摘要

自然语言是人类空间认知结果的主要表现形式，文本即是人类最常用的一种自然语言，也是一种重要的原始空间数据来源。从文本中获取未分析的、非显性的空间信息已成为当前地理信息科学迫切需要解决的问题。GIS 自然语言空间查询，自然语言路径描述，汉语文本空间关系抽取，文景转换及场景重建等领域已成为当前地理信息科学研究的热点。目前 GIS 自然语言空间关系的研究对象局限在选定的自然语言空间词汇以及有限文法的空间关系描述，通常从整体上关注“描述-解析”的映射过程。而在 GIS 自然语言空间查询及路径描述领域，受限性决定了其对文本空间信息抽取研究的局限。为此，我们提出了面向 GIS 的文本空间关系“描述-识别-解析”的整体框架，而本文通过引入中文信息处理领域的研究范式及技术，对“描述-识别”阶段中文文本空间方位关系的抽取方法进行了系统的研究。

本文研究主要分为四个部分：（1）面向信息抽取的中文文本空间方位关系表达；（2）中文文本空间方位关系的语料标注与分析；（3）基于规则的中文文本空间方位关系抽取方法；（4）基于机器学习的中文文本空间方位关系抽取方法。主要研究内容与成果如下：

（1）中文文本空间方位关系表达

首先，在分析认知语言学的空间方位参照框架及中文文本空间方位关系描述特征的基础上，通过文本与地图两类符号系统在空间关系表达上的对比，提出了中文文本空间方位关系表达的两个层次，以及受语境约束的术语在空间方位关系表达研究中的纽带作用。接着，从空间方位关系表达式、分类、类型术语的判断三个角度分析了面向信息抽取的空间方位关系表达。其中，文本意象图式与 RCC8 类型的术语映射，即能有效的指导对文本空间方位关系的分类判断，也为抽取后 GIS 空间关系的解析提供了基础。

（2）中文文本空间方位关系的语料标注与分析

在 SpatialML 标注集及标注规范的基础上，采用 GATE 平台进行空间方位关系语料的标注工程。针对地理辞典类语料的领域特征进行处理：扩展地理命名实体组合情况的标注，并设计了归一规则；扩展地理空间描述中拓扑关系判断的规范。通过对实验语料的统计，分析了空间方位关系描述中的句法特征及空间词汇的指示性，为抽取方法的研究提供了基础。

（3）基于规则的中文文本空间方位关系抽取方法

通过构建空间词汇词典，结合关系模板及匹配规则，研究基于规则的空间

方位关系抽取方法。首先使用 BootStrapping 迭代获取空间词汇，利用词汇语义型词典中概念相似度、相关度计算对空间词汇进行语义参数的设置；然后通过文本序列比对及模板聚类进行空间方位关系实例模板的泛化，结合 ANNIC 辅助人工归纳抽取模板；最后，使用 OntoGazetteer 构建空间词汇词典，JAPE 正则文法引擎编写抽取模板，并通过扩展匹配算法实现基于规则的空间方位关系抽取方法。实验表明，使用不同词性类型的空间词汇同时作为种子词，选择丰富的特征向量，通过 BootStrapping 迭代方法获取的新增空间词汇的准确率最高；使用词汇语义型词典中概念相似度、相关度计算，对空间词汇语义参数的设置具有一定的效果，但由于其通用性，还需较多的人工修正；针对特定的空间方位关系描述文本，通过半自动的构建空间词汇知识库以及适量的文法规则，可取得较好的抽取效果。

（4）基于机器学习的空间方位关系抽取方法

选用关系抽取中性能最佳的支持向量机进行基于机器学习的空间方位关系抽取方法研究。首先，引入统计学习理论与支持向量机方法，分析了其适用于空间方位关系抽取的两个特性：结构风险最小化原则与使用核函数实现到高维特征空间的非线性映射。接着，讨论了空间方位关系的实例化方法，特征向量构建及抽取流程。实验表明，基于支持向量机的抽取方法具有较好的发现性能，且不依赖于空间词汇的识别，具有一定的实用意义。

（5）空间方位关系抽取的应用研究

简要分析了空间方位关系抽取在地理知识库构建和场景重建中的应用模式，通过文本驱动的地理知识库构建以及空间方位关系的图形重建演示了相关的应用场景。

关键词：空间方位关系，空间词汇，空间关系抽取，支持向量机，BootStrapping

Abstract

Natural language is the main way for human to express the spatial cognitive results. As a most commonly used natural language, text is also an important resource of original spatial data. How to extract the non-analytical and non-explicit spatial information has become an urgent problem in geographic information science. Research on Natural Language Spatial Relation sponsored by NCGIA Language of Spatial Relation program focuses on the study of human's cognitive mechanism on space by building semantic mapping between spatial words and GIS computational models. By far, the natural language query on GIS, route description in natural language and extraction of Chinese text spatial relation have become the hot issues in geographic information science.

The previous research on GIS natural language relations is aims to selected spatial words of natural language and spatial relation description of limited grammar, i.e. the "description-resolution"mapping. And in the field of GIS natural language query and route description, this kind of limitation will definitely lead to the poor performance of spatial information extraction in text. Hence, a GIS-oriented integral frame for "description – recognition - resolution" of spatial relation in text is proposed in this paper. After introduction of the basic techniques in Chinese information processing and representation theories of GIS spatial relations, this paper studies the a few effective ways for extraction spatial relations in Chinese text. It includes the following four parts:

(1) Information-Extraction-Oriented representation of spatial relation in Chinese text

First of all, based on the analysis of the spatial reference frames of cognitive linguistics and the characteristics of Chinese text spatial relation description, this paper proposes two levels of representation of spatial relation in Chinese text and the connection effect of the context-restricted termimology on the researches of spatial relation representation, by the comparison of spatial relation representation between text and map symbol systems. Meanwhile, the researches of this dissertation and the program it belongs to have been oriented. And then, GIS-oriented representation of spatial relation is analyzed in three different aspects, including expression, classification and termimology identification of the spatial relation. As part of this study, text image-schematic structure and term mapping of RCC8 type not only can be used to effi-

ciently support the classification decision of spatial relation in text, they can also provide the basis of the resolution after GIS spatial relation extraction.

(2) Language annotation and analysis of spatial relation in Chinese text

Based on the SpatialML annotation set and criteria, this dissertation uses GATE platform to conduct the annotation of spatial relation corpus. According to the characteristics of the geographic gazetteer corpus, this study extends the annotation of the combination situations of geographic namd entity, and designs unified criteria and the identification criteria of topology relation in geospatial descriptions, and by experimental corpus statistics, analyzes the syntax characteristics in spatial relation description and the indicative property of spatial words, which lay the foundation for research on extraction methods.

(3) Approach to spatial relation extraction in Chinese text based on rule

By building spatial words dictionary, according to relation template and matching rule, this dissertation studies the ways of spatial relation extraction based on rule. Firstly it captures spatial words using BootStrapping iteration, and configures the semantic parameters using semantic relevancy computation and semantic similarity computation; Then it generalizes the instance template of spatial relation by sequence matching technique, and manually concludes extraction templates with ANNIC. Finally, spatial words is collected with OntoGazetteer, extraction templates are modified with JAPE regular grammar engine, and spatial relation extraction algorithm is presented base on rules by means of extension of matching calculation. Experimental results show that, as for Gazetteer corpus, using spatial words of different parts of speech as word seeds, choosing varied feature vector, getting newly added spatial word by BootStrapping iteration can achieve the highest accuracy. It is effective to some extent to configure spatial words semantic parameters using semantic relevancy and semantic similarity computation, however, because of its universality, much manual vertification is needed. As for specific text of spatial relation descriptions, it is effective in extraction to semi-automatically build spatial words knowledge base and adequate grammar rule.

(4) Approach to spatial relation extraction in Chinese text based on SVM

This paper studies an approach to spatial relation extraction based on machine learning by choosing the best Support Vector Machine in the relation extraction. Firstly, it introduces Statistical Learning Theory and SVM methods, and analyzes their two characteristics of being applicable to spatial relation extraction, including

Structure Risk Minimization and Kernal Function. And then, it discusses the instantiation methods of spatial relation, and the building of feature vector and extraction process. At last, by the extraction of mission discovery and recognition, it shows that discovery of spatial relation can be conducted based on SVM.

(5) Application cases of spatial relation extraction

This paper briefly analyzes the application patterns of the spatial relation extraction in geographic knowledge base building and Scene Generation. Moreover, it displays related application scene by text-driven geographic knowledge base and rebuilding graphics of spatial relation.

Key words: Spatial Relations in Chinese Text, Spatial Words, Automatic Extraction of Space Relations, Support Vector Mechine, BootStrapping

图目录

表目录

第1章 绪论

1.1 选题背景与研究意义

自然语言和地图一样，是人类空间认知结果的主要表现形式。在当前各类时空信息涌现的情况下，GIS 领域较多关注由测绘获得的基于几何坐标形式的时空信息处理和知识获取。然而，文本也是一种重要的原始空间数据来源，从文本中获取未分析的、非显性的空间信息已成为当前 GIS 迫切需要解决的问题[1]。

信息检索（Information Retrieval，IR）、自然语言处理（Natural Language Processing，NLP）和 GIS 等领域对文本中的地名解析（Toponym Resolution，TR）、地理参考获取（Georeferencing）、以及基于其上的地理信息检索（Geographic Information Retrieval，GIR）、基于位置服务（Location-Based Services，LBS）进行了大量的研究。如何对文本中的空间关系进行识别和解析，是这些研究得以深入必将面临的问题。而在面向 GIS 的文本空间关系“描述-识别-解析”的整体研究框架中，自动识别过程起着承上启下的作用。

面向 GIS 的中文文本地理空间关系的自动识别研究，具有以下三个方面的研究意义：首先，有助于文本空间信息抽取研究的深入。文本（尤其是网络文本）中蕴含着海量的空间信息，在对地理命名实体进行自动识别的基础上，需要对更深层次的空间关系进行识别。其不仅可以为地名解析算法提供更多的约束证据，也是为篇章级的空间过程信息抽取提供准备。其次，有助于文本与地图两种表意系统间空间关系进行语义映射的研究。地图综合、空间本体、虚拟环境等研究方向越来越多的关注于时空模型（尤其是地图表现）与人类空间认知的契合度，而自然语言作为人类认知的窗口，是研究认知的重要途径。文本空间关系自动识别，为其解析为地图表达提供基础。最后，在地理信息检索、基于位置服务、自然语言空间查询接口等应用层面，本研究可为突破交互语言的受限约束提供借鉴。

1.2 国内外相关研究现状

1.2.1 语言学中方位关系研究

空间关系表达是人类语言的一项基本功能，语种与文化对于自然语言中的空间关系表达有着重要影响[2]。每种语言都有一套能够完整表达空间关系的词汇系统，利用这些词汇和相关句法，人们可以组织各种各样的语句对认知空间世界的图景加以描写、叙述或说明[3]。空间语言结构都有特定的认知基础、组织原则、表达顺序和使用习惯[4]。自然语言中空间关系的表达手段比较多样化，主要包括“目的物”、“参照物”和“空间参照系统”等要素[5]。人们在运用语言表达事物

之间的空间关系时，一般都要采用词汇手段和语法手段相结合的方法[6]。

空间关系的描述与空间认知之间存在密切的联系[7]。一方面，空间方位关系作为一种物理关系是以人类自身的认知为基础的。不同类型的语言在空间关系的隐喻方面存在着很强的一致关系[8]，反映了空间语言植根于共性的空间认知。另一方面，“新沃尔夫假设”（neo-Whorfianism）认为语言部分地决定了人的认知方式，人类的空间思维在很大程度上受文化因素，特别是语言因素的影响[9]。空间方位参照框架即反映了语言社会对认知空间中方位关系的认知过程和认知方式，是一种立体的、抽象的、深层的认知结构[3]。根据图形-背景理论及空间方位参照框架理论，人们在运用语言表达事物的空间关系时，总是要把符合背景的本质特征（definitional characteristics）和联想特征（associated characteristics）的事物作为“参照物”，把符号图形的本质特征和联想特征的事物作为“目的物”，并以此为依据来确定事物之间的空间关系。

对于汉语而言，空间方位词是汉民族空间认知经验范畴化的结果[10]。在汉语语法界，人们侧重于研究方位词的种类、来源、语义、演变、性质、作用等。文炼把方位词分为单纯方位词与合成方位词，认为有的单纯方位词可以直接用在名词和名词短语前后，构成表示时间和处所的短语。其列举了十六个单纯方位词，即“上、下、左、右、前、后、东、西、南、北、里、外、内、中、间、旁”。这十六个方位词基本概括了所有的空间位置关系[11]。李静熹从功能角度出发，认为方位词是表示空间范围的形式标志，本身具有突显空间特征的功能，同时还赋予了参照物以突显空间特征的功能[12]。廖秋忠详尽地分析了现代汉语篇章中的空间、时间参考点，指出空间定位的格式一般为“（在+）名（+的）+空间方位词（+距离）”[13]。朴珉秀运用认知语言学和语义学的理论和概念，探讨了现代汉语空间系统，其主要研究对象为方位词“前”、“后”、“上”、“下”，并提出了现代汉语“空间方位”系统的结构框架，探讨了与方位词所表示的方向场有关的认知推理[14]。方经民认为方位参照包含参考点、方位词、方位辖域及其中客体、观察点等四个要素。其将方位参照分为外物参照、整体参照和自身参照三种类型[15]。郭锐指出导致水平方位词歧义的原因有三种：参照方式的不同、参照策略的不同、说话人对固有方位特征的认识不同[16]。此外，陶氏河宁指出“东”、“西”、“南”、“北”这类方向义明显的方位词，其不仅仅表示单纯的空间概念，还与中国人的思想概念、风俗习惯、日常生活等信息有关[17]。

1.2.2 GIS 空间关系研究

在以几何图形与视觉符号作为主要表意系统的 GIS 领域，空间关系是指地理实体对象间存在的具有空间特性的关系，主要描述具有一定位置、形态和属性的空间对象（包括单目标和群目标）之间的各种几何关系[18]。空间关系的语义

一般分为拓扑关系、顺序关系、度量关系。郭仁忠将空间关系分为：距离关系、方位关系、拓扑关系、相似关系、相关关系。其中，距离关系和方位关系是度量关系和顺序关系的主要方面[19]。而后两种关系则是在考虑了群目标后的扩展。Egenhofer 指出空间关系表达了空间数据之间的一种约束[20]：其中度量关系对空间数据的约束最强、而顺序关系次之、拓扑关系最弱。Mark 认为空间关系中，拓扑关系是本质的，起着定义的作用，而度量信息（大小、形状、距离和方向等）起细化的作用[21]。一些空间术语纯粹是拓扑的，与度量无关，如“进入”和“在内”等。另一方面，度量信息则决定了空间术语的语义，如“北”和“附近”都细化了“相离”拓扑关系[22]。从另一个角度看，在日常生活中，方向关系是一类比拓扑关系使用频率更高的空间关系[23]。拓扑和度量在定义空间关系特性方面，以及它们在论证空间关系过程中或描述空间关系的语言中所扮演的相关角色是复杂的[21]。

以下从三个方面论述 GIS 空间关系与本文相关的研究：基于认知的空间关系定性表达；空间谓词的语义映射；中文自然语言空间查询。

1.2.2.1 空间关系定性表达

定性空间推理（Qualitative Spatial Reasoning，QSR）的引入，为 GIS 空间关系的定性表达和推理提供了数学与逻辑的基础。4 交和 9 交模型是基于点集拓扑的一种拓扑关系表达模型，能统一、简洁、完备的表示两个简单实体间的拓扑关系[24] [25, 26] [27]。在 4 交模型区分的拓扑关系中，可进行范畴化（即可语言分类的概念）的为“Disjoint”、“Touch”、“Overlap”、“Cover/Contain”、“Equal”（“相离”、“相切”、“相交”、“包含/覆盖”、“相等”）。9 交模型由于加入了点集的外域，更细的区分了拓扑关系，但仅将“相切”范畴细化为“内相切”和“外相切”，“相离”范畴细化的“内相离”和“外相离”，而“内相切”和“内相离”则为“包含”/“覆盖”所涵盖。以此为基础的扩展模型，则有效的处理不同空间数据格式、数据模型，以及概念模糊情况下的拓扑关系形式化。陈军等将一个空间目标的 Voronoi 区域作为该目标的外域，提出了“V9I 模型”。该模型能将 9 交模型中的“相离”进一步区分为“相离”和“相邻”，并把含有空洞的“包含/包含于”与“相离”区分开 [28]。在客观地理世界，要素的边界始终是有宽度的。因此 Clementini 和 Felice 研究了在宽边界下空间目标间的拓扑关系[29]。

区域连接演算（Region Connection Calculus，RCC）是以区域作为基元，以“连接”关系谓词作为原始关系来推导空间关系的一种公理化方法[30] [31]。Cohn 等认为区域定义了一种更为自然的方法来表达与定性表现有关的不确定性，而且任何物理实体所占的空间都是一个区域而不是点，因而空间关系表达应采用区域而不是点集交来描述[32]。“连接”谓词 C(x,y)表示两个区域 x 和 y 是连通的，即

这两个区域（的闭包）至少有一个公共点。通过对 C(x,y)的逻辑演算，可得到不同的 5 种和 8 种拓扑关系，即 RCC5 和 RCC8。RCC8 的概念为“相离”（Disconnected,DC）、“邻接”（Externally Connected, EC）、“部分重叠”（Partial Overlap,PO）、“相等”（Equal,EQ）、“内相切并全包含”（Tangential Proper Part, TPP）、“不内相切但全包含”（Non-Tangential Proper Part, NTPP）以及后两者的对称关系。RCC5 较 RCC8 粗糙，不考虑一个区域的边界，即不能区分 DC 和 EC、TPP 和 NTPP（图 1.1）。RCC 区分的拓扑关系与 4 交和 9 交模型具有完全相同的图形语义。

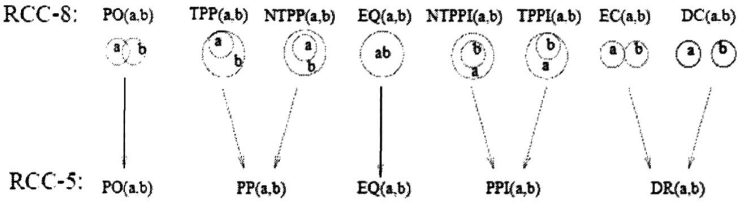


图 1.1 RCC8 与 RCC5 的映射

与宽边界类似，卵黄（Egg-Yolk）算法是描述不确定区域拓扑关系的方法[33]。将一个不确定区域的各个部分标记为卵（Egg）、卵黄（Yolk）和卵白（White），其分别对应于完整的区域、区域明确的部分以及区域完全不确定的部分。如图 1.2，（a）和（b）的拓扑类型是确定的，而（c）的拓扑类型则无法使用边界无宽度的 RCC 模型唯一确定。

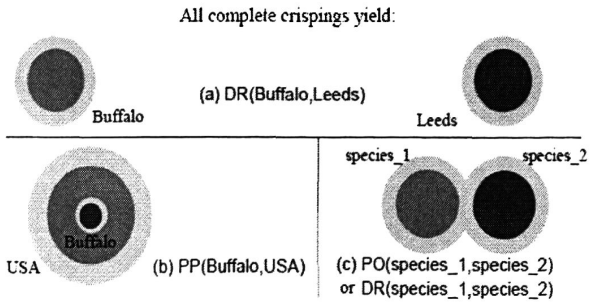
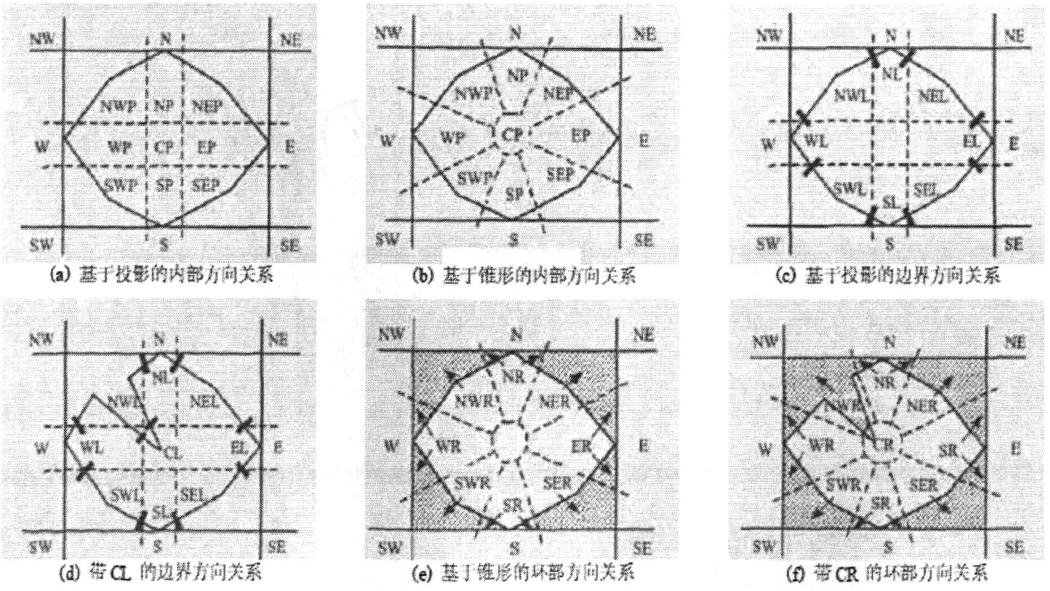


图 1.2 “Egg-yolk”模型中 RCC 类型的不确定

方向关系可看作两个空间目标间互为源目标和参考目标的互指关系，可分为定性和定量两种方法描述。定量描述是指使用方位角、象限角等精确地给出目标间的方向关系值。定性描述是用有序尺度数据概略的描述方向关系[23]，使用四方向、八方向、十六方向这些有层次的可范畴化的概念来表达。矩形模型（最小约束矩形（MBR）模型）和锥形模型及其扩展是目前理论较为完善的方向关系模型，其中锥形方向模型比较符合人们对于方向的认知习惯。此外 OPRAm

（Oriented Point Relation Algebra）[34]，DRA（Dipole Relation Algebra）[35]，DCC（Double Cross Calculus）[36]，TPCC（Ternary Point Configuration Calculus）[37]模型则以逻辑演算为目标。当以点为参考位置时，人们习惯用该点到所见目标（源目标）的连线来描述空间方向；当以线为参考位置时，人们习惯用源目标是否在线的左右边来描述空间方向；当以面为参考位置时，人们习惯用源目标在面内（或外）的哪个方位来描述空间方向[38]。不同于拓扑关系能从几何角度相对清晰的判断，方向关系的判断则需要综合目标自身的位置和形态，以及目标间的大小比例距离、拓扑关系等多个参数[23]。Goyal 使用基于投影的方向关系矩阵（Direction-relation matirx model）来描述方向，通过计算目标对象落在每个方向区域内的面积比例，可以顾及目标对象的形状和大小[39]。杜世宏等提出了细节方向关系[40]（图 1.3），其包括内部、边界和环部三种方向关系，得以表达 9 个内部方向关系，如“东部”（EP）、“东北部”（NEP）、“中部”（CP）；9 个边界方向关系，如“东部边界”（EL）、“东北部边界”（NEL）、“中部边界”（CL）；9 个环部方向关系，如“东部环”（ER）、“东北部环”（NER）、“中部环”（CR）。方向关系矩阵描述对象 MBR 以外的方向概念，细节方向关系描述对象 MBR 以内的方向概念，两者互为补充。通过把外部方向关系、内部方向关系和拓扑关系的组合，可表达诸如“一条道路从公园的东部穿过”这类更符合自然语言描述的空间关系[41]。



(a) 基于投影的内部方向关系　(b) 基于锥形的内部方向关系　(c) 基于投影的边界方向关系

(d) 带 CL 的边界方向关系　(e) 基于锥形的环部方向关系　(f) 带 CR 的环部方向关系

图 1.3 细节方向关系表达模型[40]

距离关系主要为定量化描述，包括欧氏距离、广义距离、契比雪夫距离及统计学中的斜交距离和马氏距离等[42]。使用欧氏空间的距离度量是进行绝对度量最常用的方法。而人们习惯上也用定性的概念来表达距离，如通过近、中、远、

