文章编号: 1671-8860(2005)12-1100-04 文献标志码: A

基于空间语义角色的自然语言空间概念提取

乐小虬[1] 杨崇俊[1] 于文洋[1]

(1 中国科学院遥感应用研究所遥感科学国家重点实验室, 北京市 9718 信箱, 100101)

摘 要: 根据空间信息的特点, 从定义的空间语义角色入手, 通过语义角色标注、短语识别以及概念模式匹配等手段, 具体分析了自然语言中的空间实体、实体间空间关系以及空间过程的表达与提取方法。

关键词: 空间语义角色; GIS; 信息提取(IE); 自然语言处理(NLP)
中图法分类号: P208

目前, 空间概念提取主要集中在空间命名实体和空间短语识别上, 如在网页空间知识提取中通过识别地理实体、区域编码、邮政编码等命名实体来分析网页的空间相关性[1], 在空间信息提取中利用地名词典提取命名实体以及由介词等构成的空间短语[2]; 在信息提取(IE)领域, 许多事件模板中均定义了 Loc 这一空间短语项。这些空间概念的提取一般具有较高的识别率, 但没有体现实体间的空间关系及空间过程, 因此, 所反映的空间概念不够完整。在一些特定的应用领域, 如人机对话、事故场景动画生成等领域中[3], 虽然涉及实体间的空间关系和过程, 但针对性强, 限制条件多, 缺乏普遍性。对于自然语言中完整空间概念的提取方法缺乏系统性研究。

概念提取通常涉及语义分析, 而语义分析又与语义角色密切相关。语义角色是与谓词相关的变元类型。在自然语言处理中常见的语义角色有施事、受事、与事、工具、结果、处所等, 但这些角色并不能完全表达复杂的空间概念, 例如, 现阶段常用的两种语义标注系统。因此, 本文根据空间信息的特点, 定义了适合于空间概念提取的空间语义角色, 并通过对这些语义角色的标注、短语识别以及模式匹配来提取复杂的空间概念。

1 空间概念表达

在 GIS 中, 地理空间的描述常围绕着空间实体、实体间的空间关系及时空过程 3 方面进行。为了从自然文本中识别出这些空间概念, 本文定义了相应的空间语义角色——实体(En)、属性(Ar)、路径(Path)、运动(Motion)、时间(Time)、拓扑关系(SRT)、方向关系(SRD)、度量关系(SRM)。这些语义角色相组合, 便可构造出三类空间概念的结构单元。以下分别对其形式化表达方式进行定义。

1) 空间实体

(En: (Name: value) (Ar: (propertyList)))

其中, propertyList 是句中描述实体的属性列表。

2) 空间关系

SRT/ SRD/ SRM: Prediction(En1, En2, …)

其中, Prediction 是空间谓词, 此处指空间关系谓词。

3) 空间过程

(SPr: (AGENT: En) (OBJECT: En) (Time: value) (Path: Prediction) (Motion: Prediction))

