费开源软件,它对自然语言处理的各个环节,包括语料收集、语义标注、系统性能评测等都提供了很好的支持<sup>[8]</sup>。

#### 4.1 自然语言处理

自然语言处理模块的主要任务是对输入系统的 Web 中文文本进行预处理,包括中文分词、中文词性标注和中文断句,从而为系统的后续处理提供良好的输入。中文文本使用的语言是汉语,汉语在形

式上与英语的最大区别在于构成句子的词之间没有明确的分隔符,句子之间由标点符号分隔,一个句子就是一个连续的汉字字符串<sup>[9]</sup>。因此,分词是中文地名识别的起点和基础。本文利用 ICTCLAS(Institute of Computing Technology Chinese Lexical Analysis System)来完成中文分词,同时完成词性标注。ICTCLAS<sup>[10]</sup>是当前世界上最好的汉语词法分析系统,由中国科学院计算技术研究所研制。中文断句则重用了 GATE 插件 ANNIE 的断句处理资源。至此,就完成了对 Web 中文文本的自然语言处理。

## 4.2 概念关系库生成

系统选用 GATE 框架的 OntoGazetteer 来完成基于本体的信息抽取。OntoGazetteer 是一个能够将词表中的词条与本体中的类相对应的处理资源,它的使用需要一个概念关系库。概念关系库根据地名本体生成,包括 3 类文件:多个 \*.lst 文件、一个 mappings.def 文件和一个 lists.def 文件。\*.lst 文件是由本体实例得到的词表文件,用于定义实体,一个实体就是一条词条,每个 \*.lst 文件代表一个待解析的实体类型,\*.lst 文件中定义的实体与地名本体中的地名实例相对应;mappings.def 文件描述 \*.lst 文件和地名本体概念之间的关系,即词条与类的对应关系;lists.def 为访问各 \*.lst 文件的索引文件,指明每个 \*.lst 文件所属的实体类型。概念关系库和地名本体一起,构成了中国行政区划地名领域知识库,成为中文地名识别与歧义消除的知识来源,系统运行过程中会将它们一次性载入 GATE 框架。

## 4.3 地理解析

Web 文本中出现的一个中文地名可以用如下五元组来表示:

$$r = (\text{location}, \text{type}, \text{value}, \text{index}, \text{instance}) \quad (1)$$

其中:location 为中文地名对应的地理位置,type 为该地理位置的行政区划类型,value 为表达地名的自然语言文本片段,index 为该文本片段在文本中的起始位置,instance 为地名本体中与该地名相对应的本体实例。

一篇 Web 文本中出现的所有可能的中国行政区划地名可以用一个五元组序列  $R(r_1, r_2, \dots, r_{i-1}, r_{i1}, r_{i2}, \dots, r_{ik}, r_{i+1}, \dots, r_{j1}, r_{j2}, \dots, r_{jm}, \dots, r_n)$  ( $k \geq 2, m \geq 2$ ) 来表示。其中  $r_1, r_2, \dots, r_{i-1}, r_{i+1}, \dots, r_n$  表示非 geo/geo 歧义地名的五元组,但是这些地名有可能存在 geo/non-geo 歧义,将由这些地名五元组构成的子序列记为  $R_1; r_{i1}, r_{i2}, \dots, r_{ik}$  和  $r_{j1}, r_{j2}, \dots,$

$r_{jm}$  分别表示存在  $k$  个和  $m$  个候选地理位置的 geo/geo 歧义地名的五元组,将一篇 Web 文本中 geo/geo 歧义地名的所有候选地理位置的地名五元组构成的子序列记为  $R_2$ 。地理解析需要明确五元组子序列  $R_1$  和  $R_2$ ,及  $R_1$  和  $R_2$  中每一个  $r$  的 value 和 index 取值。地理解析借助概念关系库和地名本体,利用 OntoGazetteer 处理资源来实现。

## 4.4 地理编码

地理编码需要明确一篇 Web 文本中所有地名构成的五元组序列及序列中每一个  $r$  的 location、type 和 instance 的取值。首先从  $R_1$  中删除具有 geo/non-geo 歧义的地名五元组,再为  $R_2$  中的每一个 geo/geo 歧义地名找到一个最佳地名五元组作为歧义消除结果,最后根据地名本体,将经过歧义消除处理的中文地名从文本空间映射到地理空间,使之对应于 GIS 的某种或某几种几何模型,明确隐含于地名中的空间特征及地理语义。