很远等可范畴化的概念来表达[43]。出于空间认知的考虑，定性空间距离可以从不同的空间粒度上进行描述。第一个粒度等级为近（close）和远（far），这两种关系将平面分为两个区域。依据其他空间粒度等级，可以得到不同的空间距离关系体系。如分为三个等级：近（close）、适中（medium）、远（far）；或分为五个等级：很近（very close）、近（close）、相当（commensurate）、远（far）、很远（very far），这些关系的界定是主观的、模糊的。此外，空间邻近关系可以看作是一种定性距离。

同一表达模型下的拓扑关系类型或方向关系类型间存在着渐变的过程，这称为“概念邻近”（Conceptual neighbourbood）[44]。概念邻近有助于拓扑关系或方向关系的排序，支持相似度计算。图 1.4 为 RCC8 的拓扑关系概念邻域图[45]，图中节点表示拓扑关系，如果它们能够直接通过连续的形变（如扩大、缩小以及移动）相互转换，那么，它们会沿着连线方向变化。依据概念领域图中两种拓扑关系间的距离，可以计算其相似程度。方向关系的概念邻近表现在层次性上，即附方向“东北”概念与主方向“东”或“北”的概念邻近。

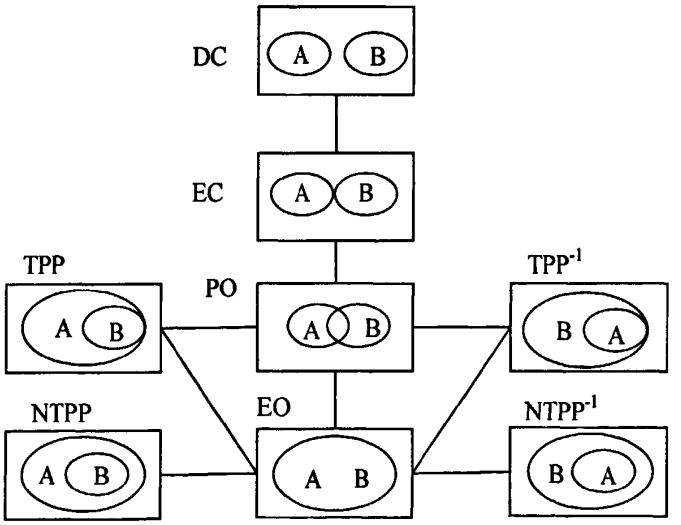


图 1.4 RCC8 的概念邻域图

1.2.2.2 空间谓词的语义映射

除了空间定性表达中使用的名词、方位词性的术语，在自然语言中还存在许多动词性的词汇，其同样具有描述空间关系的能力，形如“穿越（Cross）”、“进入（Enter）”等，称为空间谓词（Spatial Predicates）。相对定性表达中的术语而言，空间谓词具有更为丰富的空间语义，更能反映出空间语言的模糊性和空间认知的多样性。美国国家地理信息与分析中心（NCGIA）由“空间关系语言（Language of Spatial Relations）”开始的一系列课题旨在通过建立空间谓词和

GIS 计算模型间的语义映射，研究人类对于空间的认知机理[46]，提出了“自然语言空间关系（Natural-Language Spatial Relation）”的概念。

空间谓词的语义映射主要采用认知心理学的研究方式。Mark 使用认知实验方法测试了“the road cross the park”及其他相似的英文句子的几何空间语义，校正拓扑关系与自然语言的一致性程度[47]。Egenhofer 在线-面 9 交模型与其概念领域图（图 1.5）的基础上，提出了空间谓词语义量化模型，即 IAS、ITS 等 15 个线-面空间关系语义量化评价系数，并给出了英语中常见的 59 个空间谓词的量化表示[48]。这些语义量化参数一般通过实体的长度（Length）和面积（Area）的函数关系实现。为了缩小空间关系表达模型所表达的关系与人们在自然语言中所使用空间词汇之间的差异，Shariff 等用度量指标（分裂度、相近度和大体走向）结合 9 交模型的方法校正了 54 个描述线-面关系英语空间词汇与度量属性的关系[49]。Nedas 等通过定义一组反映线-线关系的度量指标（分裂率、接近率），并结合 9 交模型形式化了线-线关系的英语空间词汇[50]。类似的，Xu Jun 也定义了一组可以用来表示线状物体空间拓扑和度量关系的定量指标（角度、分裂度、距离、重合度）。在社会调查了解英语自然语言对两个线状物体特征空间关系描述的基础上，使用定量化的度量指标以及拓扑关系模型，应用数据挖掘的算法，形式化了描述线状物体空间关系的英语自然语言词汇，得到了可以表达定性的语言描述的定量化计算规则[51]。由于拉丁语系的相似性，相关研究被扩展到德语和西班牙语的跨语言比较研究[52][53]。

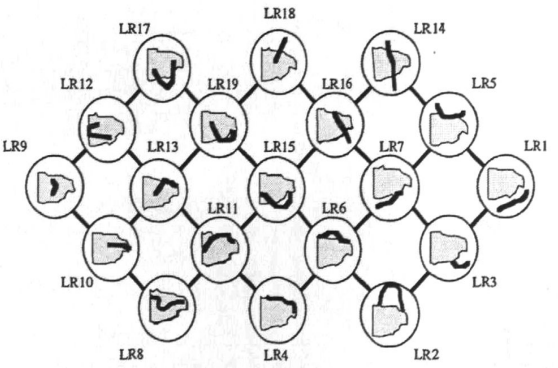


图 1.5 19 种线-面拓扑关系的概念领域图

在中文领域，许珺通过母语为英语和汉语的人群在描述线-线空间关系时使用空间谓词的双语比较，探索了不同语言和文化背景的人群对空间关系的自然语言描述的理解上的差异[54]。马林兵选取了 Egenhofer 提出的 15 个语义量化系数中的 8 个（IAS、ITS、ETS、PA、LA、ON、IC、IPA）以及扩展了 4 个新系数（APA、ALA、LHS、LVS），用于汉语空间谓词的语义量化，区分了“经过”、“横贯”、“纵贯”、“穿越”、“沿着”、“围绕”、“环绕”所描述的线-面空间关系[55]。

例如：

（1）“围绕”的语义量化公式为：

DISJOINT & APA(d)>=0.75 （1-1）

（2）“环绕”的语义量化公式为：

DISJOINT & APA(d)>=0.95 （1-2）

其中，DISJOINT 通过 9 交模型获取。而 APA（Apart Perimeter Alongness）为面域周长吻合度（图 1.6）。

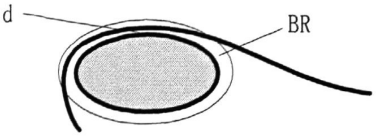


图 1.6 Apart Perimeter Alongness 示意图

系数计算方法如公式 1-1，其中公式中 L 表示线实体，BR 为面实体 R 的缓冲面，缓冲值 d 的确定取决于具体的情况，length 表示计算长度。而°、∂则是 9 交模型定义的面的内部和边界。

$$APA = \frac{length(L \cap BR^{\circ})}{length(\partial R)} \quad (1\text{-}3)$$

马林兵进而提出了一种空间关系动态性和模糊性的描述方法，该方法顾及 0 至 2 维空间目标。在空间关系语义描述中，引入语言学的“施动”和“受动”概念，将语义的产生归因于“施动者”空间目标对“受动者”空间目标的作用，将空间目标间空间关系的语义空间分解为分割度、覆盖度、接近度和环绕度，并给出了基于这些量度参数的表述和计算方法[56]。刘瑜在受限汉语的 GIS 路径重建研究中，对 GIS 路径的自然语言表述（Natural Language Representation of Path, NLRP）中使用的核心动词通过 9 交模型进行了语义映射（图 1.7）[57]。

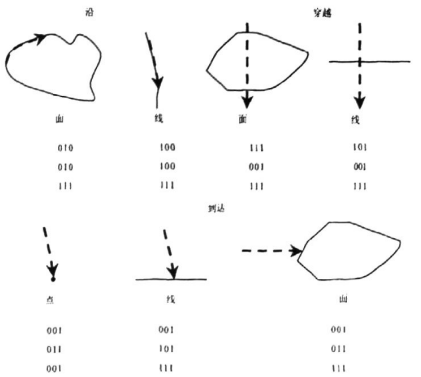


图 1.7 NLRP 核心动词的空间关系图示及 9 交模型矩阵

1.2.2.3 自然语言空间查询

处于应用层面的自然语言空间查询研究大致可分成互为关联的两个方面[58]：（1）空间查询自然语言涉及的词法、句法和语义内容的研究；（2）自然语言空间查询的空间关系图形语义及其细化的研究。其中后一方面为上两节研究的具体应用及扩展，如马林兵对汉语空间谓词的语义量化即是为自然语言空间查询服务的。邓敏利用拓扑和度量相结合的方法，描述了 28 种面目标间的拓扑关系，通过给出合理的几何度量参数及计算方法，增强了空间关系的可区分性，进而提高了空间分析的准确性[59]。陈学工提出了一组支持带孔面对象的面对象空间关系度量参数[60]。类似的研究都是通过扩展语义量化系数，细化空间关系的类型，进而更好的匹配空间查询自然语言的语义。

在空间查询的自然语言处理方面，张连蓬通过自然语言中的关键词构造模式库来映射查询语句，并对空间操作算子、空间查询语句、模式库结构等关键问题进行了讨论[61]。马林兵提出了空间信息自然语言查询接口（SINLQI），并讨论了基于 E-R 语义的词典建立、中文分词、查询文法规则及其应用领域等主要问题[62]。吴静等在正向最大匹配分词法的基础上提出来首字扩词分词法，用于对受限自然语言的查询语句进行分词处理[63]。付胜博研究了利用语义型词典 WordNet 扫描空间数据表名和字段名，建立同义词集，构造从自然语言转化为标准 SQL 语言的规则库和模板库，得以通过自然语言处理技术检索空间数据库[64]。徐爱萍分别采用中文查询语言采用的“中间栈”结构和语义查询树的中间语言结构两种方法，通过分词处理和句子结构分析获得一种中间语言结构，然后通过文法处理规则，实现查询语句向 SQL 语句转换的算法[65]。

1.2.3 中文文本中空间关系抽取研究

1.2.3.1 信息抽取

信息抽取（Information Extraction，IE）是指从文本中抽取指定的事件、事实等信息，形成结构化的数据并存入数据库，供用户查询和使用的过程[66]。广义上信息抽取处理的对象还包括声音、图像、视频等其他载体的内容，而文本也分为结构化文本（如数据库中的文本）、半结构化文本（如 Web 中的网页文本）、自由文本（合乎自然语法规则的文本），本文特指自由文本信息抽取。由于用户一般只关心有限的感兴趣的事实信息，而不关心文本意义的细微差别以及作者的写作意图等深层理解问题，因此信息抽取过程中广泛的采用浅层自然语言处理技术（Shallow Natural Language Processing Technology，SNLPT）。

信息抽取研究的开展得益于由美国国防高级研究计划委员会（DAPRA，the Defense Advanced Research Project Agency）资助的 MUC（Message Understanding

Conference）系列会议（1987-1998）的召开。MUC 显著特点不是会议本身，而在于对信息抽取系统的评测。目前推动信息抽取研究进一步发展的动力主要来自美国国家标准技术研究院（NIST）组织的自动内容抽取（Automatic Content Extraction, ACE）评测会议[67]，这项评测研究的主要内容是自动抽取新闻语料中出现的实体、关系、事件等内容，即对新闻语料中实体、关系、事件的识别与描述。与 MUC 不同，ACE 评测不针对某个具体的领域或场景，采用基于漏报（标准答案中有而系统输出中没有）和误报（标准答案中没有而系统输出中有）为基础的一套评价体系。

一个通用的信息抽取系统大致由五个步骤组成：1）词次化和标记（Tokenization and Tagging）。对输入的文档进行分段、分句后进行词性标注，对中文文档而言，词性标注与自动分词同时进行。2）句法分析（Sentence Analysis）。识别文本中的名词短语等语法结构，并进行句法分析，以识别命名实体（Named Entity）。中文的命名实体识别在 1）完成。3）抽取（Extraction）。利用与领域相关的抽取模式（Extraction Pattern）根据任务将信息抽取出来，并填入到输出模板的槽（Slot）中。4）指代合并（Merging）。主要处理待处理文本中命名实体的指代消解（Coreference Resolution）。5）模板生成（Template Generation）。根据抽取任务并结合领域知识进行对抽取的信息进行推理，生成最终的模板[68]。

信息抽取系统的构建中，最重要的是如何实现抽取模式的获取。根据获取抽取模式方式的不同，信息抽取的构建主要有两种方法，即知识工程方法（Knowledge Engineering Approach）和机器学习方法（Machine Learning Approach）[69]。知识工程方法依靠人工编写抽取模式，是系统能处理特定知识领域的信息抽取问题。这种方法要求编写抽取模板的知识工程师对该知识领域有深入的了解。机器学习的方法是利用机器学习技术让信息抽取系统通过训练文本来获取抽取模式，实现特定领域的信息抽取功能。任何对该知识领域比较熟悉的人都可以根据约定规则来标记训练文本。利用这些训练文本训练后，系统能够处理没有标记的新文本。

1.2.3.2 中文关系抽取技术

MUC 会议在最后一届（MUC-7）上首次提出了关系抽取任务（即模板关系，Template Relation）[70]。ACE 则将关系识别（RDC，Relation Detection and Recognition）表述为：探测和识别文档中特定类型的关系，并对这些抽取出的关系进行规范化表示[71]。关系抽取技术可分为以下几类：基于知识库的关系抽取；基于弱监督的关系抽取；基于特征向量的关系抽取；基于 Kernel 的关系抽取。其中以机器学习方法为研究的主流。以下按类型论述中文关系抽取的相关研究：

（1）基于知识库（模式匹配/词典）的关系抽取

邓擘等在使用模式匹配技术的基础上引入了词汇语义匹配对汉语实体关系进行抽取，并比较了一般模式匹配技术和词汇语义模式匹配技术在汉语实体关系抽取任务中的性能。实验表明，词汇语义模式匹配技术更适合于处理汉语实体关系抽取[72]。

（2）基于弱监督的关系抽取

弱监督的关系抽取是从未标注（或少量标注）的语料中学习关系抽取模式，以实现实体关系的抽取。这类方法分为两种：主动学习（Active Learning）与互训练（Co-Training）。目前研究较多的为后者，BootStrapping 是其研究的热点之一。这类方法根据最初的几个种子实例从文本中提取模板，然后将这些模板应用于文本，从而获取新的实例。重复该过程，最后获得更多的实例和模板。Ellen Riloff 将 BootStrapping 过程分为两步：在 mutual-bootstrapping 阶段挑选好的语义模板；在 meta-bootstrapping 阶段保留可靠的类型实例[73]。

姜吉发等提出了一种自举的二元关系和二元关系模板获取方法 BRPAM，该方法能够根据用户初始给出的几个种子二元关系从一个大的自由文本集合中抽取出更多的二元关系[74]。张素香等提出了一种基于 bootstrapping 的实体关系获取方法，利用由种子词和种子模板组成的知识库建立学习器，采用标量聚类的方法，通过种子模板抽取更多的与种子词语义相似的特征词，进而迭代获取更多的抽取模板[75]。陈晓颖等提出了一种 STG（slim template getter）的 bootstrapping 训练方法，该方法采用生物信息学中的序列比对技术计算上下文的语义模板，使用一定的评估机制筛选模板，有效的扩充元组以提高下一轮训练的质量。实验结果表明，STG 生成的模板不仅能覆盖大量的元组，而且正确率可达 99%[76]。陈文亮等提出了一种基于 bootstrapping 的领域词汇的自动获取方法。该方法可从大规模无标注真实语料中，自动获取领域词汇知识。实验结果显示，从人民日报语料中学习效果比专业领域语料好[77]。

（3）基于特征向量的关系抽取

基于特征向量的关系抽取基本思想是，对关系样例进行特征提取并表示为特征向量，然后选择学习算法来构建抽取模型，以实现实体关系的抽取。车万翔等以 2004 年 ACE 评测训练数据作为实验数据，使用两种基于特征向量的机器学习算法 Winnow 和 SVM 进行实体关系抽取，并指出在关系抽取时，应集中精力寻找更好的特征[78]。Zhou Guodong 等基于 SVM 模型利用词汇、语法、语义等各种知识实现了基于特征向量的关系抽取系统，取得了较好的实验效果[79]。徐芬等针对中文实体关系的特点，设计了一系列的特征，包括词、词性标注、实体和出现信息、包含关系和知网（HowNet）提供的概念信息等，以构成实体关系的上下文特征向量并使用 SVM 方法进行了中文实体关系抽取。根据分级实验的结

果，考察了各种特征集对识别性能的影响，得到下一步研究的方向[80]。

刘路等根据中文命名实体关系抽取的特点，从中文的形态学、语法及语义等几个方面选取特征并构建特征向量，然后将符合特定实体关系模板的候选命名实体对抽取出来并分为正反例。通过SVM对正反例样本进行训练，以此来判断候选命名实体对的关系类型[81]。董静等将中文实体关系划分为包含实体关系与非包含实体关系，对其采用不同的句法特征集。使用CRF进行的抽取实验表明新的特征能有效率提高汉语实体关系抽取任务的性能[82]。

基于特征向量的学习方法中，特征的选择对最终系统的性能影响较大，因此在选择了具体学习算法后，应尽量根据领域知识选择相关特征构建特性向量。而在某些应用中，同样的特征向量选择情况下，不同的学习算法性能差别不大。

（4）基于Kernel的关系抽取

基于Kernel的学习方法，最早是从SVM中被引入的[83]。在计算关系之间的距离时不再使用特征向量的点积而是使用核函数。任何核函数均在高维特征空间中隐式的计算特征向量的点积。因此，基于Kernel的方法在表示实体关系时更加灵活，即通过核函数的映射可以综合更多的知识信息。此外，由于核函数具有优良的复合特性，因此最终的实体关系距离可以由多个不同信息源的核函数复合而成。典型的核函数有子序列核（Subsequence Kernel）[84]和卷积树核（Convolution Kernel）[85]。刘克彬等应用改进的语义序列核函数，结合KNN机器学习算法构造分类器来分类并标注关系的类型，实现了基于Kernel的中文实体关系抽取系统。通过对ACE评测定义的3大类6子类实体关系的抽取，平均F值达到88%[86]。潘坤选取卷积树核实现了一个实体关系抽取的原型系统，在ACE2004上关系检测和7个关系大类抽取的F值分别取得72.5%和75.4%。此外，针对长距离的关系识别性能较低的情况，提出语篇分析树与语义关系树相结合的方法，实验证明其可行性[87]。黄瑞红等系统探讨了Kernel方法在中文关系抽取上的有效性问题：1）研究了在卷积树核中使用不同的语法树对关系抽取性能的影响；2）通过构造复合核检查了树核与平面核之间的互补效果；3）改进了最短路径依赖核，将核计算建立在原最短依赖路径的最长公共子序列上，以消除原始最短路径依赖核对依赖路径长度相同的过严要求，在ACE2007标准语料集上的实验结果表明卷积核方法对中文关系抽取有效，而最短路径依赖核对中文关系抽取效果不明显[88]。周俊生等针对中文句法分析研究现状，设计了一种能够直接利用浅层语言特性的混合谱核来描述关系实例的上下文，并给出了基于广义后缀树的高效核计算方法；然后再通过与实体核的组合生成合成核，实现了一种基于新的合成核的中文关系抽取系统。实验证明其明显优于基于特征向量的关系抽取方法[89]。

1.2.3.3 中文空间关系抽取

从信息抽取的角度，空间关系抽取是实体关系抽取的一部分。ACE 关系 Physical 大类中的 Located、Near 小类和 Part-Whole 大类中的 Geographical 小类即包含着地理空间关系的内容。对文本空间关系抽取的研究主要集中在文景转换（Text-to-scene Conversion）领域。文景转换指从一种高级描述自动生成场景，且该场景能够反映高级描述的内容。其中高级描述包括脚本、对话、故事、文本及报告等，而场景包括二维静态动态或三维静态动态。文景转换大致可分为自然语言中空间物体及其关系的信息抽取和空间物体的场景摆放两个部分。其中前一部分中空间物体间关系的抽取对场景构建起着重要的作用。

美国的 Wordseye 系统采用词性标注和句法分析工具，在对文本进行初始标注和分析的基础上，将句法分析树转换为依存结构，显示空间对象之间的空间依存关系，通过语义函数把依存结构转换为语义表示结构，以实现基于语义的英文文本空间关系抽取[90]。Reinbergerr 等通过类似“subject-verb-object”的句法结构以及领域知识分别对小语料进行空间关系的抽取，通过重建场景与真实场景的比较评测，讨论了语料及领域知识的规模对抽取结果的影响[91]。陆汝钤院士等采用基于中文扩展转移网络的句法分析方法，配合知识库进行定性空间信息规划，对用受限中文描述的童话故事进行自然语言单句理解[92]。李晗静在对简单句子级别空间关系抽取方法进行分析的基础上，提出了基于线性分类器的从真实文本中、在篇章层次上抽取空间关系方法。该方法只使用词性标注的结果就可以达到篇章层次认知概念抽取的目的，适用于句法分析尚不深入的中文空间关系抽取[93]。

方位参考点恢复是在篇章中找出方位词的参考点并进行补充，得到完整的空间表达式，其是自然语言空间语义理解中的重要问题。李晗静等提出了基于有限知识的方位参考点恢复方法，在句法分析基础上，以知网为常识库，结合有限知识识别空间表达式以及恢复方位参考点[94]。赵纪元等利用汉语空间关系表达中射体的概念，结合语料和统计数据详细分析了射体的语法、语义、结构特点以及特殊用法。在此基础上提出了基于 Winnow 算法的射体识别策略，并结合射体的语言特点，给出了一套较为完整的特征方案。实验结果显示，该方法封闭测试 F 测度可达 63.16%以上[95]。

在 GIS 领域，与地理命名实体识别（Geographical Namded Entity Recognition）以及地名解析相比，文本空间关系的抽取研究相对较少，且多与前者作为空间概念统一的进行研究。乐小虬从定义的空间语义角色入手，通过人工归纳汉语中构成空间语义角色的短语规则和句法模式，采用正向最大匹配法和有限状态机分别实现了汉语文本地理实体和空间关系的抽取，对 30 篇样本进行抽取的准确率和

召回率分别为86%和58%[96, 97]。刘元凤在对《庐山志》文本进行地理空间数据挖掘与可视化时，对古代地理描述文本中地物以及方位、拓扑关系的用词进行了分析，并通过构建数据字典进行了抽取[98]。刘瑜等通过分析带有路径表述信息的汉语文本，建立了汉语的NLRP句法模型，其是由带有空间语义的动作以及作为动作对象的地理要素构成的集合，并基于NLRP句法模型定义了受限汉语的NLRP文法，在此基础上，描述了路径重建算法PRA[57]。张雪英对自然语言空间关系研究的基本问题进行了阐述，并详细讨论了汉语中描述各种空间关系的空间词汇及其句法模式，探讨了自然语言空间关系查询请求表达的句法模式及其解析方法[99]。

1.3 问题分析

1.3.1 GIS与语言学的空间关系研究

面对作为空间认知外在表现的自然语言与地图，语言学与GIS领域各有侧重的对这两个符号系统，即文本序列和几何要素进行研究，揭示空间关系的认知机理。下面从两个方面综述相关研究：(1) GIS中自然语言空间关系的研究；(2) GIS与语言学在空间关系研究上共同的认知倾向。

(1) GIS中自然语言空间关系的研究

严格意义上讲，空间关系定性表达不属于自然语言空间关系的范畴。空间定性表达的初始目标是数理完备的对空间关系进行某种几何的定性划分，从而服务于空间定性推理。空间定性表达的一个研究倾向是更多的区分空间关系，如4交模型到9交模型，再到V9I模型。对区分出的空间关系类型能否范畴化，即赋予自然语言词汇或更为专业化的术语定义，并不关注，许多细化的空间类型没有对应的空间词汇。另一方面，可范畴化的空间关系类型使用的词汇具有GIS领域的术语性，如具有同样图形语义的RCC8中的“PO”和9交模型中的“Overlap”，在特定的表达模型和推理语境下是不能互换或由同义词替代的。因此，从人工智能中引入的QSR研究在很多情形下有“为形式化而形式化”的倾向，而忽略了地理空间以及人的空间知识表达的具体特点[100]。基于认知的空间关系定性表达则具有了一定的自然语言描述特性。如考虑了地理现象和关系不确定性的卵黄模型；细节方向关系模型类似于方位词一样，兼顾拓扑和方向的语义。