其中, AGENT 为过程的主体, OBJECT 为受体, 两者取值均为空间实体 En。Path 与 Motion 的取值用谓词表示。

2 提取方法

2.1 基本流程

将空间概念提取看成是一个原子空间语义角色标注、短语识别、句法模式匹配的三级信息提取与过滤过程。图 1 是系统的流程框图。

其提取过程如下。

收稿日期: 2005-11-03。
项目来源: 国家 973 计划资助项目(G2000077906)。

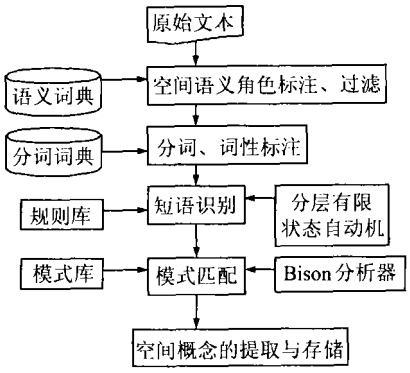

图 1　空间概念提取流程框图

Fig. 1　Framework of Spatial Concept Extraction

1) 分析有关文集, 构建空间语义词典, 搜集规则库、模式库。

2) 利用空间语义词典, 采用正向最大匹配法识别出文本中的空间语义角色, 并进行标注。不含标注的语句被过滤掉。

3) 对语句中非标注片段进行分词处理, 并进行词性标注。

4) 利用分层有限状态自动机进行短语分析, 识别出空间语义角色短语。

5) 构造空间概念表达模式的识别文法, 采用句法分析器进行模式识别。

6) 根据句法模式分析空间语义, 并提取相应的空间概念。

2.2　空间语义词典

在自然语言中, 空间概念涉及的实体、关系、过程是通过词、短语、句法表达出来的, 所以识别过程需要参照相应的词典和语法规则。词典通常包含词性和词义两项内容, 为了提取复杂的空间语义, 本文根据上文定义的空间语义角色对词义进行扩展, 并构建了相应的空间语义词典。同时, 对构成空间语义角色的短语规则和句法模式也进行了归纳, 并存储于库中。以下对词典中各语义项分别进行描述。

1) SE: 描述空间实体角色, 包含 Ec(实体类) 和 En(实体个体) 两种类型。如“街道”为 Ec, 而“王府井大街”则为 En。实体主要来源于 GIS 属性库。

2) SR: 描述空间关系角色, 含 SRT、SRD、SRM 三类角色。

其中拓扑关系角色的内容参照了 Kaoru 和 Femke(2004) 根据九交模型得出的 8 种地理本体拓扑关系——相等(equals)、相离(disjoint)、相交(intersects)、相接(touches)、穿越(crosses)、内部(within)、包含(contains)、重叠(overlaps)。

方向关系角色的内容参照八方向关系系统(东(isEastOf)、南(isSouthOf)、西(isWestOf)、北(isNorthOf)、东南(SouthEastOf)、东北(NorthEastOf)、西南(SouthWest)、西北(NorthWest)) 以及相对方向关系(前(isBeforeOf)、后(isBehindOf)、左(isLeftOf)、右(isRightOf)、上(isUpOf)、下(isDownOf)) 等。

度量关系含空间距离和时间距离, 分别用 withinDistanceOf 和 withTimeOf 表示。

3) SP: 指空间过程角色, 含 Path 和 Motion 两种类型。路径角色 Path 定义了 5 种谓词——从(From)、到(To)、经由(Via)、朝(Toward)、沿(Along); 运动角色 Motion 定义了走(Walk)、跑(Run)、去(Go)、来(Come)、离开(Left)、移动(Move) 等谓词。

4) AR: 指属性角色。该语义项有很多类型, 可根据具体的应用而确定。一般参考 GIS 属性库中字段的设置。本文选择了形状(Shape) 和颜色(Color) 两种属性, 属性值用关键字或代码表示。

5) LA: 指单词的词性, 与通常词典的内容相同。词性标注选用北大语言所制定的现代汉语语料库加工规范。

以上语义项中前四项均涉及空间语义角色, 每种角色均对应着许多实例。在词典构建时需收集大量用于表达这些语义角色的汉语词汇, 包括对应的同义词、近义词等。第 5 项 LA 用于模式分析。

2.3　标　注

由于汉语词间没有分隔, 信息提取前通常应先进行分词处理和词性标注。分词的方法有很多种, 如基于词典的分词、基于语料库的分词以及两者相结合的方法等。由于空间概念提取只涉及具体领域, 文中有大量与空间概念无关的语句, 这些语句无需进行分词与标注处理, 所以本文采用先进行过滤、语义角色标注, 后进行分词和词性标注的处理模式。具体做法是利用空间语义词典采用正向最大匹配法进行空间语义角色的识别, 含语义角色的语句被存入缓存并标注相应的语义角色, 否则被丢弃; 然后依次取出缓存中的语句, 对其中未标注部分采用基于语料库的隐马尔可夫模型(HMM) 进行分词和词性标注, 其具体算法参见文献[6]。这种处理模式针对性强, 不仅减少了无效语句的处理量, 而且能克服因分词结果不正确而丢弃的有效语句, 同时也避免了在分词基础上进行语义角色分析时出现的词语不一致的问