### 4.4.1 geo/non-geo 歧义消除

geo/non-geo 歧义消除是指从  $R_1$  中删除不具备地理含义的地名五元组,得到无歧义地名五元组序列  $R'_{\text{unamb}}$ 。

中国行政区划地名出现 geo/non-geo 歧义主要有以下 4 种情况:

#### (1) 特称地名与人名相同

中国行政区划地名中存在一些以人名命名的地名,如靖宇、左权等。此外,Web 文本中还有可能存在普通人名与特称地名相同的情况,如李北京等。原型系统对这类 geo/non-geo 的歧义消除主要借助人名列表和中文词性标注来完成。

#### (2) 特称地名与普通名词相同

普通名词经常出现在行政区划地名中,如灯塔、元宝等。不过,这种类型的歧义在省会城市、直辖市和大城市发生的情况非常少,原型系统对这类 geo/non-geo 歧义没有做处理。

#### (3) 特称地名出现在机构名中

特称地名尤其是城市特称地名常常包含在机构名称当中,作为机构的修饰词,如“兰州银行”。此类 geo/non-geo 歧义消除通过判断地名的后相邻词是否属于构成机构名称的特征词(如公司、有限公司等)来完成。

#### (4) 特称地名或特称加通称地名用作转喻

很多情况下,特称地名或特称加通称地名用来修饰另一个名词,如“北京队赢了上海队”,“从去年的人均年收入来看,四川省超过了甘肃省”。此时,

地名作为其他名词的属性类修饰词,没有必要对其进行识别及地理编码。排除地名是否用作转喻,是 geo/non-geo 歧义消除的一个难点,有些情况可以结合词性标注来完成。原型系统对这类最复杂的 geo/non-geo 歧义没有做处理。

#### 4.4.2 geo/geo 歧义消除

本文 geo/geo 的歧义消除遵守以下 3 个基本规则:

(1) 在文本中出现多次的地名,其每次出现都指向同一个地理位置。

该思想源于词的歧义消除。William 等<sup>[11]</sup>在研究词的歧义消除的过程中发现,96% 的多义词在同一篇论文中使用的是同一个词义,因此,提出了“Single Sense Per Discourse”的理论。Amitay 等<sup>[7]</sup>将这一理论引入到空间信息识别领域并验证了它的正确性。

(2) 同一文本中出现的地名往往存在着一定关系。

Martins 等<sup>[12]</sup>认为同一篇文本中的地名之间通常存在一定的关系。这意味着,同一 Web 中文文本出现的地名可能是相同的,如“兰州”和“金城”;可能是兄弟关系,如“武威”和“张掖”;可能是部分—整体关系,如“甘肃”和“兰州”;可能是相邻接的关系,如“北京”和“河北”等。

(3) 具有多个候选地理位置的特称地名出现在 Web 文本中的时候,行政区划等级最高的候选地理位置作为地名歧义消除结果的概率最大。这是因为,行政区划级别高的地理位置由于人口多,经济发达,更容易被关注。

中国县级以上行政区划同名异地的情况有两种:有隶属关系的行政区划使用同一个地名特称,如新乡市/新乡县,安阳市/安阳县等;无隶属关系的行政区划使用同一个通称加特称的地名或地名特称,如白云区(广州、贵阳)、昌邑市(潍坊)/昌邑区(吉林市)等。本文综合考虑了这两种类型的 geo/geo 歧义。下面是关于 geo/geo 歧义消除算法的 4 点说明:

(1) geo/geo 歧义消除算法的起点是 geo/non-geo 歧义消除后得到的无歧义地名五元组序列  $R'_{\text{unamb}}$  及 geo/geo 歧义地名所有候选地理位置的地名五元组序列  $R_2$ ,为了更直观地体现  $R_2$  的含义,算法中将  $R_2$  记为  $R'_{\text{amb}}$ ,即  $R'_{\text{amb}} = R_2$ 。

(2) geo/geo 歧义消除算法的终点是  $R'_{\text{amb}}$  为空(即完成了所有 geo/geo 地名的歧义消除处理),得到一篇 Web 中文文本所有的地名五元组构成的序