概念领域图的引入，使离散的空间关系定性表达连续化，为语义量化参数的引入提供的保证。“空间谓词的语义映射”，从另一个角度也称为“地理要素空间关系的自然语言理解”，其反映了GIS自然语言空间关系的研究始终是以图形认知为基准的。由于“点-点”、“点-线”、“点-面”的空间关系相对简单，而“线-线”、“线-面”、“面-面”关系由于地物形态的复杂性而使关系的认知和表达较为

复杂，因此研究多在后类空间关系中展开。从某种意义上讲，语义量化参数的目的也是为了区分空间关系，但其不是单纯为了图形分类，而是为了研究空间谓词与地图图形间的映射。语义量化参数的设置与选择是先验的，而研究的重点在于通过认知实验与数据挖掘获取相关量化参数的取值范围及量化公式。例如“围绕”与“环绕”的语义量化公式中，拓扑关系“DISJOINT”及语义量化参数 APA(d)是预先设置的，而“DISJOINT&APA(d)>=x”公式本身以及对“围绕”取值 0.75，对“环绕”取值 0.95 则应是通过社会调查获取的，其反映了人类对 GIS 空间关系的自然语言理解背后的认知规律。但这类研究具有认知心理学研究方式所具有的通病，即人文参数过多与对解释的依赖。目前只有 Shariff 和 Xu Jun 等人通过调查方法计算出具体提出的相关语义量化参数的取值范围。

在应用层面，空间谓词的语义映射的研究，为自然语言空间查询中空间词汇到空间关系图形的映射提供了保证。由于查询中空间谓词是语义受限的，因此量化参数的取值可以由某个原型进行确定，这样量化参数及量化公式更多的是考虑查询时的效率问题，如地理要素的数据量及关系判定的运算量。

（2）语言学与 GIS 在空间关系研究上共同的认知倾向

在语言学尤其是认知语言学领域，方位词及空间关系描述的研究目标是其语言现象背后的空间认知过程。阅读者会将一个表述空间关系的语句参照场景内包含的丰富信息简化为简单的意象图式结构（image-schematic strutrue，以下简称意象图式），这一过程称为图式化（Schematization）[101]。意象图式是存在于人类的感知和身体运作程序中一种反复出现的动态模式，具有相对简单的结构以及各种空间方位关系，前者如容器图式、路径图式等，后者如上-下、前-后、部分-整体、中心-边缘等[102]。认知语言学认为，隐喻不是一种简单的语言现象，而是一种认知方式。当一个意象图式通过隐喻被投射到一个非空间概念上时，该图式的内在逻辑在投射过程中被保留，成为非空间的目标概念的抽象逻辑，这使得我们可以运用空间思维来思考和理解非空间概念。因此，作为认知语言学最活跃的空间隐喻研究，在致力于揭示空间隐喻在不同语言和文化中的普遍性的同时，也诠释了具有人类共性的空间认知过程。

GIS 构建过程是从现实世界（Real World）到最终的以几何要素为基础的要素集合世界（Feature Collection World）[103]。在这一过程的抽象阶段，概念世界是人类自然语言的世界。而真实世界与概念世界间的交互称为认知接口（Epistenmic Interface）。不同的描述者对同一地物的认知及其描述不同，这对后续的抽象过程影响巨大，是 GIS 语义冲突的根本原因[104]。而在概念世界与地理空间世界间的选择接口也是一个图式化的过程。概念世界的复杂性在地理空间世界中用简单的、浅显的抽象来代替。在最终的要素集合世界符号化为地图后，对

GIS 空间关系的认知也依赖于相应的意象图式及抽象概念。语言是认知的窗口，对 GIS 空间关系的认知结果可通过（也只能通过）自然语言得以表述。

可见，认知语言学通过空间方位的隐喻研究空间关系的文本符号认知，GIS 则通过几何要素空间关系的自然语言理解研究空间关系的地图符号认知（图 1.8）。

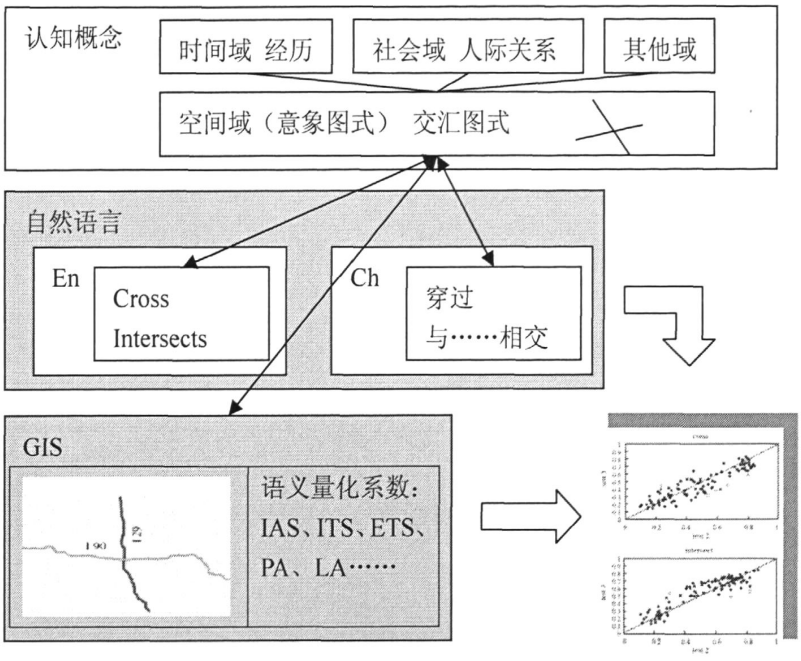


图 1.8 语言学与 GIS 在空间关系研究上共同的认知倾向

语言学方位关系以及 GIS 自然语言空间关系的研究成果是本研究的基础。而目前 GIS 自然语言空间关系的研究对象是选定的自然语言空间词汇以及有限文法的空间关系描述，同时其以空间语义的映射为最终结果。在进一步针对自然语言空间关系的研究中，我们认为有必要将“描述-识别-解析”这一自然语言空间关系研究的整体框架分为两阶段，本文即针对“描述-识别”这一阶段中文文本空间关系的抽取方法进行研究。

1.3.2 中文文本中空间关系的抽取

（1）GIS 自然语言中查询解析的受限性

GIS 自然语言空间查询继承了数据库自然语言查询的特点，在词汇、句型、语义、语用等方面都是受限的。词汇受限表现在查询中使用的名词、动词等实词必须与空间数据库内容相关；句型受限表现在查询使用的句型以祈使句和疑问句为主，一般没有兼语句、连动句等难以处理的句型；语义受限表现在查询中涉及的概念只限于已有空间数据的概念模式；语用受限表现在查询的目的是从专题数

据中获取信息，因此在查询语句相对简单的情况下，仍能得出用户的查询意图。空间查询自然语言的处理普遍采用知识库的规则方法。在空间关系涉及的查询语句的形式化处理中，产生式文法多选择上下文无关文法，并在获取文法分析树的基础上，使用特征集进行语义分析。

（2）中文文本中关系抽取方法

中文关系抽取技术研究中，基于知识库与弱监督的方法研究由于不依赖熟语料，因此实验数据多为网络文本；而基于特征向量和 kernel 的研究则多在 ACE 的语料上进行。目前研究的热点是基于 Kernel 的关系抽取方法，尤其是在英文中实证有效的新核函数在中文关系抽取上的迁移。而对基于特征向量的关系抽取研究则较为成熟，多为在特征选择、样本处理、学习算法比较等方面的优化工作。关系抽取面临以下二个困难[105]：1）特定领域标引数据集的获取（即语料库的构建）。关系抽取多采用基于模式匹配和机器学习的算法来判断关系是否存在以及关系的类型，而使用这些算法的先决条件是需要预先通过对一个特定领域手工标引的数据集进行学习以获取领域内关系类型的各项特征。此外，这些用于学习的数据集的大小和标引质量都会影响到关系抽取的效果。2）模式的获取。基于模式匹配原理的关系抽取方法在很多关系抽取系统中得到了应用。然而，定制特定领域的恰当的关系模式存在较大困难。

在文景转换领域的文本空间关系抽取中，考虑到后续场景的生成，因此较 ACE 的抽取内容而言，需抽取更多的空间信息。对于空间关系中方位参考点及射体的恢复，在英语句法分析实用的基础上，多通过依存结构进行抽取。而中文多在词性标注的基础上使用有限知识或线性分类器进行获取。由于文景转换处理的描述对象多为小范围的场景，因此对大尺度的地理空间关系的描述缺少相应的研究。

（3）中文文本空间关系抽取的应用

无论信息抽取自文本，还是声音、图形等其他表意符号，其最终目的在于有效的格式化，而服务于信息检索。在地理信息检索领域，目前主要是以地理命名实体为核心进行时空信息的组织与检索，如 ADL[106]、Getty[107] 等各种地名词典，空间关系仅作为属性设置。地名解析即是通过地名词典等地理知识库，将文本中的地理命名实体与其时空位置参考进行关联的过程[108]。这一过程沿承传统地址编码（Geocoding）的定义，也称为 GeoReferencing[109]。从信息载体的角度看，即文本获取的其描述对象的时空参考，而时空模型则通过链接文本获取了更广泛的语义。在此基础上，地理信息检索得以在自由文本与时空模型间一致的进行信息的组织与检索。

句子或篇章层面的空间方位关系抽取的引入，首先为地名解析以及地物的空

间定位提供更多的约束信息。MetaCarta 的 GTS 根据地名的共现进行建模，从而获取某一地名的空间参考置信度。直观的看，在适当的上下文窗口中，地名“新街口”与“南京”共现的频度如果高于与“北京”共现的频度，则其南京新街口的空间参考置信度要高。在地名解析中，目前多通过这种地理命名实体间的文本或空间距离建立统计约束模型[110][108]，通过抽取的空间方位关系，可以构建更为准确的约束模型用以地名解析。此外，在优化地名解析模型的同时，除了原先由空间参考数据计算获取的空间关系，文本描述的空间关系也得以同等级别的引入地理知识库，而不再仅是地名的一个属性。这样即有望构建以空间关系为组织核心的地理知识库，形成新的应用模式。简单用例如下：当用户通过文本或图形查询一个地标（LandMark）周围的兴趣点（Point of Interests，POI）时，在地理知识库中即可由空间参考数据做缓冲区获得查询结果，也可直接由空间方位关系类型为 EC（“附近”）的结点返回结果。而通过面向网络文本的统计模型进行 POI 的获取，较传统方法在实时性与置信度上都有更好的性能[111]。

综上可见，GIS 领域尚缺少通过引入自然语言处理技术，对中文文本空间关系进行信息抽取的专题研究。而通过借鉴文景转换等领域成熟的中文关系抽取技术，实现文本中地理空间关系的有效抽取，是地理信息检索和 GIS 自然语言查询等应用研究的迫切需要。

1.4 研究内容与技术路线

1.4.1 研究目标

针对中文文本中空间方位关系“描述-识别”阶段的研究需要，分析中文地理空间方位关系的描述特征，构建面对信息抽取的空间方位关系表达；设计空间方位关系标注集及标注规范，进行空间方位关系的语料标注；借鉴中文信息处理领域的研究范式与技术，研究基于知识库规则与基于机器学习的中文文本空间方位关系抽取方法，通过评测分析其适用性。本文的研究将为中文文本空间方位关系“识别-解析”阶段的研究做好准备。

1.4.2 研究内容

（1）中文文本中空间方位关系的表达及语料标注

在认知语言学的空间方位参照框架及中文空间方位关系描述特征的基础上，通过文本与地图两类符号系统在空间关系表达上的对比，分析面向信息抽取的中文文本空间方位关系的表达。以此设计标注集及标注规范，选取空间方位关系语料进行标注。分析语料中的地理空间方位关系描述特性，并扩展标注集及标注规范。

（2）基于空间词汇的空间方位关系规则抽取

采用自然语言处理中“理性主义”的研究范式，研究以空间词汇为核心的基于规则的空间方位关系抽取方法。首先使用 BootStrapping 迭代获取空间词汇，利用语义型词典设置语义参数。其次，通过文本序列比对及模板聚类进行空间方位关系抽取模板的归纳。最后，通过构建空间词汇词典，结合抽取模板及匹配规则，实现基于规则的中文文本空间方位关系的抽取方法。

（3）基于支持向量机的空间方位关系抽取

采用自然语言处理中“经验主义”的研究范式，研究通过统计学习进行空间方位关系抽取的方法。选用在关系抽取中表现最佳的支持向量机作为学习机，通过设计体现空间方位关系描述的特征向量，构建抽取模型，并对其推广能力进行评测。

1.4.3 技术路线

针对本文的研究内容，拟采用以下技术路线（图 1.9）：

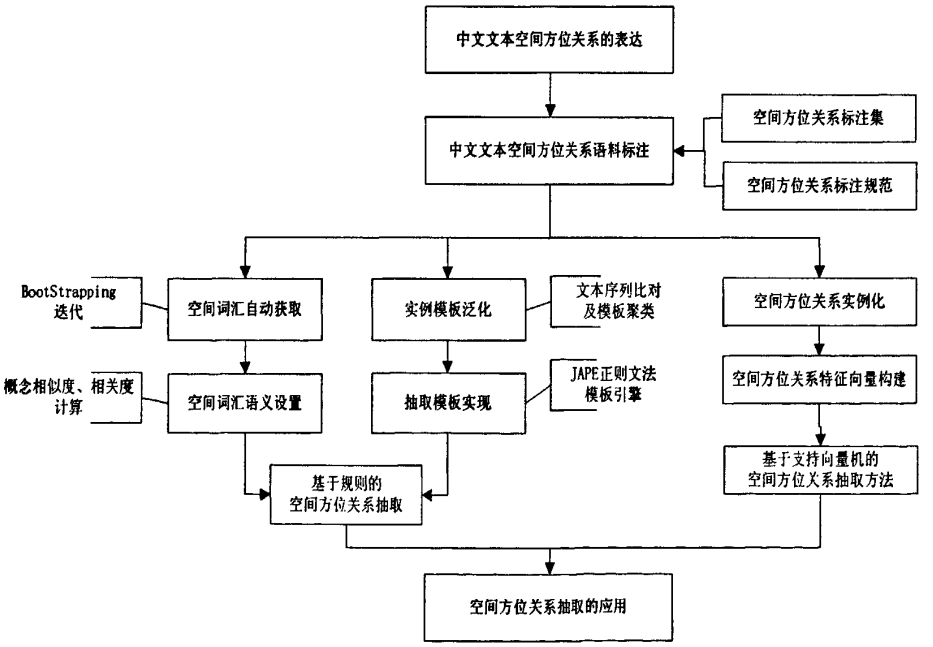


图 1.9 技术路线图

1.5 论文组织

论文的章节安排如下：

第一章 绪论。综述在语言学方位关系以及 GIS 自然语言空间关系的领域与本文相关的研究，阐述中文文本空间方位关系抽取方法的研究基础，以及本研究在“描述-识别-映射”这一自然语言空间关系解析框架中的阶段目标。综述并分

析作为本文技术支撑的中文文本空间信息抽取的相关研究。

第二章 中文文本空间关系的描述与表达。在分析认知语言学的空间方位参照框架及汉语空间方位关系描述特征的基础上，通过文本与地图两类符号系统在空间关系表达上的对比，阐述了本文对空间方位关系表达层次中定位。通过表达式、分类以及类型术语的判断三个角度，分析面向信息抽取的中文文本空间方位关系的表达。

第三章 空间方位关系的语料标注与分析。设计标注集及标注规范，选取空间方位关系语料进行标注。分析语料中的地理空间方位关系描述特性，并扩展标注集及标注规范。

第四章 基于规则的空间方位关系抽取方法。采用自然语言处理中“理性主义”的研究范式，通过构建空间词汇词典，结合抽取模板及匹配规则，实现基于规则的中文文本空间方位关系的抽取方法。

第五章 基于支持向量机的空间方位关系抽取方法。采用自然语言处理中“经验主义”的研究范式，研究通过统计学习进行空间方位关系抽取的方法。选用在通用关系抽取中表现最佳的支持向量机作为学习机，通过设计体现空间方位关系描述的特征向量，构建学习模型，并对其推广能力进行评测。

第六章 空间方位关系抽取的应用。简要分析了空间方位关系抽取在地理知识库构建和场景重建中的应用模式，通过文本驱动的地理知识库构建以及空间方位关系的图形重建演示了相关的应用场景。

第七章 结束语。总结文章研究工作的主要结论、成果及创新点，并就下一步工作提出展望。

第2章 中文文本中空间关系的描述与表达

2.1 空间方位关系的描述

Talmy 率先将格式塔心理学提出的图形（Figure）和背景（Ground）原则引入认知语言学研究[112]。他认为知觉场始终被分为图形和背景两部分。图形和背景具有不同的本质特征和联想特征。图形具有未知的、需要确定的时空特征，而背景则具有可以标明图形之未知性的已知特征。图形的联想特征包括可动、较小、较简单、较晚进入场景或意识、受到较多关注、凸显性强、依赖性强等，而背景则具有较固定、较大、较复杂、较熟悉、受到较少关注、背景性强、确定性强、独立性强等联想特征[113]。Langacker 使用“射体”（Trajector）和“界标”（Landmark）这对术语来表达相同的含义。语言中的空间关系指射体（图形）和界标（背景）之间随着时间推移而形成的存在或位移关系[114]。

由于“方位”（Orientation）这一术语常使用于语言学领域，其具有比“方向”更大的外延，同时语言中的空间关系一个主要的描述要素即是“方位词”，其可以同时描述方向关系和拓扑关系。因此，为了区分于 GIS 中的空间关系——拓扑、方向、距离关系，本文使用“空间方位关系”来特指中文文本中的地理空间关系。此外，“语音”作为自然语言的物理特性，在书面形式中表现为文字的“形”，而在口头表述中表现为声音特性。声音和图形作为语言表述的物质材料，存在一定的差异。如前者是线性的，后者是二维甚至三维的；前者在音高、音重、音长、音质上的不同也具有相应的语义。本文将研究对象限定在文本中的空间关系，以区分于语音导航等领域的空间关系描述。

2.1.1 空间方位参照体系与参照框架

2.1.1.1 空间方位参照体系

不同语言的人所面对的物理空间相同，在同一观察点上对同一三维空间所感知的认知空间也相同，但是将该认知空间概念化的过程和方式不同，从而形成了不同的空间方位参照体系。一般认为，空间方位参照体系由方位词、叙述者、观察点、方位参照点四个要素组成。方位参照是叙述者选择观察点，利用方位词跟相关的方向参照点、位置参照点的关系确定空间或时间的方位辖域[3]。

（1）方位词

方位词主要描述了界标/射体间的方向、拓扑关系。现代汉语方位词可分为五组：

A）水平方向： 前、后、左、右；

B）水平方向： 东、南、西、北；

C）垂直方向： 上、下；

D）辐轴方向： 里（内、中）、外；

E）泛方向：旁、旁边、间、中、中间、附近、周围。

A 组、B 组都表示水平方向，A 组反映了人类对自身和周围相对水平空间关系的认识，它们以语境中的人、朝向、拟人化的物为方向参照点，人面对的方向、物体的朝向或前进方向被视为前，背靠或背离的方向被视为后，靠左手一侧被视为左，靠右手一侧被视为右。B 组反映了人类对宇宙中的绝对水平空间关系的认识。它们以宇宙中确定的点为方位参照点，以日出、日落的方向定东、西，以南极、北极的方向定南、北。C 组反映了人类对地球引力的认识，以顺应地球引力的方向为下、以背离地球引力的方向为上。D 组反映了人类对空间范围、界限关系的认识，以向心、封闭的方向为里，以离心、开放的方向为外。E 组反映了人类对物体的相对位置距离关系的认识，为泛方向[3]。

（2）叙述者

叙述者是说话者或写作者。说话或写作既可以采取第一人称的叙述方式，也可以采取第三人称的叙述方式。

（3）观察点

观察点是叙述者在表达方位参照时所选择的心理视点，即所预设的观察者的位置和角度。以第一人称叙述时，叙述者作为当事人选择自身所处的位置和角度为观察点，称为“自观”；以第三人称叙述时，叙述者作为局外人以旁观者的立场和角度为观察点，称为“旁观”；无论是第一人称的叙述方式还是第三人称的叙述方式，叙述者还可以以叙述中出现的或假设存在的叙述者以外的当事人所处的位置和角度为观察点，可称为“他观”。

（4）方向参照点

方位参照点有两种：一种是位置参照点；一种是方向参照点。位置参考点指方位参照定位时选择的参照点，方向参照点指定向时选择的参照点。

2.1.1.2 空间方位参照框架

从认知语言学的角度，人类主观识解空间关系，进而使用界标、射体和方位词进行描述的过程，其自身具有的内在结构，称为空间参照框架（Spatial Frames of Reference）。空间参照框架是一种有界空间域[115]。常用的二元对立分类为：1）相对空间与绝对空间参照框架；2）以自我为中心的与以他人为中心的参照框架；3）以观察者为中心的与以物体为中心的参照框架；4）受方向约束的与不受方向约束的参照框架；5）指示性与内在性参照框架。Levinson 在对前人的有关研究进行概括和归纳的基础上，抛弃了常用的二元对立的分类方法，从认知角度提出

了相对参照框架（Relative Frame of Reference）、内在参照框架（Intrinsic Frame of Reference）和绝对参照框架（Absolute Frame of Reference）的分类[116]。

相对参照框架涉及 3 个要素，即观察者（V）、射体（OR）和界标（OL）。这是从观察者的视角出发来识解 OL 和 OR 之间的空间关系。如“图书馆的左边是学生宿舍”。虽然 V 是隐含的，但 OL（“图书馆”）和 OR（“学生宿舍”）的空间关系“左”是由 V 的观察视角确定的（图 2.1）。

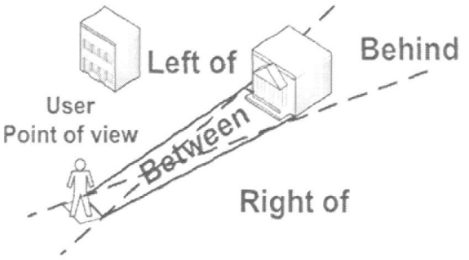


图 2.1 相对空间参照框架[117]

内在参照框架涉及 OL 和 OR 两个要素。根据 OR 的“前/后/左/右/里/外”等本质特征来识解 OL 和 OR 间的空间关系。如“图书馆前的小路”。这里的“前”由“图书馆”这一建筑的内在意象图式结构特征（inherent image-schematic structure）所确定的，而不以 V 的观察视角转移（图 2.2）。而由河流、道路朝向产生的“左岸”，“上/下游”也属于该参照框架。

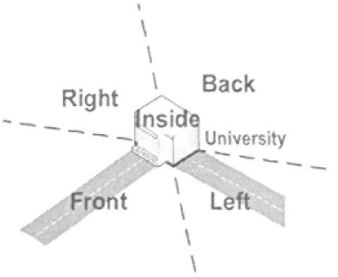


图 2.2 内在空间参照框架[117]

绝对参照框架是地理空间关系描述中最常用的参照框架，其涉及 OL 和 OR 两个要素。通过空间基本方向（Cardinal Directions），如“东/南/西/北”为基准来判断 OL 和 OR 之间的空间关系（图 2.3）。

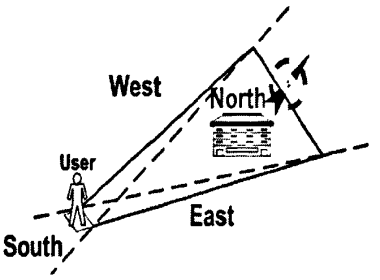


图 2.3 外在空间参照框架[117]