题。

2.4 短语分析

经标注后的语句虽然识别出空间语义角色,但这些语义角色只是原子级片段,孤立地分布于语句中。如果直接以这些片段为基础提取空间概念,则会因为自然语言的复杂性而使分析难以进行。所以应在相邻片段间建立联系,形成语义角色短语。

Sameer (2004)将语义角色短语分析看成是一种多分类问题,并利用支持向量机(SVM)进行分类识别,在试验中取得了现今所有分析方法中最好的结果[7]。但它是通过训练 PropBank 语义标注库实现的,由于目前还没有类似的汉语库,该方法对汉语还无能为力。所以本文仍采用常用的基于分层有限状态自动机的浅层分析技术进行空间语义角色的短语分析。有限状态自动机是具有离散输入与输出系统的一种数学模型[8],利用状态转换图能构造出高效的短语结构分析程序。分层有限状态自动机通过将第 $n-1$ 层自动机的输出变为第 n 层自动机的输入,使短语结构逐层得到识别。

本文采用两层有限状态自动机识别空间语义角色短语。这些短语包括空间实体短语(ENTRY_P)、拓扑关系短语(SRT_P)、方向关系短语(SRD_P)、路径短语(PATH_P)、数量短语(QUANTITY)、时间短语(TIME_P)。短语的构成规则由标注的词性和空间语义角色确定。由于有限状态自动机和正则表达式是等价的,并且可以互相转换,以下用正则表达式表示识别部分短语的有限状态自动机。

1: ENTRY_*P*→(AR *u*?)* En+ (*c* En)*

TIME_*P*→*p*? *t*+

QUANTITY→*m q*

2: SRT_*P*→(*p*? ENTRY_*P*? *u*? SRT+(*c* SRT)*)

|(*p* ENTRY_*P*)

SRD_*P*→*p*? ENTRY_*P*? *u*? SRD+(*c* SRD)*

PATH_*P*→Path ENTRY_*P u*?(SRT*)|(SRD*)

式中,数字表示自动机的层号;"→"右边的表达式描述了左部短语的构成模式;小写字母 p、m、q、c、t 为词性,分别代表介词、数词、量词、连词、时间词;"*"表示前面的表达式可出现零次或多次;"+"表示出现一次或多次;"|"表示逻辑或;"?"表示出现零次或一次。

2.5 识别文法

语义角色短语的识别使句中空间概念的粒度从词级别上升到短语级别,但仍然停留在局部语义片段的处理上,要得到完整的空间概念还需在更高层次上进行分析。由于完全句法分析与篇章分析至今仍很困难,且大多效率不高,本文尝试用句法模式进行识别。这种句法模式不是传统意义上的句法分析树,它是汉语表达空间概念时空间语义角色与语言成分间(如词性)的组合模式,相当于空间概念表达模板,它由语义角色短语、词性、关键字组合而成。本文将其分为静态和动态两种类型,前者概括了汉语表达静态空间概念的语言模式,后者表述了动态空间概念的表达模式。只要归纳出这些模式,就可通过构造相应的句法分析程序,达到在语句层次上识别空间概念的目的。

2.6 概念提取

利用 Bison 分析器处理上述句法分析程序便可对空间概念进行模式匹配。选用 Bison 分析器主要出于性能上的考虑。因为汉语表达空间概念的模式有很多种,如果采用常规的匹配方法(如 KMP, B-M),则需为每个模式 n 构造一个有限状态自动机,然后逐个地匹配原文 m,其最优的线性复杂度为 $O(m+n)$;而 Bison 是一种 LALR(1)句法分析器,只要能构造出无冲突的分析程序,则无论有多少模式,其线性复杂度都是 $O(m)$。

由于每种模式中空间语义角色与语言成分的构成方式都是确定的,在根据上述文法构造 Bison 分析程序时,每种模式均可返回相应的空间概念。这样,一旦某语句被成功匹配,则其中的空间概念即可被方便地提取出来。这些概念按照§2 中定义的表达形式分别存储于库中,供其他系统分析使用。