列  $R'_{\text{unamb}}$ 。

(3) 对于已经消除 geo/geo 歧义的地名,将其结果的地名五元组加入  $R'_{\text{unamb}}$  中的同时,从  $R'_{\text{amb}}$  中移除它的多个候选地理位置的地名五元组,并终止歧义消除算法,否则按算法顺序继续排除。

(4) 算法中明确提出针对“歧义地名”的步骤,对特称加通称地名和特称地名的歧义消除都适用;算法中明确提出针对“特称地名”的步骤,只适用于特称地名的歧义消除。

中文地名识别与歧义消除原型系统采用如下 geo/geo 歧义消除算法:

(1) 遍历  $R'_{\text{amb}}$  和  $R'_{\text{unamb}}$ 。如果在  $R'_{\text{amb}}$  和  $R'_{\text{unamb}}$  中存在相同的地名五元组  $r$ ,说明  $r$  作为无歧义地名在 Web 文本中的其他位置出现过。根据规则(1),可以明确  $r$  对应的地理位置,从而完成歧义消除。例如,文本中某处出现了无歧义地名“新乡市”,另一处出现了 geo/geo 歧义地名“新乡”,则“新乡”指的是“新乡市”。

(2) 对于歧义地名  $\text{value}_i$ ,存在  $n(n \geq 2)$  个候选地理位置  $r_{i1}, r_{i2}, \dots, r_{im}, \dots, r_{in}$ 。如果  $\text{value}_i$  的前相邻词或后相邻词为无歧义地名,即  $r_{i-1} \in R'_{\text{unamb}}$  或  $r_{i+1} \in R'_{\text{unamb}}$ ,那么根据中文地名的表达习惯,地名与其前相邻词或后相邻词之间往往存在部分—整体关系,即  $\exists r_{ik} \in R'_{\text{amb}} (1 \leq k \leq n)$ ,且  $\text{location}_{ik}$  是  $\text{location}_{i-1}$  的一部分(前相邻词无歧义的情况)或  $\text{location}_{i+1}$  是  $\text{location}_{ik}$  的一部分(后相邻词无歧义的情况),从而选择  $r_{ik}$ 。例如,“兰州市城关区”中的“城关区”,“甘南舟曲”中的“甘南”。

(3) 对于歧义地名  $\text{value}_i$ ,存在  $n(n \geq 2)$  个候选地理位置  $r_{i1}, r_{i2}, \dots, r_{im}, \dots, r_{in}$ 。如果  $\exists r_j \in R'_{\text{unamb}}$ ,使得  $\exists r_{ik} \in R'_{\text{amb}} (1 \leq k \leq n)$  且  $\text{location}_j$  与  $\text{location}_{ik}$  具有部分—整体关系,那么选择  $r_{ik}$ 。例如,Web 文本中某处出现了歧义地名“城关区”,而另一处出现了无歧义地名“兰州市”,则“城关区”指的是兰州市城关区,而不是拉萨市城关区。

(4) 对于歧义地名  $\text{value}_i$ ,存在  $n(n \geq 2)$  个候选地理位置  $r_{i1}, r_{i2}, \dots, r_{im}, \dots, r_{in}$ 。如果  $\exists r_{ik} \in R'_{\text{amb}} (1 \leq k \leq n)$  且  $R'_{\text{unamb}}$  中存在一个或多个地名与  $r_{ik}$  是“兄弟”关系,那么选择  $r_{ik}$ 。例如,Web 文本中某处出现了歧义地名“城关区”,而另一处出现了无歧义地名“七里河区”,它们都是兰州市的市辖区,存在“兄弟”关系,则 Web 文本中的“城关区”指的是兰州市城关区,而不是拉萨市城关区。

(5) 对于歧义特称地名  $\text{value}_i$ ,存在  $r_{i1}$  和  $r_{i2}$  两

个候选地理位置,  $location_{i1}$  是  $location_{i2}$  一部分, 即两个候选地理位置存在行政隶属关系。如果  $type_{i2}$  为地级行政区划“地区”或“自治州”,  $type_{i1}$  为县级行政区划, 且  $location_{i1}$  为  $location_{i2}$  的首府。自动内容抽取会议 ACE 2005 年制订的中文命名实体识别准则指出<sup>[13]</sup>, 大多数情况下, 人们提到这类歧义特称地名时, 都是指地区或自治州, 即选择  $r_{i2}$ 。如, Web 文本中某处出现了歧义特称地名“阿克苏”, 则指的是阿克苏地区, 而不是阿克苏市。