有时，上述三种参照框架依次也称为：以观察者为中心的参照系、以物体为中心的参照系、以环境为中心的参照系。可见，在这几类空间参照框架中，观察者 V、界标物 OL 和环境（environment）都可用来对事物的空间方位实施定位。此外，在绝大部分语言中，空间参照框架都是语法化（grammaticalization）和词汇化（lexicalization）的结果[118]。

2.1.2 空间方位关系的描述特征

汉语空间方位关系的描述特征表现在词汇与语法两个层面：

2.1.2.1 空间词汇

空间方位关系描述通过一系列的自然语言的词汇来完成，这些词汇称为“空间词汇”。空间实体的类型、形状、大小、比例尺，以及叙述者的文化、教育背景等，都会影响其对空间词汇的选择[49]。汉语中空间词汇从词性的角度，可粗略分为三类：

（1）方位词

方位词是空间方位参照体系的组成部分，通常和方位介词一起使用，形如“在……东面”。方位词和方位介词都为闭合词。但通过方位词自身的组合以及与其他词的组合，可形成更为丰富的方位描述，如“中南部”。除辐轴方向与泛方向外，方位词主要描述方向关系，并通过附加名词来隐含的描述拓扑关系，如“南侧”多为拓扑相切，“南部”多为拓扑包含。

（2）动（谓）词

即上述的空间谓词。其主要描述拓扑关系，也可以与方位词结合附加描述方向关系，如“自南流入”。在空间方位关系的描述中，空间谓词语义最为丰富，不仅依赖地理命名实体的类型，还依赖其他空间词汇，如“流经……附近”。

（3）名词

指与地理空间语义密切相关的名词，如“上游”、“边境”、“分水岭”。这些词与地理命名实体的类型具有相关性，常通过描述实体的形态来表述空间方位关系。一般情况下，反映实体间的整体-部分关系。

此外，人们在描述方位关系时常使用了类比、拟人等方式。如使用表盘时间度来表示方向的“十二点钟方向”；地图视角的“上北下南左西右东”；拟人化的“半山腰”等。

2.1.2.2 文法模式

汉语句法结构较为灵活，在空间方位关系的描述中，主要有以下四个特征：

（1）省略

省略是语言中的普遍现象。而在文本描述中，由于阅读的可回溯性，省略更是被大量使用，尤其是在辞典等要求描述简练的文本中（图 2.4）。

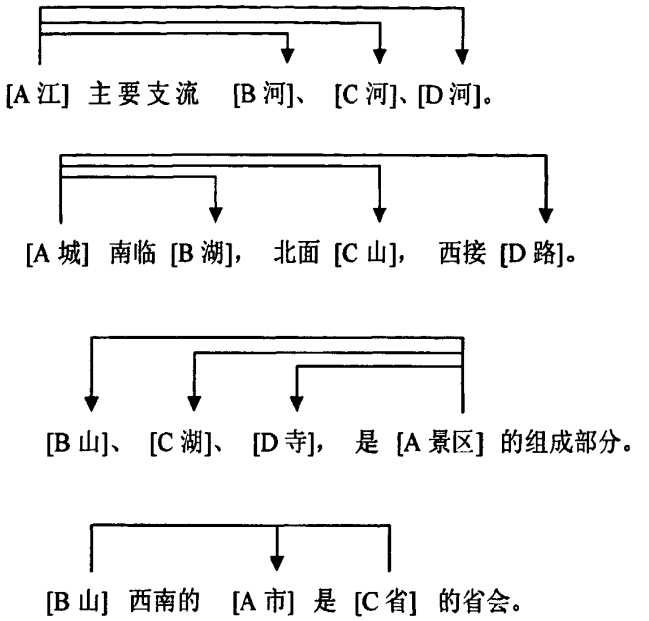


图 2.4 射体和界标一对多的省略情况

（2）空间焦点转移

空间焦点是阅读者在空间上关注的当前一个或几个实体，以及它或它们的空间位置[119]。如果阅读者在通过文本序列的阅读构建一组射体/界标的空间方位关系的过程中，需临时构建另一组空间方位关系，则称为空间焦点转移。

例 2-1

[青龙镇]位于[上海市]西部[青浦县]东北境。

例 2-1 中，阅读者在构建“青浦县”/“青龙镇”这组方位关系过程中，需构建“上海市”/“青浦县”这组方位关系。“青浦县”在前一关系中是界标，在后一关系中是射体。临时转移的空间焦点常用来对外部空间焦点中的某个实体进行约束。

（3）地理命名实体的组合

在实际的空间方位关系描述中，存在多个实体参与的关系，典型如三元关系，即“A 在 B 和 C 之间”。依据射体/界标理论，可将“B、C”作为一个组合，而在整体上看作界标。而中文的“之间”与“Between”不仅相同，可以是多者之间。

例 2-2

A） [白洋淀]位于[安新]、[高阳]、[任丘]、[雄县]等县境。

B） [大庆市]介于[哈尔滨]和[齐齐哈尔市]之间。

C） [大娄山]是[赤水河]与[乌江水系]的分水岭。

例 2-2 中，A）中“安新”、“高阳”、“任丘”、“雄县”形成一个实体组合，注意其与省略的差别，即组合涉及的地理命名实体需同时存在，而“[白洋淀]位于[高阳]”与整句的含义是不一致的。C）中空间词汇“分水岭”更进一步的指示了需类型为水系的两个实体组合为界标，而类型为山脉的实体作为射体。

（4）路径描述

自然语言中路径描述一般指导航路径，由按时间顺序排列的一系列地物与动作组成，用以描述行进的路线，包括起点、重定向、过程和终点四个部分[120]。而当静态的描述山体走向、河流流向等自然地物的形态时，也常使用类似的描述，如例 2-3。

例 2-3

[无定河]发源于[陕西省定边县白于山]北坡，上源称[红柳河]，东北流至[内蒙古乌审旗巴图湾]，折向东流复入[陕西境]，经[横山县]至[榆林县]的[鱼河堡]折向东南，至[清涧县]入[黄河]。

路径描述常使用省略，不同的是其具有顺序，不能前后置换。此外简洁的路径描述中常包含了空间焦点转移的情况（图 2.5）。

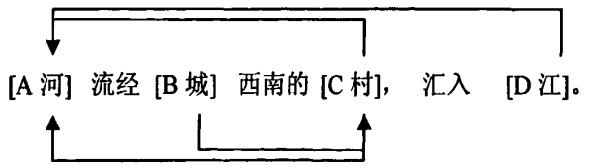


图 2.5 空间焦点转移与路径描述重叠的情况

空间方位关系的描述中还需符合图形-背景理论，以及空间关系层次性和尺度相应原则[121]。形如“长江在亭子边”即不符合图形-背景理论，而“亭子在南京”则跨越多个层次而不合乎常识。

2.2 空间方位关系的表达

2.2.1 形式化与符号三角形理论

从人工智能（Artificial Intelligence，AI）的研究角度，形式化有广义和狭义两层意思。从广义的角度，一切能被感知的事物，以及被意识到的精神活动，都存在符号和其他某种形式在意识中形成对应物（图形、声音、文字等符号），这些东西总是表现为一定的形式，而人类的思维活动就是通过这种素材的组织而完成[122]。广义形式化强调可描述性。而狭义的形式化则是由逻辑学规定的，将某一过程形式化，也就是建立一种算法，将这一过程通过建模表达出来。狭义形式化强调的即为可计算性。由此可见，非受限的自然语言、绘画、音乐可以看作是广义的形式化，而由视觉变量约束的地图、受限语言、数理模型则是狭义的形式化。其中，受限语言即包括计算机编程语言等形式语言，也包括自然语言查询中的受限自然语言。本文针对自然语言等广义形式化采用“描述/表述”（Description）这一术语，针对可计算模型的狭义形式化采用“表达”（Representation）这一术语。从信息处理的角度，所有广义形式化的描述都必须通过狭义形式化的表达，才能进行进一步的数理计算。

无论是广义或狭义形式化，作为符号表意系统，都遵循符号三角形理论（Meaning Triangle）。该理论用以解释在任何形式化语言（受限的或自然的）中，符号（Symbol）如何被人类理解。也就是说，符号如何触发人产生概念（Concept），而这个概念指向真实世界的实体（Object）[123]。

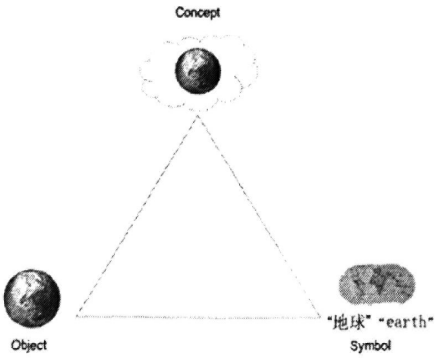


图 2.6 符号三角形理论

如图 2.6，文本符号“地球”、“Earth”以及世界地图这一地图符号，对于一个具有适当上下文信息的阅读者来说，这两类符号都能触发阅读者产生“地球”的概念，进而联想到诸如“世界”、“环境”等概念。“符号-概念”的对应关系即

为经典的“能指（Sense）-所指（Reference）”关系。而任何符号都是对实体的一种近似。

2.2.2 文本符号与地图符号

从认知论的角度，语言是人类所独有的一种种属属性，它似乎是人类心理过程唯一的窗口[21]。有人退一步但坚持认为，我们只有通过语言的研究才能把握思想。而地图作为真实世界最重要的形式化方式，与文字相比而言，是一种模型或图形符号，与图像相似。但与图像相比，地图又有其特性，即数学基础的严密性、表达内容与范围的选择性和符号的强调性等。

从方法论的角度，语言学的分析方法不仅仅应用于语言研究本身，而是成为许多现代科学的研究范式。杜清运将空间信息领域所发展的语言学模型分为两类：一类即为 NCGIA 的“空间关系语言”，其是从自然语言中的空间概念分析入手，以改善目前空间信息概念模型上的不足。一类是将空间信息看成是一种空间形式的特殊语言形式，并利用语言学的语法、语义和语用研究手段来分析和认识空间信息的内部结构[124]。例如将现代语言学中的语法结构观点应用于 GIS 空间关系，诠释空间关系的语音变体、语法变体和语义变体。

由于作为广义形式化的语言是空间认知的窗口，因此在空间关系的文本描述与地图表达的研究中，常从认知的层面进行比较，强调空间关系的地图表达需符合文本描述的常识习惯。如图形-背景理论在地图符号的视觉变量与语言学的方位参照框架中的一致性；空间关系描述的层次性和尺度相应原则等。另一方面，空间关系的文本描述与地图表达有着符号形式上的本质区别：语言是一种序列性的符号，不同语言的人会使用不同的扫描方式来描述空间关系，如中文“桌子上的书”与英文“book on the table”在界标/射体先后顺序上的不同。而地图作为图形符号则可以在同一时间以二维或三维的方式展开和并列所有空间关系，界标/射体是同时呈现的。而这种形式上区别进而引发了更深层的不同。对于“The chair is in the corner of the room（椅子在屋角）”这一空间方位关系，可以说是模糊的（Vague），但却是确定的（Certainty）。但当在地图上绘出该空间关系时，由于几何图形的约束，位置总是精确的（Precision），但却是不确定的（Uncertainty）[125]。在日常生活中，我们需要语言的这种模糊描述，在共有的语境下传递交流双方足以确定的信息。而在几何计算模型的表达中，数理的精确性决定了对真实世界表达的不确定。

2.2.3 空间方位关系的表达层次

自然语言描述的空间方位关系作为一种广义形式化，必须通过狭义形式化的表达，才能格式化而得以数理计算。

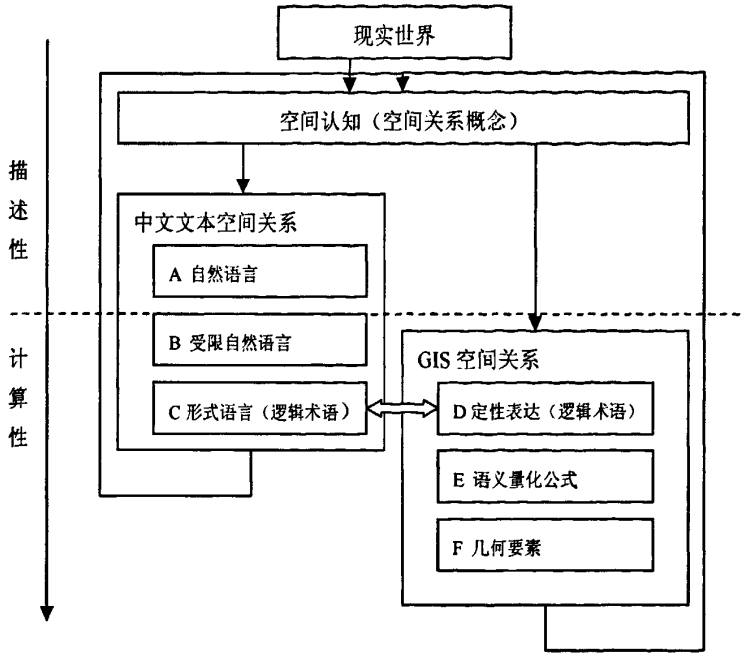


图 2.7 空间方位关系表达的层次

如图 2.7 所示，人类对空间关系的认知可以直接由现实世界获得，也可以从文本符号与地图符号获得。空间方位关系的形式化表达可有两个层次：

（1）从自然语言（A）到语义量化公式（E）的映射。这也是空间谓词的语义映射的研究路线：通过语义量化参数的设置与认知实验，统计获取相关量化参数的取值范围，从而对空间关系的自然语言理解进行分析。当应用于自然语言空间查询时，在应用上下文的约束下，自然语言（A）成为受限自然语言（B），而通过语义参数取值范围的人为设定，语义量化公式(E)得以精确为几何要素(F)。这样，在查询时则为受限自然语言（B）到几何要素（F）的映射。而在空间数据库的具体运算中，则是形式语言（C）与几何要素（F）间在数理层面上的映射。

这里自然语言（A）、受限自然语言（B）、形式语言（C）在符号能指上完全的一样的，不同在于其所处的应用语境，从而在所指上逐步精确，得以计算。比如“附近/Near”一词在日常对话问询、自然语言空间查询接口、空间数据查询语句中的不同。

（2）从自然语言（A）到形式语言（C）的映射。这是信息抽取/自然语言处理领域的研究路线：将篇章分为句子，将句子分为词汇。再由词汇构建句子结构，由句子结构获取语义角色，从而形成一个句子上下文无关的意义，即其逻辑形式。最后根据上下文形成最终的意义，一般假设一阶谓词演算(FOPC, first-order

predicate calculus）即为精确的知识表示。自然语言处理领域将“语义”限定在一阶谓词演算的表达能力范围内。

对于“平型关位于山西省繁峙县东北边境”这一空间关系描述的句子而言，最终的逻辑表达可能为：

Location(NE_BORDER,(OL:Fanzhixian,OR:Pingxingguan))

&Location(SHEN_XIAN,(OL:Fanzhixian,OR:Shanxisheng)) （2-1）

其中“&”、“:”等演算符号的含义是统一的，而“NE_BORDER”、“Location”等谓词的名称及含义则依赖具体的抽取目标而设置。习惯上，常用具有一定描述性的词汇（如“NE_BORDER”）来精确化自然语言词汇（如“东北边境”、“东北边”等）。而使用英文则是因为其是计算逻辑的通用语言。本文将形式语言中精确为某个特定语境中的词汇称为“术语”（Terminology）。

由 GIS 空间关系自然语言理解的研究方法可知，从空间认知的角度，空间方位关系形式化的第一层次似乎可以不包含第二层次。但具体到自然语言空间查询等应用层面，形式化的第一层次则需要第二层的自然语言处理技术的支撑，从而通过逻辑术语的精确化来进行 GIS 空间关系的映射。而在中文文本空间关系与 GIS 空间关系两个能指形式不同的符号系统里，形式语言（C）和定性表达（D）是唯一的接口。

空间方位关系形式化的第一个层次是本课题的最终目标，即针对自然语言空间关系词汇及语法到 GIS 空间关系语义的“描述-识别-解析”框架。而为了突破目前研究对象仅为选定的空间词汇以及有限文法的局限，我们首先需对空间方位关系形式化的第二个层次进行研究，即为“描述-识别”。而与信息抽取领域不同，我们对空间关系术语的设置需参考 GIS 空间关系中定性表达（D）的内容，以满足下一阶段的研究需要。

2.2.4 空间方位关系的语言结构

空间方位关系表达的第二层次的主要任务就是将中文文本中的空间关系结构化，映射为形式语言。从自然语言处理的角度，空间方位关系的语言结构可通过文法分析与浅层语义解析两个阶段获得。而在实际信息抽取的应用中，未必需要在完整分析出文法结构和论旨角色的情况下，才能确定空间方位关系。

（1）乔姆斯基文法与产生式规则

没有一种自然语言可以对其进行足够精确的表达，但为了知识表达而假设为某种形式语言时，即可以进行精确的、数理化的特征表示。形式语言是一种为特定目标设计的人工语言。Chomsky 将其定义为按一定规律构成的句子或符号串的有限或无限的集合[55]。任何一种形式语言的文法 G 可定义为如下四元组：

$$G=(VN,\ VT,\ S,\ P) \qquad (2\text{-}2)$$

其中 VN 是非终结符的集合，VT 是终结符的集合，VN 和 VT 不相交。S 是起始符，P 为产生式规则集。给定一个文法 G，可以从起始符 S 开始，应用产生式规则推导出语言 L(G)，这一过程称为生成（Generative）。由相应的产生式规则推导可获得句法分析树，推导过程即可以自顶向下，也可以自底向上。

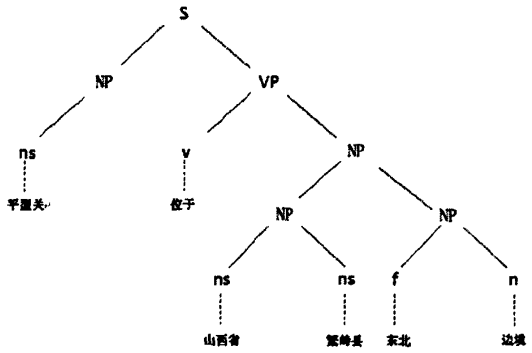


图 2.8 空间方位关系描述的句法分析树

产生式规则如下：

1) S →NP VP　　7) ns→平型关

2) NP→ns　　8) v →位于

3) VP→v NP　　9) ns→山西省

4) NP→NP NP　　10) ns→繁峙县

5) NP→ns ns　　11) f →东北

6) NP→f n　　12) n →边境

基于产生式规则的不同文法形式化体系可以根据其生成能力进行比较，形成文法的层次结构，称为乔姆斯基层次体系（Chomsky Hierarchy）。按生成能力的降序，四种类型的文法如下：

1）0 型文法：可以接受任何形式的产生式规则；

2）1 型文法：上下文相关文法（Context Grammar，CG）；

3）2 型文法：上下文无关文法（Context-free Grammar，CFG）；

4）3 型文法：正则文法（Regular Grammar）；

0 型文法约束最弱。1 型文法对语言的表达能力强于 2 型文法，但计算处理复杂度高。3 型文法约束最强，在产生式规则的右端只能出现一个非终结符，仅能表达语言中各部分的排列顺序，而不能表达层次结构。2 型文法，即上下文无关文法，能处理嵌套、递归结构，足以表达大部分的自然语言结构；同时，其又足够严格，可以在其基础上搭建出分析句子的高效分析器。因此，当前大多数研究多采用 2 型文法来设计自然语言空间查询的文法规则[55][58, 124]。但由于中文自

然语言在句法结构上的灵活性，目前采用概率上下文无关文法（PCFG）的中文句法分析效果尚不能达到类似分词一样的实用水平[126]。

（2）框架语义学：论旨角色

框架语义学（Frame Semantics）是在早期格语法理论的基础上提出的。该理论认为，要理解词语的意义，就必须首先具备概念结构，即语义框架的知识。语义框架是有关概念结构、信仰、习俗、意象等模式的图示化表征，此图示表征能够为某一语言团体提供意义交流的基础。词义是用框架来描写的，词语可以通过其所在的语言结构，按照一定的原则或方式选择和突出基本的语义框架的某些方面[127]。在句法分析的基础上，可以通过 FrameNet[127]和 Penn PropBank[128] 这类框架语义型词典获取谓词相应的论旨角色，从而构建空间方位关系。常见的论旨角色有施事、受事、与事、工具、结果、处所等。

如在 FrameNet 中，对于词汇单元“meet”，可参与多个框架，作为动词可参与“Meet_with_response”、“Locative_relation”、“Congregating”等的框架，作为名词“meeting”可参与“Social_event”、“Discussion”框架。而“Locative_relation”框架则是“Relation”、“Trajector-Landmark”的子类框架，其核心论旨角色为 Figure（Figures）和 Ground，非核心论旨角色为 Direction 和 Distance、Time。针对句法树中的“meet”周围叶节点的词汇，使用适当的框架，即获取各词汇在以“meet”为核心的语义框架中的论旨角色，从而构建完整的语义框架。

2.3 面向信息抽取的空间方位关系表达

从信息抽取的角度，空间方位关系的表达应尽量清晰简洁，如 ACE 定义的 Part-Whole 大类的 Geographical 小类中明确定义了 Arg1 及 Arg2 的命名实体类型为 FAC/LOC/GPE，且抽取的关系在能指和所指上都很明确，如表 2.1 为其中 GPE-GPE 的示例。清晰简洁的表达模型有助于语料标注及抽取方法的实现。

表 2.1 ACE 中 Geographical 关系的 GPE-GPE 示例

Moscow, Russia			
Class	Type	Argument1	Argument2
Formulaic Asserted Unspecified	Part-Whole.Geo	*Moscow, Russia*	*Russia*

而从形式化表达的角度，空间方位关系的表达应尽可能的反映出文本描述的空间关系语义。GUM（Generalized Upper Model）是一个独立于具体领域的语言学本体，由德国 Bremen 大学本体研究组负责研制及维护，目的在于实现自然语言语义的形式化表达[129]。GUM 提供了较完善的信息分类方式，充分考虑自然语言的灵活性，使得个应用领域可以有连贯一致的表达模型。其中 GUM-Space 部

分针对英文从四个方面对自然语言中的空间关系进行表达：1）GeoML（Static Spatial Descriptions），侧重于对空间实体静态位置的表达；2）Maptask（Spatial Instructions），侧重于对空间实体形态及其关系的表达；3）IBL 与 4）Trains（Route Instructions），从运动局部及整体的角度对路径的描述进行的表达。GeoML 的示例如下：

例 2-4

A Building 5 miles east of Fengshan.