3 试验结果与讨论

本文选用非结构化和半结构化文本(网页)对上述提取方法进行了测试。测试样本 30 篇,其中 5 篇为非结构化文本,来源于军事文书,因为此类型文书中含丰富的空间概念;另外 25 篇为网页,获取方法是随机抽取 25 个地名,分别用搜索引擎进行搜索,取排位为 1 的网页源文件作为样本。评估对象为三类空间概念,即空间实体、空间关系、空间过程。评价指标为信息提取中常用的准确率和召回率两指标。前者是正确提取的结果数与所有结果数量的百分比;后者是正确提取的结果数与所有正确结果数量的百分比。

试验方法是先将这些样本进行预处理,去除网页中的标记符后取出正文部分;然后先以人工

的方式将其中的空间概念提取出来, 分别统计空间实体、空间关系、空间过程的出现次数; 再将经预处理后的样本以文件的形式输入到系统中, 并输出系统提取所有空间概念, 与人工提取的结果相比较便可得到正确提取的结果数。利用上述计算方法即可得到系统的准确率与召回率。

经测试, 本方法提取的空间实体准确率为 87%, 召回率为 91%, 二元关系准确率为 86%, 召回率为 58%。与通常的信息提取系统性能(二元关系准确率和召回率约 60~ 70 %)相比, 该方法提取的实体识别性能相当, 只要词典库能得到更新, 两指标均有提高的空间。二元关系中准确率有所提高, 召回率不足。这主要是因为采用了原子空间语义角色标注、短语识别、模式匹配三层过滤机制所致。受人工归纳的空间概念表达模式的影响, 目前召回率还有待提高, 这主要是因为表达模式在短期内很难达到完备性, 需在试验中不断地进行修正和补充。

致谢: 感谢刘冬林老师在系统开发中给予的帮助和指导。

参 考 文 献

1 Morimoto Y. Extracting Spatial Knowledge from the Web. http://www.morimo.com/morimo-ken/pub/120_morimoto_y.pdf, 2003

2 Schilder S. Extracting Spatial Information: Grounding, Classifying and Linking Spatial Expressions. http://www.geo.unizh.ch/~rsp/gir/abstracts/schilder.pdf, 2004

3 Johansson R. Carsim: A System to Visualize Written Road Accident Reports as Animated 3D Scenes. http://www.cs.lth.se/home/Pierre_Nugues/Articles/acl2004/acl2004.pdf, 2004

4 Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet Project. http://framenet.icsi.berkeley.edu/~framenet/papers/acl98.pdf, 1998

5 Gildea D L, Hockenmaier J. Identifying Semantic Roles Using Combinatory Categorial Grammar. http://www.cs.rochester.edu/~gildea/gildea-emnlp03.pdf, 2003

6 俞士汶. 计算语言学概论. 北京: 商务印书馆, 2003

7 Pradhan S. Semantic Role Parsing: Adding Semantic Structure to UnstructuredText. http://oak.colorado.edu/~spradhan/publications/pradhan-icdm-2003.pdf, 2003

8 杜淑敏, 王永宁. 编译程序设计原理. 北京: 北京大学出版社, 2001

第一作者简介: 乐小虬, 博士生。主要研究方向为空间语义 Web、智能搜索引擎。
E-mail: xiaoqiule@yahoo.com.cn

Spatial Concept Extraction Based on Spatial Semantic Role in Natural Language

LE Xiaoqiu[1]　*YANG Chongjun*[1]　*YU Wenyang*[1]
(1 State Key Laboratory of Remote Sensing Sciences, IRSA, CAS, POX. 9718, Beijing 100101, China)

Abstract: This paper presents an approach for extracting complex spatial concepts from unstructured text. By defining several spatial semantic roles based on the characteristics of geospatial information in natural language, the authors extract spatial entities, spatial relationships among entities and spatial procedures by means of spatial semantic annotation, semantic phrases recognition and pattern match. It tries to solve the problem of unable to get deep spatial semantics in common IE. The primary experiment shows that it has a good precision and a similar recall comparing to common IE systems.

Key words: spatial semantic role; GIS; IE; natural language process(NLP)

About the first author: LE Xiaoqiu, Ph.D candidate. His main research interests are spatial semantic web, intelligent search engine. E-mail: xiaoqiule@yahoo.com.cn