(6) 对于歧义特称地名  $value_i$ , 存在  $n(n \geq 2)$  个候选地理位置的地名五元组  $r_{i1}, r_{i2}, \dots, r_{im}, \dots, r_{in}$ 。如果  $\exists r_{ik} \in R'_{amb}(1 \leq k \leq n)$  且  $r_{ik}$  的行政区划等级最高, 那么根据规则(3), 选择  $r_{ik}$ 。例如, 对于歧义特称地名“海南”, 它的 3 个候选地理位置分别为海南省、海南藏族自治州(青海省)和海南区(内蒙古海市), 其中海南省的行政等级为最高的省级, 因而选择海南省最具歧义特称地名的最佳义项。

(7) 通过上述地名歧义消除处理, 现在只剩下极少量的县级行政区划同名引发的地名歧义, 包括多个市辖区同名(特称加通称地名相同或特称地名相同)和县与市辖区同名(特称地名相同)。受 Google 距离的启发, 可以通过 Google 为这些歧义地名赋予默认的地理位置, 从而达到地名歧义消除的目的。

Google 距离是由 Cilibrasi 等<sup>[14]</sup>提出来的。它是将 Web 中的数据作为语义词典, 在 Google 中输入词汇进行查询, 利用 Google 返回的匹配记录数来计算两个概念间的语义距离, 即 Google 距离, 用 NGD(Normalized Google Distance)表示, 它是从语义上分析两个词语的相似性的, Google 距离越大, 语义相似度越小。Google 距离的计算需要用到 Google 搜索引擎索引的页面总数, 由于 Google 公司 2005 年从首页上撤下了索引规模的数字, 使得计算 Google 距离并不容易。本文在 Google 中输入歧义地名作为关键词来搜索 Web, 统计搜索结果列表中各个候选地理位置的页面总数, 将页面总数最多的地理位置设定为歧义地名的默认地理位置。

## 5 实验与评测

实验以 50 篇 Web 中文文本作为语料, 它们是以歧义地名作为关键字, 使用百度新闻搜索引擎和百度搜索得到的 Web 网页。原型系统的性能评测借鉴了消息理解系列会议确立的量化评价指标, 即准确率  $P$ (Precision)、召回率  $R$ (Recall) 和  $F$

指数(F-Measure)<sup>[15]</sup>。召回率等于系统正确抽取的结果(即系统正确识别与歧义消除的地名数量)占有所有正确结果(即人工从 Web 文本识别出的地名数量)的比例; 准确率等于系统正确抽取的结果占有所有抽取结果(即系统识别与歧义消除的地名数量)的比例。

F 指数的计算公式如下:

$$F = \frac{(\beta^2 + 1.0) \cdot P \cdot R}{\beta^2 \cdot P + R} \times 100\% \quad (2)$$

其中:  $\beta$  为召回率和准确率的相对权重, 是一个在进行系统评测的时候预先设定的值。  $\beta = 1$  时, 召回率和准确率同样重要;  $\beta > 1$  时, 准确率更重要;  $\beta < 1$  时, 召回率更重要。在对原型系统进行评测的过程中, 将  $\beta$  设定为 1。通过使用 GATE 框架提供的评测工具 Corpus Benchmark Tool, 得到原型系统的准确率为 80.72%, 召回率为 91.15%,  $F$  指数为 85.62%。实验结果证明了原型系统设计的合理性, 及 geo/non-geo、geo/geo 歧义消除算法的可行性和有效性。同时, 存在 3 方面的因素影响系统的准确率和召回率: ICTCLAS 的分词不能保证 100% 正确, 存在误分的情况, 而且某些分词错误会传递, 即一处的分词错误会导致下一处的分词错误; 系统消除 geo/non-geo 歧义的能力不强; 系统消除 geo/geo 歧义的时候, 存在一定比例的错误。

## 6 结 语

本文研究的地名识别与歧义消除是基于地名本体的, 地名本体的好坏直接影响识别及歧义消除性能。地名不是一成不变的, 随着社会的发展, 旧地名逐渐消亡, 新地名不断涌现。在地名本体中考虑地名的变化能够增加原型系统的实用性。今后将要进一步研究的内容包括: 不断完善地名本体, 增加其他地理实体的同时, 考虑地名变化的因素, 甚至加入时间特征; 进一步改进歧义消除算法, 使其能够应用于其他类型中文地名的歧义消除。