SpatialLocating SL1(locatum "building",

placement GL (hasSpatialModality

EastExternal + quantitativeDistantExtent "5miles",

Relatum "Fengshan"))

其中 locatum/relatum 即为射体/界标。GL 是 GeneralizedLocation 的简称，表示通用的位置，SpatialModality 子类对英文中的空间关系进行的详细的定义（图 2.9）。如 EastExternal 表示外部的东方，quatitvativeDistantExtent 表示定量的距离范围等。

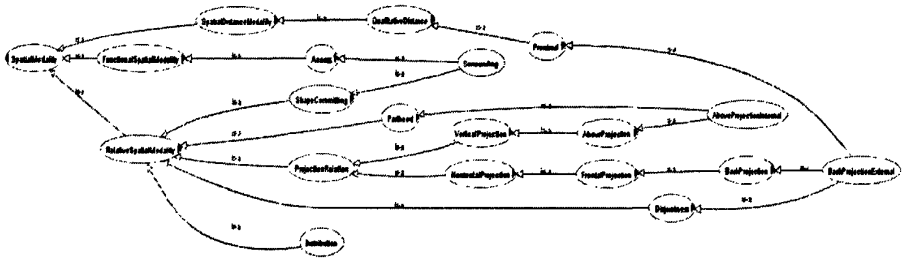


图 2.9 GUM-Space 中 SpatialModality 的子类

可见，空间方位关系的表达一方面需有效的结构化文本描述的空间关系信息，为抽取后 GIS 空间关系的解析提供基础；另一方面，表达必须考虑到信息抽取方法的适用性。

2.3.1 空间方位关系表达式

Mark 认为在空间关系中，拓扑关系是本质的，起着定义的作用，而度量信息（大小、形状、距离和方向等）起细化的作用。拓扑和度量在描述空间关系的语言中所扮演的相关角色是复杂的[21]。闫浩文认为在日常生活中，方向关系是一类比拓扑关系使用频率更高的空间关系[23]。事实上，人类对空间认知经验进行范畴化的过程中，并没有刻意的对应区分拓扑关系和方向关系。在不同的语言中，几乎都有专门描述方向关系的词汇，如大部分语言中都有与汉语“东南西北”相应的词汇。但在描述拓扑关系的词汇上，则较为复杂。如英语中使用两个不同的词汇“on”和“above”分别表示两种不同拓扑的（有接触和无接触）“在……上”的方向关系；而汉语使用“上面”和“上方”这类附加名词的方式来描述这

种拓扑关系的区别，而在一些情况下，“上面”也可为无接触。比较“日本位于中国东部”和“江苏省位于中国东部”，这里单纯从词汇“东部”无法区分拓扑关系，拓扑关系需通过实体类型进行区分。可见，在日常交流语境下，拓扑关系作为本质的关系，常不通过范畴化的某个具体空间词汇进行描述，而是隐含于方向关系的描述，或者蕴含于交流的语境。

根据空间方位参照体系及框架，兼顾信息抽取与 GIS 空间关系的表达，将空间方位关系中的拓扑关系和方向关系分离，且将距离关系作为方向关系的一个属性。定义空间方位关系表达式如下：

(Topology(topologyType), Direction(directionType,frame,Distance))

(Trajector, Landmark) （2-3）

其中，frame 为空间参照框架，Trajector/Landmark 为射体/界标。由于文本符号与地图符号的不同，在 GIS 空间关系中容易定性的拓扑关系类型，从信息抽取的角度，由于常没有明确的词汇指示而较难识别，成为空间方位关系抽取的难点。相反，在 GIS 空间关系中受目标自身位置和形态影响而较难定性的方向关系，由于有明确的方位词指示，所以较易抽取。

2.3.2 空间方位关系的分类

针对空间方位关系表达式中的拓扑关系和方向关系的分类，一种方法是使用 GUM 这类语言学本体的分类结构，但针对汉语目前尚没有类似的知识库，且不便于信息抽取方法的实现。另一方面，考虑到与 GIS 空间关系的接口，我们参考图 2.7 中空间关系定性表达（D）中的分类方式。而面临的主要问题在于，GIS 空间关系定性表达是根据图形语义分类的，而不是根据文本语义分类。这导致了符合图形语义分类的术语未必符合文本语义分类的语境，这也就是“同名不同胚，同胚不同名”的情况。这里的“同胚”即为具有同一 GIS 空间关系定性表达的术语，严格说是具有相同的 9 交模型或一致的语义量化公式，而“同名”则是指使用相同的自然语言空间关系词汇。如图 2.10，I 和 II 的 9 交矩阵是不一样的，但都可以使用“覆盖（Overlap）”这一空间词汇，即“同名不同胚”；II 和 III 的 9 交矩阵相同，但 III 的情况，我们习惯上不称为“覆盖（Overlap）”，而称为“相交（Cross）”，这是“同胚不同名”的情况。后一种情况可以通过语义量化参数进行细化，从而符合自然语言的区分。

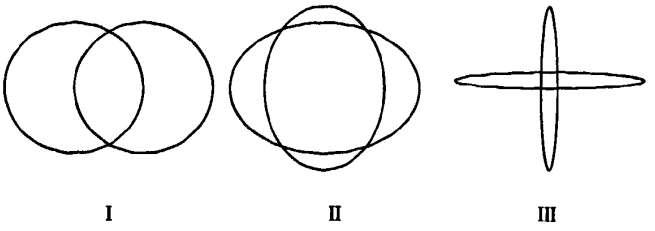


图 2.10 同名不同胚，同胚不同名

即使在区分“覆盖（Overlap）”和“相交（Cross）”定性表达术语的 GIS 空间关系分类中，这种分类模型也无法适用于自然语言。在 GeoDataBase 中，将空间拓扑关系分为 7 类，其中区分了相交（Cross）与覆盖（Overlap）（如图 2.11）。但其定义参照面和比较面不存在“相交”关系，这种关系属于“覆盖”类型。而在自然语言中，我们并不做这种区分。

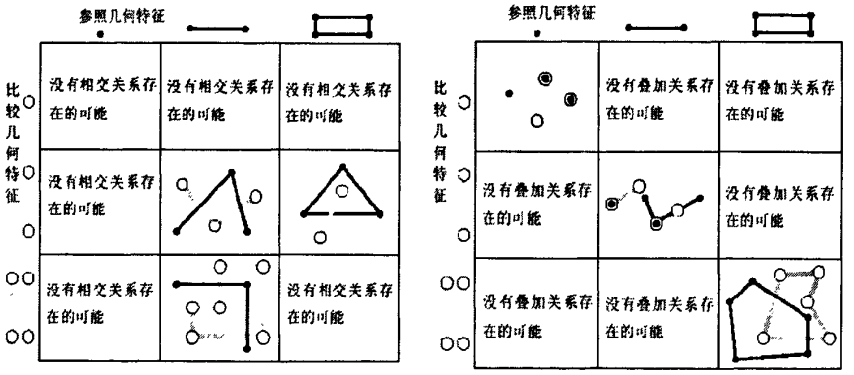


图 2.11 GeoDataBase 拓扑关系中的“相交（Cross）”与“覆盖（Overlap）”

因此，我们参考 GIS 空间关系定性表达中最为基础的分类模型，即 RCC8 来对空间方位关系中的拓扑关系进行分类(表 2.2)。严格意义上，是引入了 RCC8 的拓扑关系类型术语，而通过对这些术语在文本语义中的解释（参见 2.3.3 节）而使其适合于文本语义分类的语境。在图 2.7 的表达层次中，这可以看作是在不改变定性表达（D）类型术语的图形语义情况下，将其借用于形式语言（C）类型术语的方法。因此，本文中我们使用 RCC8 的英文简写作为类型术语，而不使用同义的中文术语。

表 2.2 空间拓扑关系分类

拓扑关系类型	示例
TPP(tangential proper part)	[青铜峡市]位于[银川平原]南端。
NTPP(non-tangential proper Part)	[大庆市]地处[松辽拗陷带]中心部位。

EC(external connection)	[苗栗北境]与[新竹]交界处。
DC(discrete connection)	[白云鄂博]位于[包头市]正北约 150 公里处。
PO(partial overlap)	[密云县]有[京通铁路]过境。
EQ(equality)	[墨脱]又称[白马岗]。

方向关系的类型术语则与文本能指基本一致。分类如表 2.3：

表 2.3 空间方向关系分类

方向关系类型	示例
Behind	[青龙镇]镇后有南宋[酒瓶山]遗址。
Front	[乌拉山]与[大青山]前[昆都伦河冲积扇]上。
Above	[锡山]上有[龙光寺]。
Below	坐落在[天寿山中锋]下的[长陵]地面建筑最宏伟。
Right	[北港镇]在[北港溪下游]右岸。
Left	[裕溪口港]位于[长江]左岸。
N/NE/ENE... (sixteen-point compass rose)	[北戴河]西起[戴河口]，东到[鹰角亭]。

2.3.3 空间方位关系的判断

分类学的目标在于对外延实例的有效区分。阅读者进行空间方位关系的理解判断是从文本描述，经意象图式，到类型概念的过程。由于方位词为方向关系提供了明确的范畴界定，因此标注者对方向关系类型的判断较为清晰。而标注者从文本描述主观产生的拓扑关系意象图式，与 RCC8 定义的类型中显性精确的图形表达未必一致。本文根据原型理论，以 RCC8 定义的类型图形作为原型，借鉴定性空间表达中的“概念邻近图”的思想，对由文本描述触发的拓扑关系意象图式与 RCC8 类型术语作如下映射：

（1）意象图式中 EQ 强调实体 A 和 B 在空间域上的形态相似且面积相近。即使图形表达情况为 PO 类型（如图 2.12），也作为 EQ 类型理解。除了两个实体的比较外，文本描述中的 EQ 概念还包括所指为在不同时间阶段中的同一实体。由于现实世界边界的模糊性，如同下文 EC 的相切问题一样，TPP 与 NTPP 在实际标注时难于区分包含者与被包含者是否接触，因此将其统一为 IN 类型。而意象图式中 IN 强调了部分-整体关系（Mereology）、隶属关系，而不关注是否内切。

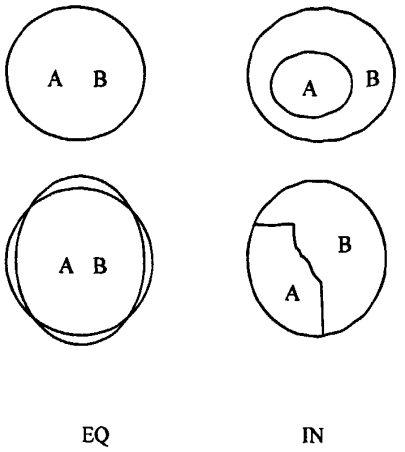


图 2.12 空间拓扑关系的 EQ/IN 类型意象图式

例 2-5 是触发 EQ 与 IN 类型意象图式的例句：

例 2-5

A）EQ [云林]，旧名[斗六]。

B）IN [北海]为镇建制，隶属[广东省合浦县]。

（2）将 RCC8 中 DC-EC-PO 的概念领域看作是界标/射体间由远到近直至重叠的运动过程，作为原型的精确 DC、EC、PO 是这一运动过程中的三个邻近状态，而 EC 作为 DC 与 PO 的临界状态而得以区分三者。由于现实世界边界的模糊性，意象图式的 DC-EC-PO 间的区分则较为复杂（如图 2.13）。

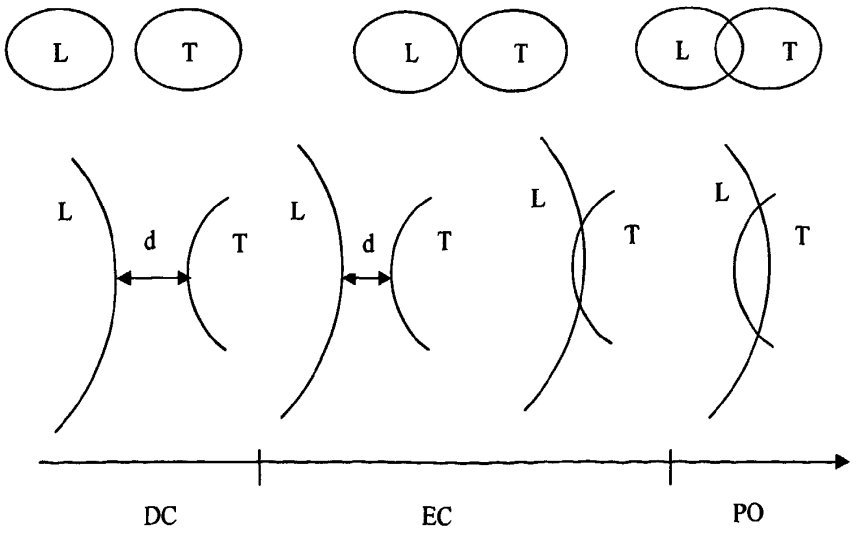


图 2.13 空间拓扑关系的 DC-EC-PO 类型意象图式

设界标 L 与射体 T 间的距离为 d，阈值 D：

则当 d>>D 时，为 DC 类型；

当 d<<D 时，则为 EC 类型。

其中>>表示“远大于”，<<表示“远小于”。

设界标 L 与射体 T 重叠后，相交面积为 s，阈值 S：

则当 s<<S 时，为 EC 类型；

当 s>>S 时，为 PO 类型。

其中 D 与 S 的设定，完全取决于阅读者的主观设定。意象图式中的 DC-EC-PO 的边界较为模糊。从运动的角度，意象图式中的界标/射体更倾向于朝原型 DC-EC-PO 中的哪个状态运动，则为此类型。因此，意象图式中的 DC 强调界标与射体间存在距离，EC 强调界标与射体间邻近、接触，PO 强调界标与射体存在重叠部分。以下是触发 DC-EC-PO 意象图式的例句：

例 2-6

A）DC [白沙镇]镇西 6 公里处的[紫金滩]。

B）EC [镇江]和[扬州]仅一江之隔。

例 2-7

A）EC [班公错]是[中国西藏自治区]西部边境著名界湖。

B）PO [北盘江]发源于[云南省宣威马雄山]西北麓，流经滇东。

比较例 2-6 中的两例，A）为明显的 DC，而 B）如果在图形表达情况下应为 DC，由于文本描述强调的是接近，因此归于 EC 类型。比较例 2-7 中的两例，B）为明显的 PO，A）如在图形表达情况下应为 PO，而由于文本描述强调的是边界、接触，因此归于 EC 类型。

由于阈值 S 和 D 的主观性、以及文本描述触发的意象图式的隐含性，因此上述意象图式的示例为拓扑关系类型的判断提供了参考，起到一定的指导作用。

2.4 本章小结

本章首先在分析认知语言学的空间方位参照框架及中文文本空间方位关系描述特征的基础上，通过文本与地图两类符号系统在空间关系表达上的对比，提出了中文文本空间方位关系表达的两个层次，以及受语境约束的术语在空间方位关系表达研究中的纽带作用。同时对本文及所在课题的研究进行了定位。然后，从空间方位关系表达式、分类、类型术语的判断三个角度分析了面向信息抽取的空间方位关系表达。

第3章 空间方位关系的语料标注与分析

3.1 空间方位关系的语料标注

无论是基于模式规则，还是基于统计学习，在自然语言处理领域，语料库的构建是研究的基础。空间方位关系的语料标注包括以下两个部分：（1）在空间方位关系表达式的基础上，定义能进一步反映语言特性的标注集和作为人工标注操作依据的标注规范；（2）确定适于空间方位关系标注的存储格式和系统平台，选择领域语料进行标注，并根据领域文本特征进行标注集与标注规范的扩展。

3.1.1 空间方位关系标注集

空间方位关系的表达式定义了语料标注最终的语义结构，这个结构是基础的，具有普适性的。而为语料标注设计的标注集在此基础上，还需要考虑两方面的因素。首先是对文本序列和语义元素间的关联层次。在语料标注这一结构化的过程中，适度的标注层叠要求即能反映文本到语义间映射过程的语言学特性，对于标注人员又有一定的可操作性。其次，是标注集语义的具体分类粒度。在空间方位关系范畴和表达定义的分类间，需要通过类型本身的内涵分析与外延示例来进行权衡。

在语言空间信息标注领域，除了语言学本体支撑的 GUM 外，还有以下几种常用的标注集：（1）TRML（Toponym Resolution Markup Language）是 TAME（Toponym Annotation Markup Editor）系统采用的一种地名标注集。其在标识文档、段落、句子及词语的基础上，可对地名的候选空间位置进行标注。（2）GeoTagger 是 MetaCatra 公司 MetaCarta Geographic Text Search(GTS)智能地名搜索中采用的标注集，可根据文档上下文判断某一地名的置信度，从而获取更为智能的空间位置信息标注。（3）TESLA（The gEoSpatial Language Annotator）是一种适用于实时路径描述的语音语料标注集[130]。（4）SpatialML（Annotation Schema for Marking Spatial Expressions in Natural Language）由美国著名的非盈利研究组织 MITRE 制定，是目前国际上第一个较为系统的、适用于多语种的自然语言地理信息标注体系[131]。ACE 已将其采纳为其文本空间信息标注的标准之一，并在扩展的英文文本标注指南基础上，建立了英文语料库。

上述的标注集都是以地名标注为核心的，面向的主要还是地名解析及地名词典信息发布。其中 SpatialML 在清晰定义地理命名实体标注集的基础上，开创性的使用 RLINK/LINK 来标注文本中的方向关系与拓扑关系，同时使用 SIGNAL 来关联语言特征[132]。我们在 SpatialML 的基础上，针对中文描述的特征扩展

SIGNAL，形成空间方位关系的标注集（如图 3.1）。

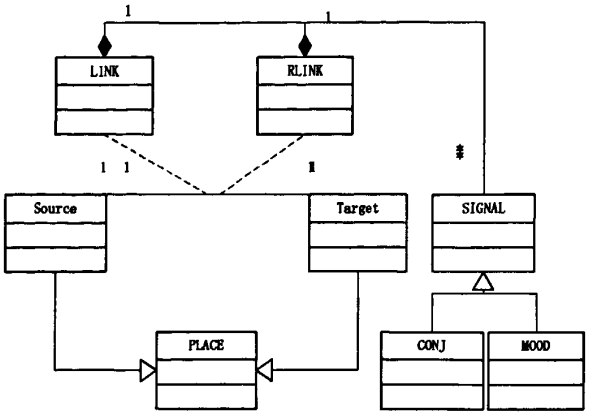


图 3.1 空间方位关系标注集

其中，PLACE 为地理命名实体，RLINK/LINK 分别作为关联类标注 Source（界标）/Target（射体）的空间方位关系，并采用上一章的拓扑与方向关系分类体系作为 LINK 与 RLINK 的分类标准。而 SIGNAL 则对指示空间方位关系的空间词汇进行标注。此外，CONJ 用于标识空间方位关系表达中常用的介词和连词等。MOOD 用于标识描述强弱程度及模糊概念的词汇。标注示例如例 3-1：

例 3-1

[白云鄂博]位于[包头市]正北约 150 公里处。

<PLACE id=1 typecode=510000 form=NAM>白云鄂博</PLACE>

<PLACE id=2 typecode=510000 form=NAM>包头市</PLACE>

<SIGNAL id=3 type="DIRECTION">正北</SIGNAL>

<SIGNAL id=4 type="DISTANCE">150 公里</SIGNAL>

<CONJ id=5 >位于</CONJ>

<MOOD id=6>约</MOOD>

<RLINK id=7 source=2 target=1 signals=3,4 conj=5 mood=6 frame=EXTRINSIC direction=N />

<LINK id=7 source=2 target=1 signals=3,4 conj=5 mood=6 linktype=DC />

3.1.2 空间方位关系标注规范

空间方位关系的标注以句子为单位，即一个空间方位关系实例涉及的所有地理命名实体及起指示作用的空间词汇必须位于同一句子中。除地理命名实体及其已标注的类型外，在标注过程中不应再参考其他先验知识。在实际标注工程中的操作规范主要包含以下几点：

（1）RLINK/LINK 标注与 PLACE 标注的划分粒度

PLACE 标注中包含属性 mod，用以标注地理命名实体自身的方位限制，如“华南”的 mod 为 S。在英语中，由于地名单词的首字大写，较易判断出现的方位词是置于 PLACE 的 mod，还是作为 RLINK/LINK 的 SIGNAL。而在中文中，

需从 RLINK/LINK 标注的角度出发标注，同时也尽可能的从句子层面标注出指示 RLINK/LINK 的 SIGNAL，以保证地理命名实体的相对稳定性，如例 3-2。

例 3-2

[巴里坤湖]位于[天山东段]北坡。

<PLACE id=1 typecode=130000 form=NAM>巴里坤湖</PLACE>

<PLACE id=2 typecode=610000 form=NAM mod=E>天山东段</PLACE>

<SIGNAL id=3 type="DIRECTION">北坡</SIGNAL>

<RLINK id=4 source=2 target=1 signals=3 direction=N />

<LINK id=5 source=2 target=1 signals=3 linktype=IN />

而对于地理命名实体（尤其是地址）中，各地址要素出现的逐层包含的情况，需将相邻包含的地址要素间设置 LINK 为 IN。在实际标注操作时，在将各地址要素标注为 PLACE 的基础上，再将地址作为整体标注为一个 PLACE，而各地址要素间的逐层 IN 关系不再显式标注。

（2）Source/Target 的不对称性

从空间方位参照体系中射体/界标的角度，即使是 LINK 中的非 IN 关系，空间方位关系中 Source/Target 都是不对称的。对于例 3-3，一般情况下，Source/Target 对应关系在 A）中为“密云县”/“京通铁路”，B）中为“滕格诺尔”/“艾不盖河”，C）中为“包头市”/“白云鄂博”，D）中为“白马岗”/“墨脱”。但更为复杂的情况是，从篇章的角度看 B），如果文章通篇是在介绍“滕格诺尔”，则 Target 应为“滕格诺尔”。对于 RLINK 而言，这种由于叙述产生的复杂性则被方向具有的明确 Source/Target 指示性所消除。

例 3-3

A）PO [密云县]有[京通铁路]过境。

B）EC [艾不盖河]向北汇入[滕格诺尔]。

C）DC [白云鄂博]位于[包头市]正北约 150 公里处。

D）EQ [墨脱]又称[白马岗]。

（3）LINK/RLINK 关系的显著度

在文本描述过程中，空间方位关系未必作为浅层语义表现出来，如例 3-4。

例 3-4

A）IN/SW [绵阳市]位于[四川盆地]西北部。

B）IN [绵阳市]是[四川省]辖市。

C）IN/SW [绵阳市]是[四川盆地]西北部交通枢纽。

D）IN 铁路年客运量在[四川省]内仅次于[成]、[渝]两地。

E）IN/W [阿克苏市]这颗镶嵌在[祖国]西陲的明珠熠熠生辉。

由 A）到 E）的文本描述中，空间方位关系的显著度逐步减弱。A）通过空间词汇“位于”、“西北部”明确指示了空间方位关系。B）、C）通过行政区划和道路交通相关的专题空间词汇“辖市”、“交通枢纽”间接指示了空间方位关系。而 D）不仅仅限定在交通专题领域，更要通过一定的推理来获取空间方位关系。

E）则包含了比喻。为了和 GIS 中空间关系的外延一致，本文将 A）、B）、C）设置显著度为 1（在具体标注操作时不设置），对于 D）类需要在不包括先验知识情况下进行推理的空间方位关系，设置显著度为 2，而对于 E）则不进行标注。

（4）LINK 关系的类型判断

如同拓扑关系的类型判断一样，在空间方位关系描述触发的隐性意象图式中，较难界定实体的点、线、面特征。考虑到现实世界中面状地物的本原性，在上一章空间拓扑关系类型判断指导下，要求标注者尽可能的以面来抽象地理命名实体。地图中的线状地物认为条状面（即类线地物，Line-Like Object），点状地物认为小面积的圆（即类点地物，Point-Like Object）。而对于更具领域性、复杂性的空间方位关系描述的理解，则尽可能与日常的地图进行对应，而获取较为一致的类型判断。

3.1.3 空间方位关系的语料库构建

3.1.3.1 语料标注格式

语料标注是在标注规范的约束下，人工进行文本识别与语义解析的过程。作为逻辑模型的标注集，其结构的复杂度决定了标注所采用的物理模型，即语料标注格式。同时，语料标注格式根本上还要受到文本序列性的制约。针对空间方位关系的语料标注，现有的标注格式大致可分为以下三种：

（1）纯文本方式

在词性标注、命名实体识别为主的语料中，纯文本方式是使用最为广泛的标注格式。例如“白云鄂博/ns 位于/v 包头市/ns 正北/f 约/d 150/m 公里/q 处”。而在进行空间方位关系标注时，由于文本跨度（Span）的叠置，需要嵌入首尾标识。例如[93]：

……有/vg 一/m 天/q ，/wo 鹰/ng[TR3] 看见/vg 农夫/nc[TR1][TR2] 坐/vg 在/d #1 将要/vz 倒塌/vg 的/usde #1 墙/ng[LM1][LM2] 下/f #r[SE1]#r[SE2]，/wo 就/d 立刻/d #1 朝/p 下/f #r[SE3]飞/vg 去/vg

其中 LM 表示界标，TR 表示射体。SE 表示空间表达式，其作为空间关系的一个组成部分。#1 和#r 为空间表达式的首尾标识。可见，在词性标注的基础上，为了对同一句子中的不同空间方位关系涉及的界标/射体对及其空间表达式进行区分，需加入较多的附加标识。

纯文本的标注格式在标注文本不存在叠置且标注集内容较少的情况下，具有形式简洁、解析简单的优点。因此，其便于标注内容的共享与交换。但在进行空间方位关系的标注，尤其在长距离省略和焦点转移这类存在文本叠置及标注对象较多的情况下，纯文本的标注格式既不便于人工标注及阅读，也不便于标注集的解析。

（2）Stand-On XML 方式

使用 XML 替代纯文本有助于更好的格式化标注集内容。TRML 即采用 Stand-On XML 的格式。例如对句子“EU rejects German call to boycott BRUSSELS lamb.” 的标注格式为：

```
<s id="s1">
  <w tok="EU" pos="NNP" chk="I-NP" ne="I-ORG" />
  <w tok="rejects" pos="VBZ" chk="I-VP" ne="O" />
  <w tok="German" pos="JJ" chk="I-NP" ne="I-MISC" />
  <w tok="call" pos="NN" chk="I-NP" ne="O" />
  <w tok="to" pos="TO" chk="I-VP" ne="O" />
  <w tok="boycott" pos="VB" chk="I-VP" ne="O" />
  <w tok="BRUSSELS" pos="NNP" chk="I-NP" ne="I-LOC" />
  <w tok="lamb" pos="NN" chk="I-NP" ne="O" />
  <w tok="." pos="." chk="O" ne="O" />
  <candidates>
    <cand id="c1" src="NGA" lat="-23.3833333" long="29.15"
    humanPath="Brussels &gt; (SF04) &gt; South Africa" />
    <cand id="c6" src="NGA" lat="50.8333333" long="4.3333333"
    selected="yes" humanPath="Brussels &gt; (BE02) &gt; Belgium" />
    ...
  </candidates>
</s>
```

使用 Stand-On 格式的 XML 将文本序列拆分后与标注集统一格式化，有助于标注集的解析与共享，但在处理标注文本叠置的问题上，与纯文本方式一致。XML 标注方式的主要问题在于文件容量过大及语料共享时的格式转换，但目前主流数据库对 XML 的广泛支持则弥补了这些不足。

（3）Stand-Off XML 方式

使用 Stand-Off XML 作为在标注格式是目前通用语料标注中逐渐被采纳的方式。其实现了文本序列与标注集的解耦，即标注集内容不再镶嵌在文本序列中。以下文 GATE 使用的标注格式为例：

```
...
<TextWithNodes>
...<Node id="73" />白云鄂博<Node id="77" />位于<Node id="79" />包头市<Node
id="82" />正北<Node id="84" />约<Node id="85" />150<Node id="89" />公里
<Node id="92" />处<Node id="93" />...
</TextWithNodes>
<AnnotationSet Name="SpatialML">
  <Annotation Id="750" Type="PLACE" StartNode="73" EndNode="77">
    <Feature>
      <Name className="java.lang.String">id</Name>
      <Value className="java.lang.String">3</Value>
```

```
    </Feature>
    ...
  </Annotation>
  <Annotation Id="754" Type="LINK" StartNode="73" EndNode="93">
    <Feature>
      <Name className="java.lang.String">target</Name>
      <Value className="java.lang.String">3</Value>
    </Feature>
    <Feature>
      <Name className="java.lang.String">linkType</Name>
      <Value className="java.lang.String">DC</Value>
    </Feature>
    ...
  </Annotation>
  ...
</AnnotaitonSet>
```

通过使用 Stand-Off XML 的方式将文本序列表示为 XML 中的 Content 而不是 Element，从而保持文本序列的非结构化特征，然后根据实际的文本段在序列中标识其首尾的节点（StartNode/EndNode），得以结构化的方式表示标注集内容。这些节点对于文本序列是唯一的，同时可被多个标注引用。该方式充分利用了 XML 对半结构化信息的表达能力，能很好的处理文本叠置及复杂标注集的情况。此外，在便于标注内容解析的基础上，提供了对文本叠置的拓扑运算，有助于全文索引及内容检索。

3.1.3.2 语料标注平台

GATE（General Architecture for Text Engineering）是目前被广泛采用的一个开源自然语言处理平台[133]。通过基于插件的软件框架和面向对象的编程模型，GATE 为语言工程提供了语料标注、信息抽取、信息检索、算法评测、资源集成等各个层面的功能组件，这些组件称为 CREOLE(a Collection of Reusable Objects for Language Engineer)，其中语料资源（Language Resource）提供了对不同格式来源和编码的文本语料的统一管理，并可通过 XSD（对应于 Schema Annotation Editor）或 Ontology（对应于 Ontology-based Corpus Annotation Tool，OAT）进行标注集架构的预定义。处理组件（Processing Resource）则以管道（Pipeline）的方式进行组合，形成具体应用（Applications）。本文的抽取方法将通过构建此类组件进行集成。GATE 语料标注操作界面如图 3.2。

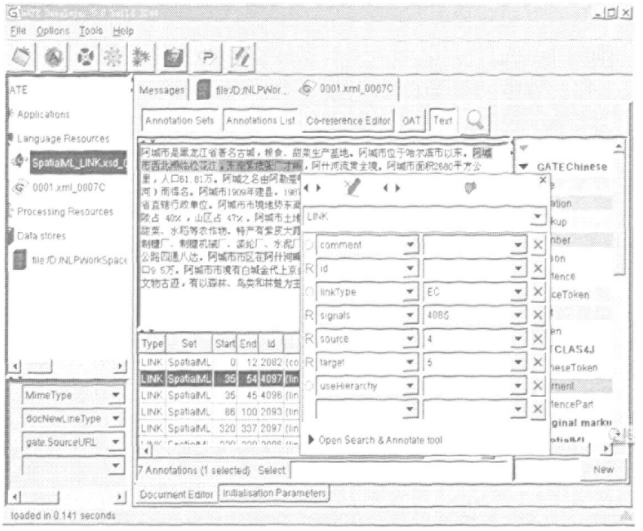


图 3.2 GATE 语料标注操作界面

本文采用上文所述的标注集和标注规范进行空间方位关系的语料标注，扁平化后的 SpatialML XSD 如图 3.3。在标注过程中，根据语料的领域特征进行标注规范的扩展，标注结果以 Stand-Off XML 格式进行存储。

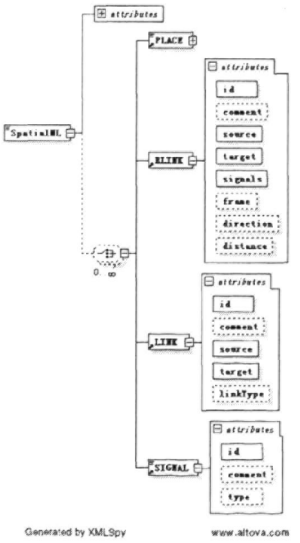


图 3.3 SpatialML XSD

在 GATE 语料标注工程中，标注语料由多文档（Document）组成。每个文档可包括多个不同格式的标注集合（Annotation Set）。这些标注集合具有统一的文本跨度参数和不同的标注特性（Feature），后者通过键值对进行表达。通过对文本序列的拓扑运算可获取不同文本跨度中的标注集合，也可直接通过对标注特性的值查询标注集合。如图 3.4 为 LINK 关系标注集合中一个 IN 类型的实例。

图 3.4 GATE 语料标注实例

3.2 实验语料的构建与分析

实验语料采用《中国大百科全书（地理分册）》中随机选取的 200 篇富含空间方位关系的文本。该语料为典型的地名辞典类文本，通常针对一地名，通过地理位置、社会人文、气候环境、经济政治等方面进行详细的描述。文体严谨简练，具有相当的领域代表性。标注工程采用 GATE 平台，基于上述标注集及标注规范，招募 GIS 专业学生进行人工标注及标注结果的交叉检验。

3.2.1 实体组合的标注扩展

在语料标注过程中，现有的标注集能一致的处理省略及其焦点转移的情况，但不能较好的处理地理命名实体组合作为射体/界标的情况。对于例 3-5 的空间方位关系标注，直观的方法是使用三元组标注。但由于中文词汇“之间”与“Between”不尽相同，其指示的地理命名实体可以为多个。此外，依据空间方位参照框架理论，多元组本质上是一个根为二叉结构的树。因此，SpatialML 的规范提供了一个不扩展标注集的办法，即使用的二叉树层叠处理这类标注。

例 3-5

[大娄山]是[赤水河]与[乌江水系]的分水岭。

<PLACE id=1 typecode=610000 form=NAM >大娄山</PLACE>

<PLACE id=2 typecode=110000 form=NAM >赤水河</PLACE>

<PLACE id=3 typecode=110000 form=NAM >乌江水系</PLACE>

<SIGNAL id=4 type="TOPOLOGY">分水岭</SIGNAL>

<LINK id=5 source=2 target=3 signals=4 linktype=EC />

<LINK id=6 **source=5** target=1 linktype=EQ />

但从标注集类结构的角度，LINK/RLINK 是表示 source 与 target 空间方位关系的关联类（Association Class）而非实体类。将关系视为实体的过程隐含了将关系涉及的两个 PLACE 进行空间拓扑运算获取新空间域（Domain）的步骤，而新的空间域作为实体参与更高一层的方位关系构建。例 3-5 中，关系为 EC 的“赤水河”与“乌江水系”在进行求交集（Intersect）后形成的新的空间域，进而与“大娄山”形成 EQ 的关系。而使用二叉树的层叠标注则默认 EC 蕴含的为求交集运算。但除交集运算外，还可能存在着其他的空间拓扑运算，如例 3-6。

例 3-6

[云岭山体北段]在[西藏自治区]及[四川省]境内。

<PLACE id=1 typecode=610000 mod=N form=NAM >云岭山体北段</PLACE>

<PLACE id=2 typecode=510000 form=NAM >西藏自治区</PLACE>

<PLACE id=3 typecode=510000 form=NAM >四川省</PLACE>

<SIGNAL id=4 type="TOPOLOGY">境内</SIGNAL>

<LINK id=5 source=2 target=3 linktype=EC />

<LINK id=6 **source=5** target=1 signals=4 linktype=IN />

该例中，“西藏自治区”与“四川省”也为 EC 关系，通过求并集（Union）运算获取新的空间域，而“云岭山体”则与该空间域形成 IN 的关系。为了清晰的反映出空间方位关系描述中的地理命名实体的这种空间拓扑运算的组合过程，需对标注集进行扩展（如图 3.5），即 Target/Source 可由多个 Place 组成。

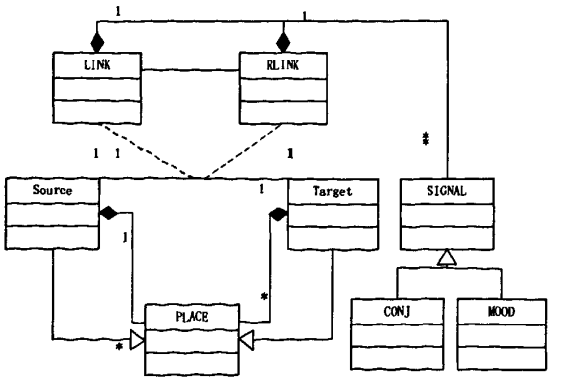


图 3.5 空间方位关系扩展标注集

由于人类的空间认知容量及其语言范畴的限制，虽然从句法结构上允许，但

这种组合的层叠不会过深，不然描述就不自然且难于理解。此外，就地图表达的空间拓扑算子而言，裁剪（Clip）、切割（Cut）以及差集（Difference）和异集（Symmentic Difference）在文本描述组合中的使用较为少见，因此扩展的空间拓扑运算如表 3.1。

表 3.1 地理命名实体组合中空间拓扑算子类型

空间组合拓扑算子	涉及实体的拓扑关系	标注符号
Intersect	EC 或 PO	&
Union	EC 或 PO	|
Convert Hull	DC	_

显而易见，只有存在 EC 与 PO 关系的实体才能进行 Intersect/Union 操作。其中，EC 关系的实体进行 Intersect 操作形成降一维的空间域。DC 关系的实体进行 ConvertHull 操作包含了两点成线和多点成面，因此可一致的处理“之间”这类空间词汇的标注。上述两例新的标注如下：

例 3-7

[大娄山]是[赤水河]与[乌江水系]的分水岭。

<PLACE id=1 typecode=610000 form=NAM >大娄山</PLACE>

<PLACE id=2 typecode=110000 form=NAM >赤水河</PLACE>

<PLACE id=3 typecode=110000 form=NAM >乌江水系</PLACE>

<SIGNAL id=4 type="TOPOLOGY">分水岭</SIGNAL>

<LINK id=5 source=2 target=3 signals=4 linktype=EC />

<LINK id=5 **source=2&3** target=1 linktype=EQ />

例 3-8

[云岭山体北段]在[西藏自治区]及[四川省]境内。

<PLACE id=1 typecode=610000 mod=N form=NAM >云岭山体北段</PLACE>

<PLACE id=2 typecode=510000 form=NAM >西藏自治区</PLACE>

<PLACE id=3 typecode=510000 form=NAM >四川省</PLACE>

<SIGNAL id=4 type="TOPOLOGY">境内</SIGNAL>

<LINK id=5 source=2 target=3 linktype=EC />

<LINK id=6 **source=2|3** target=1 signals=4 linktype=IN />

对于两个实体组合进行 Convert Hull 标注的实例如例 3-9。

例 3-9

[大庆市]介于[哈尔滨]和[齐齐哈尔市]之间。

<PLACE id=1 typecode=510000 form=NAM >大庆市</PLACE>

<PLACE id=2 typecode=510000 form=NAM >哈尔滨</PLACE>

<PLACE id=3 typecode=510000 form=NAM >齐齐哈尔市</PLACE>

<SIGNAL id=4 type="TOPOLOGY">之间</SIGNAL>

<LINK id=4 **source=2_3** target=1 signals=4 linktype=IN />

对多个实体组合进行 Convert Hull 标注的实例如例 3-10。该例中，对于实体

组合的拓扑关系判断需要先验知识，因此不标注。如果实体组合为 EC，则这种情况下 Union 等价于 ConvertHull，也可标为 ConvertHull。

例 3-10

[大相岭山体]蜿蜒于省境西部[雅安]、[荥经]和[汉源]之间。

<PLACE id=1 typecode=610000 form=NAM >大相岭山体</PLACE>

<PLACE id=2 typecode=510000 form=NAM >雅安</PLACE>

<PLACE id=3 typecode=510000 form=NAM >荥经</PLACE>

<PLACE id=4 typecode=510000 form=NAM >汉源</PLACE>

<SIGNAL id=5 type="TOPOLOGY">之间</SIGNAL>

<LINK id=4 **source=2_3_4** target=1 signals=5 linktype=IN />

3.2.2 组合标注的归一规则

通过对标注集及标注规范的扩展，显性的标注出空间方位关系描述中地理命名实体组合的拓扑运算过程。而在空间方位关系的表达中，要求在不加入先验知识的情况下，尽可能的结构化成为 Source/Target 实体对的空间拓扑关系。通过加入归一规则可实现这一目标（见表 3.2）。

表 3.2 组合标注的归一规则表

Domain / LINK	Intersect(A, Bn)	Union(A, Bn)	ConvexHull(A, Bn)
IN(C,Domain)	EC(C,A); EC(C,Bn)	PO(C,A) ; PO(C,Bn); If(n=1) then EC(A,B);	DC(C,A); DC(C,Bn)
IN(Domain, C)	PO(A,C); PO(Bn,C)	IN(A,C); IN(Bn,C)	IN(A,C); IN(Bn,C)
EQ(C, Domain)	EC(C,A); EC(C,Bn)	IN(A,C); IN(Bn,C)	IN(A,C); IN(Bn,C)
PO(C,Domain)	PO(C,A);PO(C,Bn)	PO(C,A); PO(C,Bn);	
DC(C,Domain)	DC(C,A);DC(C,Bn)	DC(C,A);DC(C,Bn)	DC(C,A);DC(C,Bn)
EC(C,Domain)		EC(C,A);EC(C,Bn)	EC(C,A); EC(C,Bn)

其中，通过(A,Bn)表示多元实体组合。由于图式上较复杂，且在具体描述中情况不定，因此 PO(C,ConvexHull(A,Bn))及 EC(C,Intersect(A,Bn))无确定的归一规则。对于由归一规则获取的新拓扑关系，除上述阴影部分外，显著度设置为 2。其中阴影部分的规则，在理解上类似于省略的情况。例 3-7 和例 3-8 归一后的标注结果如例 3-11 和例 3-12。

例 3-11

[大娄山]是[赤水河]与[乌江水系]的分水岭。

<PLACE id=1 typecode=610000>大娄山</PLACE>

<PLACE id=2 typecode=110000>赤水河</PLACE>

<PLACE id=3 typecode=110000>乌江水系</PLACE>

<SIGNAL id=4 type="TOPOLOGY">分水岭</SIGNAL>

<LINK id=5 source=2 target=3 signals=4 linktype=EC />

<LINK id=6 **source=2&3** target=1 linktype=EQ />

<LINK id=7 source=2 target=1 linktype=EC significance=2/>
<LINK id=8 source=3 target=1 linktype=EC significance=2/>

例 3-12
[云岭山体北段]在[西藏自治区]及[四川省]境内。
<PLACE id=1 typecode=610000 mod=N>云岭山体北段</PLACE>
<PLACE id=2 typecode=510000>西藏自治区</PLACE>
<PLACE id=3 typecode=510000>四川省</PLACE>
<SIGNAL id=4 type="TOPOLOGY">境内</SIGNAL>
<LINK id=4 **source=2|3** target=1 signals=4 linktype=IN />
<LINK id=5 source=2 target=1 linktype=PO significance=2 />
<LINK id=6 source=3 target=1 linktype=PO significance=2 />

IN(Intersect(A,Bn), C)的情况也较为常见，如例 3-13。
例 3-13
[岷江]于[宜宾]和[金沙江]相汇。
<PLACE id=1 typecode=110000>岷江</PLACE>
<PLACE id=2 typecode=510000>宜宾</PLACE>
<PLACE id=3 typecode=110000>金沙江</PLACE>
<SIGNAL id=4 type="TOPOLOGY">相汇</SIGNAL>
<SIGNAL id=5 type="TOPOLOGY">于</SIGNAL>
<LINK id=6 source=3 target=1 signals=4 linktype=EC />
<LINK id=7 source=2 **target=1&3** linktype=IN />
<LINK id=8 source=2 target=1 linktype=PO significance=2/>
<LINK id=9 source=2 target=3 linktype=PO significance=2/>

3.2.3 空间关系判断规范的扩展

在上文所述的 LINK 类型判断基础上，对地理辞典类文本描述的空间方位关系中拓扑关系的判断作相应扩展，以提高标注的一致性。图式中，线为类线面（LLO），点为类点面（PLO）。

（1）线状实体的 EC 类型图式（图 3.6）

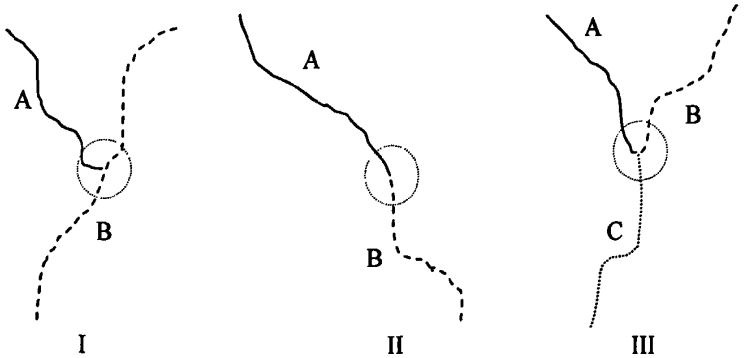


图 3.6 线状实体的 EC 类型图式

相应的例句如例 3-14：

例 3-14

I）EC [艾不盖河]向北汇入[滕格诺尔]。

II）EC [秦淮河]北源[句容河]。

III）EC [岷江]于[宜宾]和[金沙江]相汇后始称[长江]。

（2）线状实体的 PO 类型图式（图 3.7）

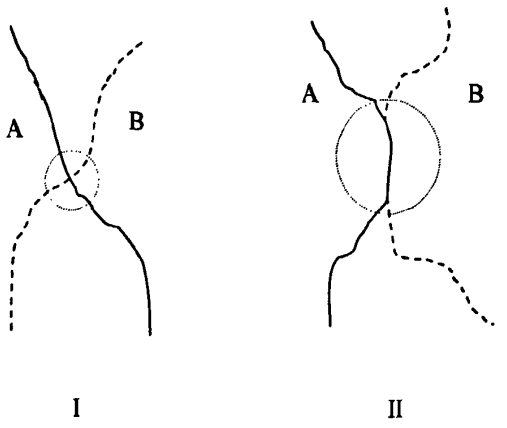


图 3.7 线状实体的 PO 类型图式

相应的例句如例 3-15：

例 3-15

I）PO [原平县]位于[同蒲铁路]和[京原铁路]交会处。

II）PO [潮河]与[白河]分别自东北和西北流入[县境]，在[县城]汇集后分别流出。

（3）线-面实体的类型图式（例 3-8）

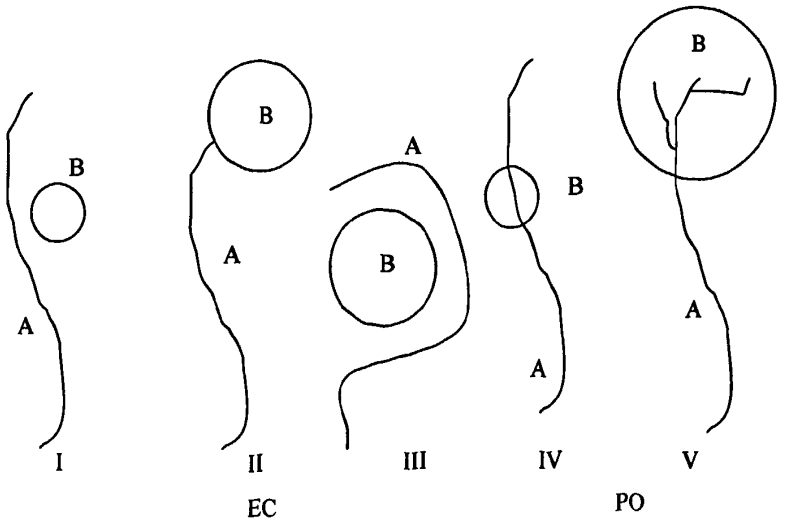


图 3.8 线-面实体拓扑类型图式

相应的例句如例 3-16：

例 3-16

I）EC [岷江小三峡]附近的[板桥驿]盛产荔枝。

II）EC [大夏河]过[临夏市]后至[康家湾]注入[刘家峡水库]。

III）EC [句容河]和[溧水河]会合后，绕[江宁县方山]西侧北流。

IV）PO [北盘江上游]流经[滇东科斯特高原]。

V）PO[大黑河]源于[内蒙古自治区中部蛮汗山东北坡骆驼脖子]和[双鹭鹮]一带。

（4）名词类空间词汇类型判断

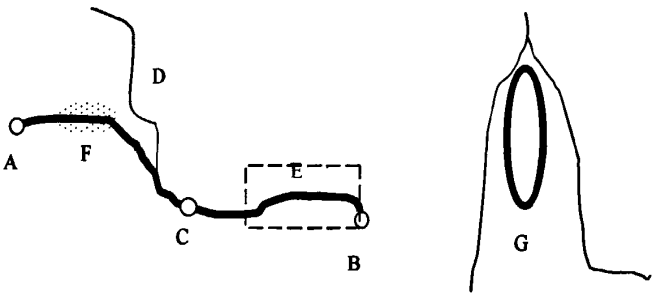


图 3.9 名词类空间词汇类型判断

名词类空间词汇触发的较复杂的意象图式主要与河流、道路等特定专题的地理实体有关（图 3.9）。相应的例句及其拓扑类型如例 3-17：

例 3-17

A）EC [大理市]为[滇藏公路]的起点。

B）EC [凭祥市]为[湘桂铁路]终点。

C）IN [平凉市]为[西（安）兰（州）公路]中途要站。

D）EC [无定河]主要支流有[芦河]、[榆林河]。

E）IN [大通河]下游[八宝川（连城、窑街间）]为重要农耕区。

F）EC [漠河县西林吉镇]位于[嫩林铁路]沿线。

G）EC [云岭]是[澜沧江]和[金沙江]的分水岭。

其中，A）和 B）定义为 EC，与 GeoDataBase 中的切触（Touch）关系一致，而 C）定义为 IN，与 GeoDataBase 中的内含（Within）相一致。