### 参考文献(References):

- [1] Li Yusen, Zhang Xueying, Yuan Zhengwu. Study on Geographical Entity Recognition in GIS[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2008, 20(6): 719-723. [李玉森, 张雪英, 袁正午. 面向 GIS 的地理命名实体识别研究[J]. 重庆邮电大学学报(自然科学版), 2008, 20(6): 719-723.]
- [2] Studer R, Benjamins V R, Fenesel D. Knowledge Engineering: Principles and Methods[J]. Data and Knowledge Engineer-

- ing, 1998, 25(1-2): 161-199.
- [3] Egenhofer M, Mark D M. Naive Geography[C]//Frank A, Kuhn W. Spatial Information Theory—A Theoretical Basis for GIS, International Conference COSIT'95, Lecture Notes in Computer Science 988. Berlin: Springer-Verlag, 1995: 1-15.
- [4] Zhang Wubing, Jin Shulan. New Practical Chinese Atlas[M]. Beijing: Sinomaps Press, 2006. [张武冰, 晋淑兰. 新编实用中国地图册[M]. 北京: 中国地图出版社, 2006.]
- [5] Casati R, Smith B, Varzi A. Ontological Tools for Geographic Representation[C]//Formal Ontology in Information System. Amsterdam: IOS Press, 1998: 77-85.
- [6] Stanford University. Protégé[EB/OL]. <http://protege.stanford.edu/>, 2011-04-25.
- [7] Amitay E, Har'El N, Sivan R, *et al.* Web-a-Where: Geotagging Web Content[C]//Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, Sheffield, UK. New York: ACM Press, 2004: 273-280.
- [8] University of Sheffield. GATE HOME[EB/OL]. <http://gate.ac.uk/>, 2011-04-25.
- [9] Zhang Xueying, Jürgen Krause. An Approach to Automatic Keyword Extraction in Chinese Text[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(4): 512-520. [张雪英, Jürgen Krause. 中文文本关键词自动抽取方法研究[J]. 情报学报, 2008, 27(4): 512-520.]
- [10] ICT of Chinese Academy of Sciences. ICTCLAS[EB/OL]. [http://ictclas.org/ictclas\\_feature.html/](http://ictclas.org/ictclas_feature.html/), 2011-04-25.
- [11] William A G, Kenneth W C, David Y. One Sense Per Discourse[C]//Proceedings of the the 4th DARPA Speech and Natural Language Workshop, February 1991, Pacific Grove, California, US. New York: ACM Press, 1992: 233-237.
- [12] Martins B, Manguinhas H, Borbinha J, *et al.* A Geo-temporal Information Extraction Service for Processing Descriptive Metadata in Digital Libraries[J]. e-Perimtron, 2009, 4(1): 25-37.
- [13] ACE Linguistic Data Consortium. ACE Chinese Annotation Guidelines for Entities[EB/OL]. [http://projects.ldc.upenn.edu/ace/docs/Chinese-Entities-Guidelines\\_v5.5.pdf](http://projects.ldc.upenn.edu/ace/docs/Chinese-Entities-Guidelines_v5.5.pdf), 2011-04-25.
- [14] Cilibrasi R L, Vitanyi P M B. The Google Similarity Distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383.
- [15] Li Baoli, Chen Yuzhong, Yu Shiwen. Research on Information Extraction: A Survey[J]. Computer Engineering and Applications, 2003, 39(10): 1-5. [李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10): 1-5.]

## Recognition and Disambiguation Chinese Toponym from Web Texts ——Take the Names of Chinese Administrative Division above County for Example

Du Ping, Liu Yong

(College of Earth and Environmental Science, Lanzhou University, Lanzhou 730000, China)

**Abstract:** This paper introduces the concepts of ontology, geographic ontology and toponym ontology, discusses some key issues on recognition and disambiguation Chinese toponyms from Web texts based on natural language processing using toponym ontology under GATE (General Architecture for Text Engineering) framework, designs and implements a prototype system. By eliminating the geo/non-geo and geo/geo ambiguities, rich semantics and precise geographical coordinates are given to the extracted toponyms which correspond to the locations on the earth's surface. At last, an experiment is made and the result is analyzed.

**Key words:** Recognition toponyms; Disambiguation; Geoparsing; Geocoding