（4）描述尺度的影响

在拓扑类型的判断中，也反映了描述尺度的影响，如例 3-18 与图 3.10 所示。

例 3-18

A）IN [常州]是[沪宁线]上的重要城市。

B）PO [沪宁线]穿越[市区]。

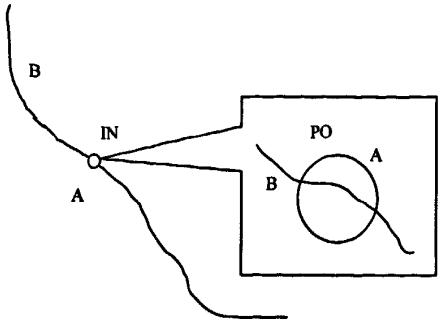


图 3.10 描述尺度对拓扑类型判断的影响

3.2.4 实验语料的分析

3.2.4.1 基本统计分析

针对 200 篇语料，首先进行预处理，包括段落整饰及编码、全半角转换，并使用中科院计算技术研究所的 ICTCLAS 进行分句、分词与词性标注。词性标注采用计算所汉语词性标记集。然后根据标注规范分阶段进行地理命名实体与空间方位关系的标注。语料包括字符 91504 个，词汇 56686 个，短句 14422 条，PLACE6972 例，LINK 关系 2436 例，RLINK 关系 515 例。LINK 关系中，IN 关系 1486 例，占 61%，PO 关系 457 例，占 18.8%，EC 关系 229 例，占 9.4%，EQ 关系 189 例，占 7.8%，DC 关系 75 例，占 3.1%。RLINK 关系中，S 关系 107 例，N 关系 105 例，W 关系 87 例，E 关系 72 例，SE 关系 24 例、NW 关系 24 例，LEFT/RIGHT 等其他关系相对较少。主方向关系与附方向关系比例为 4.2：1，仅有一例 SSW。RLINK 的参照框架为 INTRINSIC 的 7 例，其他都为 EXTRINSIC。

可见，在辞典类地理空间描述文本中，整体-部分关系（尤其是行政、山体、水体隶属关系）是射体定位常用的拓扑类型。而人们在使用方向关系进行定位时，倾向于使用主方向甚于附方向。此外，与文景转换领域经常采用的空间语料不同，地理空间方位关系语料中采用的空间方位参照框架多为与传统二维地图认知相对应的绝对参照框架。

对于 Source-Target 数量映射的统计数据如表 3.3 和表 3.4：

表 3.3 LINK 关系 S-T 数量映射统计

	1：1	1：2	1：3	1：4	1：5 及以上
Source-Target	1641	132	47	20	33
Target-Source	1369	189	66	29	47

表 3.4 RLINK 关系 S-T 数量映射统计

	1：1	1：2	1：3	1：4	1：5 及以上
Source-Target	406	28	4	1	0

Target-Source	310	35	9	9	6

Source-Target 的数量映射包括省略以及组合归一后的情况。其中 LINK 关系的 Source-Target 的数量映射比例最高可达 1：13，Target-Source 最高可达 1：14。RLINK 关系的 Source-Target 数量映射比例最高为 1：4，Target-Source 最高为 1：7。

可见，随着射体与界标映射比例的增加，对应的空间关系的数量急剧下降，但也存在大数量映射的情况。这即反映出空间方位关系描述需与人的认知能力适应，也体现了作为地理辞典类文本对空间域的描述特点。形如“[A 城]南临[B 湖]，北面[C 山]，西接[D 路]。”这类句式，简练而清晰，正是通过省略来统一的描述 A 所处的空间域。

3.2.4.2 空间词汇分析

语料中共有 SIGNAL 689 个，其中指示 DIRECTION 的词汇 96 个，指示 TOPOLOGY 的词汇 498 个，指示 DISTANCE 的词汇 26 个。共有 CONJ 48 个，MOOD 7 个。其中使用频度前五位的 SIGNAL 和 CONJ 如表 3.5。

表 3.5 SIGNAL 及 CONJ 频度前五位词汇

SIGNAL	频度	CONJ	频度
位于	120	是	240
属、称	50	有	110
南	39	为	90
驻	37	在	14
辖市	30	与	12

SIGNAL 根据指示的关系类型统计如表 3.6。

表 3.6 SIGNAL 按指示的方位关系类型统计

RLINK/LINK 类型	SIGNAL 个数	典型空间词汇示例
IN	351	省会、境内、中心部位
PO	128	纵贯、交会、流经
EC	85	河畔、濒临、接壤
EQ	37	古称、语意为、又名
DC	45	隔海相望、距 15 公里
N/S/W/E	92	东、以东、东麓、东郊
SW 等附方向	34	西南部
L/R/Behind/F/A/Below	17	左岸、上方、上头

可见，不同类型的方位关系描述采用的空间词汇与其类型所占的比例正相关，其中 DC 类型相应的空间词汇较多的原因在于其中存在数字。在地理辞典类空间方位关系描述文本中，空间词汇基本可归属于方位词类、谓词类和名词类。

在此基础上，通过组合具有更为丰富的形式：

（1）方位词+动词。例如“北沿”、“东临”。

（2）方位词+名词/名词+方位词。例如“北麓”、“东岸”；“市内”、“河畔”。

（3）动词+其他词性。例如动词+介词（“纵贯于”、“延伸至”），动词+名词（“入江”）。

（4）名词+其他词性/其他词性+名词。例如名词+代词（“中心之一”），形容词/区别词+名词（“主要岛屿”、“重要产区”）。

空间词汇的多词组合源于中文分词粒度的设置。由于中文没有形式上的字词区分，分词时可参照不同的词性标注集，而分词算法性能也影响具体的分词结果。因此，同一空间词汇在不同的词性标注集的标注下，可分出词性不同及数量不等的词汇。以动词、名词为核心，（1）、（3）的情况可归属于谓词类，（2）、（4）的情况可归属于名词类。

除了形如“南京（宁）”这类通过标点符号指示的空间方位关系外，绝大部分空间方位关系都有相应的空间词汇进行指示。而空间词汇在指示空间方位关系时，指示程度不同。尤其是方位词类空间词汇对方向关系及 IN/DC/EC 拓扑关系的指示最为明确。参看例 3-19：

例 3-19

A）PO [北盘江上游]流经[滇东克斯特高原]。

B）IN/W [平山县]位于[省境]西部[太行山区]。

C）IN [平山县]古迹有[商旧址]、[文宣王庙]。

D）IN [四川省南坪县]的[九寨沟自然保护区]，面积 6 万公顷。

空间词汇的指示程度逐步减弱。在 A）中谓词类空间词汇“流经”在射体/界标的地物类型的辅助下，能明确的指示拓扑关系 PO。同时“流经”词汇本身也指示了射体的地物类型为河流。B）中谓词类空间词汇“位于”指示作用依赖于方位词类空间词汇“西部”，如后文所示，也可将“位于”看作指示拓扑关系 IN 或 EQ。C）中的词汇“有”虽为动词，但由于其较弱的指示作用，将其标注为 CONJ。D）中的“的”即为 CONJ，指示作用最弱，在一些情况下可省略。

在具有较强指示能力的空间词汇中，依然存在指示拓扑关系的类型歧义。主要包括以下几种情况：

（1）方位词在拓扑关系上相对较弱的指示能力

例 3-20

A）IN/E [江苏省]位于[中国]东部。

B）DC/E [日本]位于[中国]东部。

（2）地理命名实体类型的不同

例 3-21

A）PO 源于[大青山]的[五贝滩河]。

B）EC/N [秦淮河]源于北面的[句容河]。

（3）三维现实世界映射到地图的二维化结果

例 3-22

A）EC/ABOVE [锡山]上有[龙光寺]、[龙光塔]等。

B）IN [保定市市境]位于[太行山山前冲积扇]上。

前两例都可在确定界标/射体的类型基础上得以消歧。2）的情况是界标/射体大类不同即能消歧，而 1）的情况是界标/射体同一类型的不同层次需确定。

综上可见，空间词汇形式丰富，在指示空间方位关系时起到重要的作用。同时，空间词汇指示强弱程度不一、侧重不一，指示的空间方位关系中的拓扑关系可能存在歧义。

3.3 本章小结

本章在 SpatialML 标注集及标注规范的基础上，采用 GATE 平台进行语料的标注工程。在对地理辞典类语料的空间方位关系标注过程中，针对领域特征进行了处理：1）扩展了地理命名实体组合情况的标注，并设计了归一规则。2）扩展了地理空间描述中拓扑关系判断的规范（更为完整的信息可参考《面向自然语言的地理信息标注规范》（NaturalGML 1.0））。在标注工程的基础上，通过对实验语料的统计，分析了空间方位关系描述中的句法特征以及空间词汇的指示性，为抽取方法的研究提供了基础。

第4章 基于规则的空间方位关系抽取方法

4.1 方法的提出

在关系抽取研究领域，普遍采用的规则方法有基于模式匹配和基于词典驱动的方法。基于模式匹配的方法运用语言学知识，在执行抽取任务前，构造出若干基于词语、词性或语义的模式集合并存储起来。当进行关系抽取时，将经过预处理的语句片段与模式集合中的模式进行匹配。一旦匹配成功，就可以认为该语句片段具有对应模式的关系类型。基于词典驱动的关系抽取方法则是在抽取模式相对稳定的情况下，通过添加新的词汇入口获取对新的关系类型的抽取能力。

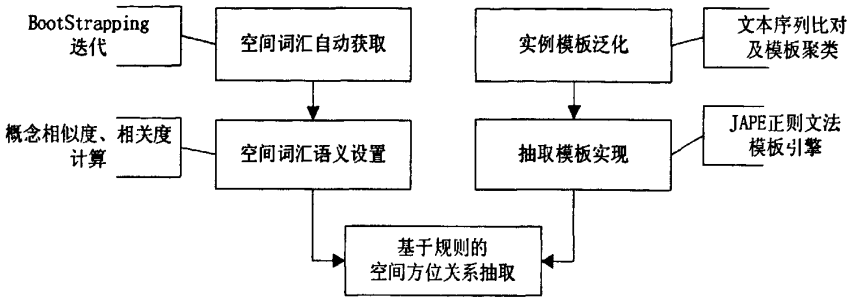


图 4.1 基于规则的空间方位关系抽取流程

通过对标注语料的分析可知，空间词汇对空间方位关系具有显著的指示作用。由此，本文提出通过构建空间词汇词典，结合关系模板及匹配规则的空间方位关系抽取方法。具体流程如图 4.1，主要分为空间词汇词典构建与空间方位关系抽取模板实现两个部分：1）通过 BootStrapping 从生语料中自动迭代获取空间词汇，利用词汇语义型词典中概念相似度、相关度计算进行空间词汇的语义参数设置，使用 OntoGazetteer 构建空间词汇词典；2）通过文本序列比对及模板聚类进行空间方位关系实例模板的泛化，结合 ANNIC 工具辅助人工归纳抽取模板，使用 JAPE 正则文法引擎编写空间方位关系抽取模板。在空间词汇词典与抽取模板的基础上，根据空间方位关系的描述特征扩展匹配算法，实现基于规则的空间方位关系抽取。

4.2 空间词汇的获取

对空间方位关系的发现与识别起着显著指示作用的空间词汇，是基于词典驱动抽取方法的主要组成部分。由于空间词汇较为宽泛，单靠人工归纳整理和通用词典整合无法高效的进行空间词汇的获取。因此，本文使用弱监督的

BootStrapping 方法针对领域语料进行空间词汇的自动迭代获取。

4.2.1 选择性约束与 BootStrapping 方法

自然语言的内容获取一般基于以下两点语言学的假设：

1）选择性约束原则（Principle of Selectional Restrictions）[134]：句法结构为其中的某些语言成分提供了相关语义内容的约束。这些由句法约束获得的新的语言成分即构成论旨角色。如图 2.8 中由句法结构可确定作为“界标/射体”的空间方位关系语义角色“繁峙县”/“平型关”，而空间谓词“位于”则是这一语义框架结构的约束核心。

2）概念的相互构建（Notion of Co-composition）[135]：在选择性约束原则下获取的各论旨角色，它们之间还存在着相互的语义构建过程。换句话说，这种共现性除了体现整体的句法约束外，也体现了在同一语义框架中的各论旨角色间相互的语义约束[136]。如图 2.8 中的界标、射体、空间词汇之间存在着相互的语义构建。以空间词汇“位于”为核心的空间方位关系语义框架中的槽，需要填入作为“界标/射体”的一组地理命名实体“繁峙县”/“平型关”，而界标/射体的概念则是相互构建的。

因此，通过分析相似的文法结构或共现现象，如句法分析树、组块结构、词的搭配等，可以归纳某一特定的语言成分。而通过观察已知语言成分的其他语言现象，则可增量的获取具有与之相似语义的其他语言成分。在语言学领域，这种通过迭代文法自举语义的过程，同时也被认为是人类获取语言能力的一种理论[137]。而在信息抽取领域，BootStrapping 是这类弱监督的自举学习方法的通称[138]。其优点在于：1）与基于熟语料库的监督学习方法不同，无需人工进行大规模的标注；2）学习结果多为丰富的、可解释的特征词或抽取模板，便于后处理；3）可以有效减少一般迭代算法循环依赖的风险[139]。

空间词汇迭代获取的流程如图 4.2：首先选择初始的种子空间词汇，将其放入种子空间词汇列表。根据设置的上下文参数，在语料中获取种子空间词汇相应的上下文模板，放入种子模板列表。然后在语料中遍历候选空间词汇，通过计算候选空间词汇上下文与种子模板的相似度，根据阈值进行筛选，获取新的空间词汇。根据新的空间词汇再获取新的上下文模板，并加入种子空间词汇和模板列表，进入下一轮迭代。当不再产生新的空间词汇及模板时，迭代停止。

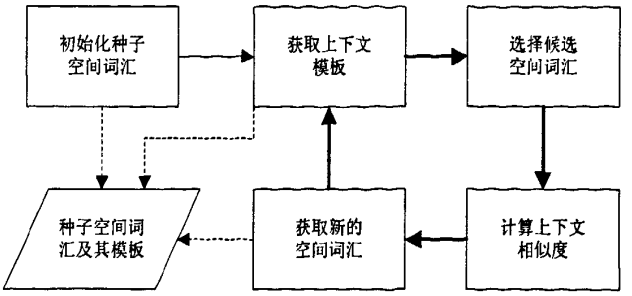


图 4.2 空间词汇的 BootStrapping 获取

4.2.2 空间词汇的迭代获取

4.2.2.1 上下文相似度计算

根据选择性约束的假设，如果两个词汇具有相似的上下文，则认为这两个词汇具有相似的语义。向量空间模型（Vector Space Model，VSM）是最常用的文本上下文相似度计算模型，在自然语言处理中有着广泛的应用。其计算公式如式 4-1：

$$sim(\vec{x},\vec{y})=\cos(\vec{x},\vec{y})=\frac{\langle\vec{x},\vec{y}\rangle}{|\vec{x}||\vec{y}|}=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}} \tag{4-1}$$

其中向量$\vec{x}$、$\vec{y}$是比较词汇的上下文特征向量。特征项的值定义如公式 4-2：

$$x_i=1-\frac{1}{d_i+2} \tag{4-2}$$

其中 d_i 是上下文中语言单元 i 与词汇的文本距离。该特征项的计算方法基于语言序列的物理距离假设，即距离特定词汇近的语言单元及其排列顺序对该词的语义约束较大。语言单元可以是字、词、地理命名实体，文本距离的计量可使用字数位移量或选定语言单元的位移量。语言单元的设置及文本距离计量方式直接影响上下文相似度的计算，进而影响迭代获取的效率与准确率。

Word	岷江	因	流经	古	青衣	国	而	得	名	。
Segment	nr	p	v	a	n	n	c	v	q	w
PLACE	110000				510000					
Segment &PLACE	110000	p	v		510000		c	v	q	w

图 4.3 上下文相似度计算中的语言单元

针对图 4.3 中的示例，以词汇“流经”为基准，如果语言单元选择分词获取的词汇，特征项选择词性，文本距离计量选择为词的位移，上下文窗口在句子的文本跨度内设为 4，则前后上下文特征向量构建为：(p-1/2, nr-2/3)与(a-1/2,n1-2/3,n2-3/4,c-4/5)。如果语言单元选择地理命名实体，特征项选择为类型编码，文本距离计量选择为字数位移，则前后上下文向量构建为：(110000-2/3)与(510000-1/2)。由上可见，词性作为特征项较地理命名实体类型编码更为丰富，而针对空间词汇的获取，地理命名实体类型编码也具有一定的约束作用，因此，可将两类标注进行合并构建新的语言单元及特征项。在本研究采用的地名辞典语料中，地理命名实体 PLACE 为人工标注，而分词 Segment 及词性标注则是通过 ICTCLAS 实现，后者的命名实体识别中虽包括地名标注，但相较地理命名实体而言，其内容较为简单。此外，针对地名辞典类文本行文简练而单句较长的特点，通用的分词软件分词性能受到一定的影响。因此两类标注的合并以 PLACE 为基准，所有文本跨度与 PLACE 没有重叠的 Segment 与之合并，记为 Segment&PLACE。对图 4.3 中的示例使用合并后的 Segment&PLACE 作为语言单元，词性或类型编码作为特征项，前后上下文向量构建为：(110000-2/3,p-1/2)与(510000-1/2)。

4.2.2.2 迭代算法

空间词汇的迭代获取算法伪码如下：

```
SpatialWords CaptureSpatialWords(Corpus, KeyWords,Parameters){
    //输入：语料 Corpus 种子词汇 Keywords 参数 Parameters
    //输出：空间词汇集 SpatialWords
(1) SpatialWords = SpatialWords + KeyWords
(2) While(SpatialWords.increase !=0)          //如果 SpatialWords 不再增长，则迭代结束
      //模板获取
(3)   (Foreach SpatialWord in NewSpatialWords)  //对每个新获取的空间词汇操作
(4)     ContextTemplate = getContextTemplate(SpatialWord)  //获取上下文作为候选模板
(5)     (Forall Pattern in PatternBase)
(6)       If(Sim(ContextTemplate,Pattern) < PatternDistictThreshold)
(7)         FilterPattern(ContextTemplate)       //候选模板过滤
(8)         PatternBase.add(ContextTemplate)  //获取具有足够区分度的模板
      //空间词汇获取
(9)   (Foreach Segment isn't in Place and in SpatialWords){
        //对每个不包含在 PLACE 中且不为空间词汇的词汇 Segment 操作
(10)    ContextTemplate = getContextTemplate(Segment) //获取 Segment 的上下文
(11)    (Foreach Pattern in PatternBase)
(12)      If(Sim(ContextTemplate,Pattern) > contextSimiThreshold)
(13)        FilterSpatialWords(Segment)          //空间词汇过滤
(14)        SpatialWords.add(Segment)          //获取具有相似上下文的空间词汇
```

在参数设置较为严格（上下文相似度阈值 ContextSimiThreshold 较大，模板区分阈值 PatternDistictThreshold 较小）的情况下，迭代次数与获取的空间词汇相对较少。反之，在较为宽松的参数情况下，迭代次数及获取的空间词汇相对较多，而空间词汇的准确率可能下降。具体的参数需根据语料情况进行设置。此外，为了提高迭代获取的准确率和效率，常通过对候选模板或空间词汇的过滤完成。对候选模板的过滤主要方法为：将候选模板重新用于文档集合中，验证其能否从文档集合中抽取出相关的文档，从而对候选模板进行评测[140]。对候选词汇的过滤则采用经验公式 TF*IDF 进行评测。其中 TF 指候选词汇对应的模板在模板库中的出现频率，即该词对模板库的代表性；而 IDF 等于语料中所有词数 N 和该词在语料中出现次数 n 的比值，反映该词的普及度[141]。如果候选词汇的 TF*IDF 值小于某个阈值，则将其过滤。由于空间词汇具有有限的词性及相对较低的词频，本文根据候选空间词汇的词性与词频进行筛选。

4.3 空间词汇的语义参数

在迭代获取空间词汇的基础上，通过设置其语义参数，构建以空间方位关系抽取为目标的空间词汇词典。本文通过词汇语义型词典的概念相似度和相关度计算进行空间词汇语义参数的设置。

4.3.1 语义型词典

语义型词典作为语言信息处理的基础，从单词、句法层面提取语义信息，并将这些信息以网状形式呈现，与传统的按字符顺序组织词汇信息的词典相比，更好的从词汇概念的角度出发，充分反映了词汇间的上下义、同义、对义、反义、部分-整体等语义关系，从而到达对关联度的可计算性[142]。目前主要的语义型词典分为两大类，一类是从句法层面提取语义的词典，如前文以格框架理论为基础的框架语义型词典 FrameNet[127]、Penn PropBank[128]；以动词的句法框架分类标准为基础的句法语义型词典 VerbNet[143]；另一类则是从词汇层面提取语义的词典，称为词汇语义型词典，应用最广泛的为 WordNet[144]。较之前一类词典，这类语义型词典虽然不是通过上下文来表达语义，在语义表达上侧重词义本身，但词汇覆盖面广，语义关系网络丰富是其优势。

由于面向英语的语义型词典在编纂内容上已具有相当的广度与深度，因此已广泛应用于信息检索、机器翻译、文本分类、自动问答等自然语言处理领域。在面向汉语的语义型词典的编纂方面，框架语义型词典的编纂工作刚刚起步[145][146]，而以“知网”（HowNet）为代表的词汇语义型词典的相关研究与应用则相对较为丰富。因此，在“句法分析-框架语义”这一信息抽取的思路在非受限

的中文自然语言中尚无法实现的情况下，充分利用词汇语义型词典获取词汇概念便成为可行的方法。

4.3.1.1 知网（HowNet）

知网（HowNet）是一个以英汉双语所代表的概念为描述对象，以揭示概念和概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[147]。知网中的“概念”是对词汇语义的一种描述，每个词可以表达为几个概念。知网中使用“义原”来描述概念以及各种语义关系，义原是用于描述一个概念的最小意义单位。知网使用其特有的“知识描述语言”(Knowledge Representation Language)形成的“语义描述式”来表达“义原-义原”、“概念-义原”、“概念-概念”间的关系，其结构图如下。目前网上可以免费下载的知网 2000 版，其通过 1503 个义原定义了 55501 个汉语词汇的 68383 个概念和 58582 个英语词汇的 76189 个概念。此外，其提供相应的编辑扩展工具及调用接口。

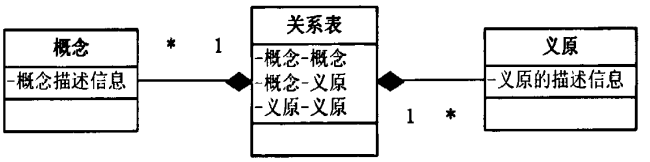


图 4.4 知网的关系网络结构图

“义原-义原”间在通过基本的上下位关系形成树状结构的基础上，进而通过同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系构成为一个复杂的网状结构。“概念-义原”及“概念-概念”间的语义描述式有三种形式：1）独立义原描述式，用“基本义原”，或者“(具体词)”进行描述；2）关系义原描述式，用“关系义原=基本义原”或者“关系义原=（具体词)”来描述；3）符号义原描述式，用“关系符号基本义原”或者“关系符号（具体词)”加以描述。关系义原类似于一种格关系，而符号义原大部分是格关系的“反关系”，表示可充当某个框架结构中的语义角色。为了可计算性，符号义原通过外延的形式来示范概念。主要的符号义原有#（与其相关)、%（部件-整体)、$/*（事件-工具，即宾格/主格)、+（隐性角色-事件)、&（指向，属性的宿主)、@（空间或时间)、?（材料-成品)、!（属性是宿主的典型属性)。

概念的描述都以“DEF=”为开始。概念中出现的所有义原或符号必须是在知网的分类体系中定义的义原或符号或者由知网知识系统描述语言所规定的特定标识符；概念描述中的第一个义原必须为该概念的基本意义，并用事件、实体、属性和属性值这四类义原中的一个描述出来；对于简单概念直接描述该概念的意

义；利用动态角色与特征（关系义原）来描述复杂概念；属性类概念必须标明其宿主；整体部分类型的概念必须标明该部分的整体等。

例如中文词汇“东西”具有四类概念：1）{Distance|距离}（030217）；2）{direction|方向}（030218）；3）{human|人}（030219）；4）{thing|万物}（030220）。其中 030217、030218 子概念描述如下：

NO.=030217
W_C=东西
G_C=N [dong1 xi1]
E_C=新加坡~30 公里南北 20 公里
W_E=from east to west
G_E=PP
E_E=
DEF={Distance| 距 离 :LocationFin={place| 地 方 :modifier={west| 西 }},LocationIni={place| 地 方:modifier={east|东}},host={earth|大地}}

NO.=030218
W_C=东西
G_C=N [dong1 xi1]
E_C=~走向，横贯~，~宽度，奔走~，~向，~两个方向
W_E=east and west
G_E=N
E_E=
DEF={direction| 方 向 :modifier={east| 东}};{direction|方向:modifier={west|西}}

以 030218 子概念为例，除了编号、拼音、英译、组合例词外，概念定义（DEF）为核心部分。其中，“{direction|方向}”、“{east|东}”、“{west|西}”为基本义原。同时，通过关系义原“{direction| 方向 :modifier}”构成“{direction| 方向:modifier={east|东}}”和“{direction|方向:modifier={west|西}}”两个的概念，而这两个概念再组合构成词汇“东西”的 030218 号子概念。

4.3.1.2 同义词词林扩展版

《同义词词林》是现代汉语较为常用的一部义类词典[148]。收录的词条按照树状层次结构分为大、中、小三类，形成一个由宽泛概念到具体词义的语义分类体系。在其基础上扩充完成的《同义词词林（扩展版）》[149]将词群与原子词群作为四、五级进行编码，形成五级层次分类，最终词表包含 77343 条词语。如下为同义词词林扩展版中 Cb02A 词群的主要词汇：

Cb02A01= 东南西北 四方
Cb02A02= 东 东方 东边 右 正东 东面 东头
Cb02A03= 南 南方 南边 正南 阳 南部 南缘 阳面
Cb02A04= 西 西方 西边 正西 左 西头 西部 西面
Cb02A05= 北 北方 北边 朔 正北 阴 北部 北缘 北头
Cb02A06# 东南角 东北角 西南角 西北角

其中第八位的标记有 3 种类型，分别为“=”，“#”，“@”。“=”表示该原子词群中的词汇相等、同义；“#”表示该原子词群中的词汇不相等、同类、相关；“@”表示该原子词群中的词汇相对封闭，既没有同义词，也没有相关词。

4.3.2 空间词汇的语义表达

对地理本体语义的研究中，文献[150]对 2005 版知网中与地理空间有关的实体

及其关系的概念进行了统计。获取实体概念 2133 个，按义原分为 13 大类：1）设置（Facility）、房屋（House）和建筑物（Building）797 个；2）场所（Institute Place）259 个；3）植物（Plant）32 个；4）土石（Stone）174 个；5）水（Water）和水域（Waters）215 个；6）地方（Place）364 个；7）陆地（Land）227 个；8）空域（Sky）31 个；9）大地（Earth）14 个；10）天象（Weather）。获取实体间语义关系概念 2007 个，按义原分为 6 大类：1）属性（Attribute）306 个；2）方向（Direction）38 个；3）位置（Location）398 个；4）部件（Part）953 个；5）材料（Material）254 个；6）现象（Phenomena）58 个。其中，实体概念及其关系概念对应的词汇中许多可作为空间词汇，按词性类型举例如表 4.1：

表 4.1 空间词汇的知网表达示例

词汇类型	空间词汇	语义表达
方位词及扩展	东南	direction|方向,east|东,south|南
	北郊	part|部件,%place|地方,surrounding|周围,#city|市,north|北
	北疆	part|部件,%country|国家,north|北
谓词	濒临	BeNear|靠近
	流经	flow|流
	纵贯	cross|越过
名词	山麓	part|部件,%land|陆地,base|根
	分水岭	part|部件,%land|陆地,space|空间,head|头
	上游	part|部件,%waters|水域
	河岸	part|部件,%land|陆地,#waters|水域,edge|边

在同义词词林的语义分类体系中，空间词汇以同义或相关的形式在同一中、小类集中体现。以动词类的空间词汇“邻接”为例，在较为接近的语义范围内，集中出现了可指示拓扑类型为 EC 及 DC 的相关空间词汇。另一方面，同义词词林中的词未进行概念的区分，如词汇“相交”没有空间关系这一类概念项，因此未与“交汇”作为同义词列出。

Id21A01= 交界 分界 接壤 邻接 毗邻 毗连

Id21B01= 靠近 邻近 接近 临近 挨近 贴近 将近 近乎 靠拢 ……

Id21B08# 沿边 沿江

Id21D01= 相距 距离 离开 离 去 距 偏离 相差

Id22B01= 相隔 分隔 隔 相间

Id23A01= 围绕 环绕 缠绕 盘绕 环抱 回环 拱抱 拱卫 围 圈 环 拱 绕 萦 缠

Id23A02= 包围 包 重围

Id23A04# 环城 环路 环线

4.3.3 概念相似度与概念相关度

4.3.3.1 概念相似度计算

从机器翻译的角度，概念相似度是指两个词在不同的上下文中可以相互替换使用而不改变文本的句法语义结构的程度[151]。概念相似度是一个主观性较强的概念。从宽泛的角度，词性的相同即体现了概念的相似度；从严格的角度，只有同义词才可以看作概念相似。因此，概念的相似度被定义为一个 0 到 1 之间的实数。概念相似度的计算分为两类，一类是基于分类体系（Taxonomy/Thesaurus）或语言学本体（Linguistic Ontology），一类是基于语料库的统计。知网和同义词词林的概念相似度计算属于第一类计算方法。

在知网特有的“知识描述语言”支持下，词汇的概念描述可表示如式 4-3。其中“[...]”表示特征结构，“{...}”表示集合。

$$
\text{实词概念}=\begin{bmatrix}
\text{第一基本义原描述}=\text{基本义原}_a \\
\text{其他基本义原描述}=\{\text{基本义原}_b,\text{基本义原}_c,\ldots\ldots\} \\
\text{关系义原描述}=\begin{bmatrix}\text{关系义原}_1=\text{基本义原}_x\mid\text{具体词}_x \\ \text{关系义原}_2=\text{基本义原}_y\mid\text{具体词}_y \\ \ldots\ldots\end{bmatrix} \\
\text{关系符号描述}=\begin{bmatrix}\text{关系符号}_1=\{\text{义原}_a\mid\text{具体词}_a,\text{义原}_b\mid\text{具体词}_b,\ldots\ldots\} \\ \text{关系符号}_2=\{\text{义原}_c\mid\text{具体词}_c,\text{义原}_d\mid\text{具体词}_d,\ldots\ldots\} \\ \ldots\ldots\end{bmatrix}
\end{bmatrix}
\tag{4-3}
$$

由于概念都最终归结于用义原（个别地方用具体词）来表示，所以义原的相似度计算是概念相似度计算的基础。根据义原的上下位关系树，可简单的通过语义距离计算其相似度：

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{4-4}$$

其中 p_1 和 p_2 表示两个义原，d 是 p_1 和 p_2 在义原层次体系中的路径长度，为一个正整数，α 是一个可调节的参数。在此基础上，考虑概念作为独立义原、关系义原和符号义原等几部分的整体，如何进行相似度的计算。设两个概念 A 和 B，A 可分解为$[A_1,A_2,\cdots,A_m]$，B 可分解为$[B_1,B_2,\cdots,B_n]$，那么这些部分之间的对应关系就有 m×n 种。因此，首先要做的工作是对两个概念整体中的各个部分进行一一对应，然后在对应部分之间进行比较。如果某一部分的对应物为空，将任何义原（或具体词）与空值的相似度定义为一个比较小的常数（δ）。最后，整体的相似度通过部分相似度的加权平均得到。可将其分为四个部分：

（1）第一基本义原描述 $Sim_1(S_1,S_2)$。其相似度按式 4-4 计算。

（2）其他基本义原描述 $Sim_2(S_1,S_2)$。对应与语义描述式中除第一基本义原描述式外所有基本义原描述式。其他义原描述式为多个且没有先验的对应关系，可按下列步骤对其分组：A）先把两个概念中的其他基本义原任意配对，计算出

所有可能的配对义原的相似度；B）取相似度最大的一对，将其归为一组。C）在剩下的基于义原的配对相似度中，取最大的一对，归并为一组，如此反复，直到所有基本义原都完成对应。

（3）关系义原描述式 $Sim_3(S_1,S_2)$。对应于语义描述式中所有的关系义原描述式。

（4）符号义原描述式 $Sim_4(S_1,S_2)$。对应于语义描述式中所有的符号义原描述式。

第一基本义原描述式反映了一个概念最主要的特征，但如果是加权平均的话，在 Sim_1 非常小而 Sim_3 或 Sim_4 较大的情况下，整体的相似度不能反映出第一基本义原的主导作用。因此，最终两个概念的整体相似度记为：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \tag{4-5}$$

其中，$\beta_i(1\le i\le 4)$ 是可以调节的参数，且有：$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\ge\beta_2\ge\beta_3\ge\beta_4$。后者反映了 Sim_1 到 Sim_4 对于整体相似度所起的作用依次递减。由于第一基本义原起主导作用，因此将其权值定为最大，一般应在 0.5 以上。

在同义词词林层次分类体系下，概念相似度可通过词汇所在类的编码进行正向匹配获得。对于进行比较的两个概念词汇，设 Sim=0，将其所在原子词群的五级编码由大到小进行匹配，相同的情况下 Sim 加 1，直到匹配不相同或匹配结束。Sim/5 即为其概念相似度。同上，可对每级编码设置权重。

4.3.3.2 概念相关度计算

概念相关度反映了具有共现可能性的两个词汇间相互关联的程度，其与概念相似度之间有着密切的联系。一般认为，两个词汇的概念相似度高，其相关度也较高；但两个词汇的相关度高，其相似度不一定高。例如：“纵贯”与“贯通”具有较高的相似度和相关度；“贯通”和“城区”则具有较高的相关度，而相似度不高。概念相关度是一个较为模糊的概念，与概念相似度一样，没有客观的衡量标准。概念相关度对基于规则的句法分析产生的结构性歧义具有较好的消岐作用[152]。参见例 4-1：

例 4-1

A）流经秦淮区的秦淮河（……）。

B）（……）流经秦淮区的夫子庙。

如果单在词性的基础上进行句法分析的话，将无法正确的区分这两个句子的句法树，因为其词性排列完全一致（vs/ns/b/ns）。而正确的句法分析树则如图 4.5。在例 4-1 中，A）是名词短语，而 B）为动词短语。

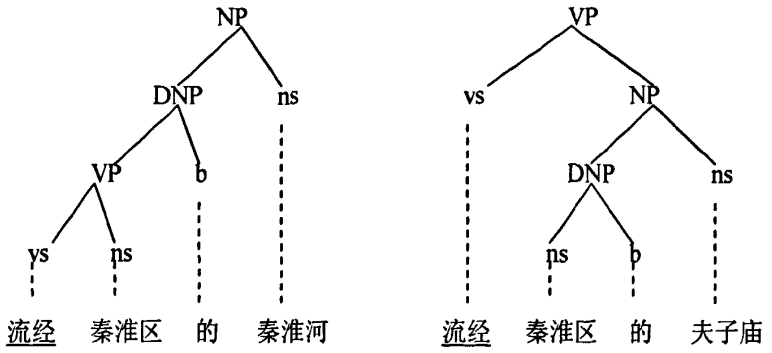


图 4.5 结构性歧义的句法树

通过概念相关度计算则可进行句法消岐，即相关度越大的两个词在句法树中的距离越短。从常识角度，通过关键字“河”、“区”、“庙”，可判断地理命名实体的类型分别为“河流”、“行政区划”、“景点”，而从空间谓词“流经”与“河流”、“行政区划”、“景点”的概念相关度可知，合理的句法结构是“流经[行政区划]/[景点]的[河流]”或“[河流]流经[行政区划]/[景点]”，即谓词动作的主体类型为“河流”，客体类型为“行政区划”或“景点”。因此 A）短语中“秦淮区”作为“流经”指示的 PO 类型关系的界标，而“秦淮河”作为射体；B）短语中，“夫子庙”作为“流经”指示的 PO 类型关系的界标，相应的射体未列出，该短语所在句的描述特征属于空间焦点转移。

知网的概念相关度计算研究相对较少。可在概念相似度的基础上，通过义原纵向和横向关系及概念的实例信息进行计算[153]。更一般化的概念相关性计算需将知网中各种语义关系全部提取出来，建立概念间、概念与义原以及义原间的内在联系，形成一个网状的信息表达结构。通过网状表示的任一信息入口，可以方便地访问与其相关的信息。网状结构生成的基本流程是[154]：（1）首先通过对各个义原文件的规范化处理，从中提取出不同义原的基本信息和义原间的关系描述，分别加入义原表和义原关系表中；（2）然后从概念词典中提取出每个概念（词语义项）的基本信息加入概念表，从各个概念的定义项中提取出概念与义原、概念与概念之间的关系及其组合关系，加入词典关系表中；（3）最后对这几张表进行排序，以关系表信息为触发设置其中的扇入扇出指针信息，形成一个完整的网状表示。

在同义词词林的层次分类体系中，其并没有特别的与概念相关对应的结构，因此其概念相关度与概念相似度计算相同，侧重通过对“#”标记的原子词群进行编码的正向匹配获得。不仅仅是“#”标记的原子词群，可将任一原子词群中词汇间的概念相关度记为 1（其概念相似度也为 1）。

4.3.4 空间词汇的语义参数设置

对于语义型词典中的空间词汇，可从其语义表达中获取用于空间方位关系抽取的参数。以空间方位关系表达式为准，该参数设为键值对字典：

{POS;Topology;Direction;Relevancy}

其中 POS 表示该空间词汇的词性。Topology 为指示的拓扑关系类型，歧义的情况需根据概率依次列出，如“东部”设置为 Topology=IN|DC。对于拓扑关系涉及的地理命名实体形成组合而可能的拓扑算子，在关系类型后声明，如“之间”设置为 Topology=IN(_)。Direction 为指示的方向关系类型。Relevancy 为该空间词汇相关的地理命名实体类型，类型划分粒度根据具体情况进行选择，并按界标/射体依次列出。本文定义粗颗粒的类型为{Water|水域,Road|道路,Land|陆地,Admin|行政区划}，则“流经”设置为 Relevancy=admin|land>water。界标/射体同类的不再重复，歧义的根据概率依次罗列或空缺。空间词汇语义参数示例如表 4.2。

表 4.2 空间词汇语义参数示例

词汇类型	空间词汇	语义参数
方位词及扩展	之间	{pos=f; topology=IN(_); }
	东南部	{pos=f; topology=IN|DC; direction=ES;}
	北疆	{pos=f; topology=EC; direction=N; relevancy=country>land;}
谓词	流经	{pos=v; topology=PO; relevancy=admin|land>water; }
	北起	{pos=v; topology=EC; direction=N; relevancy= land > mountain;}
	古名为	{pos=v; topology=EQ; relevancy=admin|land;}
名词	上游	{pos=n; topology=IN; relevancy=water; }
	分水岭	{pos=n; topology=EC(&); relevancy=water>mountain; }
	河畔	{pos=n; topology=EC; relevancy=water>land;}

语义参数中，POS 由词汇本身的词性进行设置，Direction 可根据表 2.3 空间方向关系类型对应的“前后左右上下”、“东南西北”等封闭词设置。而较为隐含的Topology与Relevancy可通过语义型词典中相关词汇的概念相似度与相关度计算进行设置。

由于知网中一个词汇具有多个概念义项，因此需选择与空间概念相关的义项进行计算，如上文的“东西”需选择 030217、030218 号义项。空间概念义项的选择可通过对上位义原的判断获取。定义空间方位关系的上位义原集合 SR 为{空间、方向、部件、界限、地方、距离值}，对于词汇所有的义项进行遍历，比较每个义项的上位义原集是否与 SR 存在交集，从而确定该义项是否被选取。

以概念相似度计算为基础，可通过 k-均值聚类进行 Topology 参数的设置[93]。设待聚类的空间词汇样本集合 $W=\{w_1,w_2,...,w_n\}$，n 为 W 所含的空间词汇数。选取具有不同 Topology 参数类型的典型空间词汇$\{c_1,c_2,...,c_k\}$作为聚类中心，k 为拓扑类型数。最终空间词汇的划分集合为$\{X_1,X_2,...,X_k\}$，而 X_i 中每个空间词汇的 Topology 参数由 c_i 确定。算法伪码如下：

(1) 根据聚类中心$\{c_1,c_2,...,c_k\}$初始化划分集合 $X_1,X_2,...,X_k$；
(2) 若 n>=1，继续；否则，转(7)；
(3) 取词汇样本集合 W 中第一个样本 x，令 n=n-1；
(4) 计算 x 与每个划分 X_i 的语义距离 S_i；
(5) 将 x 归于 S_i 最小的划分 X_i；
(6) 转(2)；
(7) 算法停止。

其中，语义距离的计算方法如下：首先使用公式 4-6 计算空间词汇 x 与划分 X_i 的概念相似度，其中 n_i 为划分 X_i 中的样本数，Sim_W(x,y)为特定语义型词典的概念相似度计算方式，针对知网采用公式 4-5 计算，针对同义词词林则通过编码正向匹配获得。接着使用公式 4-7 计算 x 与划分 X_i 的语义距离 S_i。

$$Sim\_X(x,X_i)=\frac{\sum_{j=1}^{n_i} Sim\_W(x,x_{ij})}{n_i} \tag{4-6}$$

$$S_i=\frac{1}{Sim\_X(x,X_i)} \tag{4-7}$$

针对知网的空间词汇概念表达方式，Relevancy 参数的概念相关度计算可通过义原描述式的内容与地理命名实体类型的相似度来获取。如表 4.1 中，“北郊”的符号义原“#city|市”、“山麓”的符号义原“%land|陆地”则直接指示了空间词汇相关的地理命名实体类型。

4.4 空间方位关系模板的获取

在基于模式匹配的规则抽取方法中，关系抽取模板的归纳与编写直接影响着抽取的性能。本文使用文本序列比及模板聚类进行空间方位关系实例模板的泛化，结合 ANNIC 工具辅助进行抽取模板的人工归纳。

4.4.1 文本序列比对

序列比对主要应用在生物信息中，基因学的一个主要主题就是比较 DNA 序列并尝试找出两个序列的公共部分[155]。如果两个 DNA 序列有类似的公共子序列，那么这两个序列很可能是同源的。Needleman-Wunsch 算法是一种整体联配（global alignment）算法，最佳联配中包括了全部的最短匹配序列，是一种动态

规划解决方案。如图 4.6 所示，两个不等长的序列“GAATTCAGTTA”与“GGATCGA”整体联配后等长，扩展的序列相互间重叠，字符间或字符与空格间对应。

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5	5
A	0	1	2	3	3	3	3	4	5	5	5	6

```
G_AATTCAGTTA
|  |  ||  |  |
GG_A_TC_G__A
```

图 4.6 Needleman-Wunsch 整体联配

通过对整体联配后的等长序列进行评分，可获取两个序列的文本相似度：如果一列为两个相同的字符，则给予得分 c（c>0），称为一个匹配；若为两个不同字符，给予得分 inc（inc<0），称为一个失配；若列中含有一个空格，则给予得分 g（g<0）。总分为每一列得分的和 Sum，而相似度则为 Sum/Length，Length 为等长序列的字符串长度。

4.4.2 实例模板的泛化

对于每个空间方位关系实例，可根据涉及的两个地理命名实体将其上下文分为 3 部分（对应为 S1，S2，S3）。定义实例模板为 InstancePattern(S1,S2,S3)。将实例模板的上下文中与空间方位关描述相关的标注进行合并，例如 PLACE&SIGNAL&Segment。通过对这些合并标注进行序列比对，计算其相似度，进行实例模板的聚类，辅助人工进行实例模板的泛化，从而获取抽取模板。参见例 4-2：

例 4-2

A）[白城市]市境南北有[平齐铁路]经过。

Segment: 白城市/ns 市/n 境/ng 南北/f 有/m 平/a 齐/a 铁路/n 经过/d 。/w

SpatialML: [白城市]/PLACE[市境]/SIGNAL[南北]/SIGNAL 有[平齐铁路]/PLACE [经过]/SIGNAL。

B）[青龙镇]镇后有[南宋酒瓶山]遗址。

Segment: 青龙镇镇/nr 后/f 有/v 南宋/t 酒瓶/n 山/n 遗址/n 。/w

SpatialML: [青龙镇]/PLACE[镇后]/SIGNAL 有[南宋酒瓶山]/PLACE[遗址]/SIGNAL。

在分词与空间词汇匹配的基础上，对上下文中的标注 Word、Segment 以及合并标注 SIGNAL&Segment、SIGNAL&Word 进行序列比对，并设置窗口间权重相同、c 为 1、inc 为 0、g 为 0，计算相似度。结果如表 4.3：

表 4.3 实例模板序列比对示例

合并标注	S2 序列比对	S3 序列比对	相似度（%）
Word	市/境/南北/有 <->镇/后/有	经过 <->遗址	8.34
Segment	n/ng/f/m <->nr/f/v	s <->s	70.00
SIGNAL&Segment	SIGNAL/SIGNAL/m <->nr/SIGNAL/v	SIGNAL <->SIGNAL	60.00
SIGNAL&Word	SIGNAL/SIGNAL/有 <->镇/SIGNAL/有	SIGNAL <->SIGNAL	75.00

取相似度最高的标注 SIGNAL&Word 进行实例模板泛化，简单情况下可将模板中不同部分用*代替，获取抽取模板的 BNF 范式为：

SpatialRelation::=<PLACE>(Segment)*(SIGNAL)(有)<PLACE>(SIGNAL)（4-8）

实际操作中，提供实例模板序列比对计算相似度的工具（图 4.7），以用户预设的实例模板为聚类中心，对于语料中所有其他模板进行聚类，获取相似度较高的前 n 个实例模板，辅助人工进行实例模板的泛化，获取抽取模板。

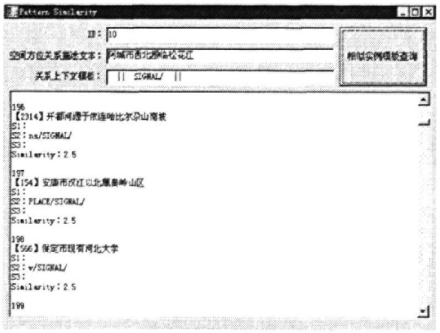


图 4.7 实例模板相似度查询界面

此外，使用 ANNIC（ANNotations-In-Context）辅助人工对实例模板中合并标注进行平行分析。通过使用 Lucene[156]对标注进行全文索引，ANNIC 不仅提供了检索叠置标注的查询规则，还提供了对标注结果的可视化（图 4.8）。

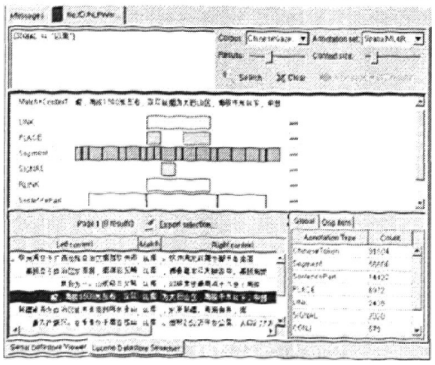


图 4.8 ANNIC 标注检索

4.5 空间方位关系抽取规则的构建

使用 OntoGazetteer 构建空间词汇词典，通过 JAPE 正则文法引擎实现抽取模板。在此基础上，根据空间方位关系的描述特征扩展匹配算法，实现基于规则的空间方位关系抽取。

4.5.1 空间词汇词典构建

通过词汇语义型词典进行概念相似度计算与相关度计算，实现空间词汇语义参数的设置，在此基础上构建以空间方位关系抽取为目标的空间词汇词典。为了便于语义参数模型以及词典本身的扩展，使用 Ontology 构建空间词汇词典。如图 4.9 所示，空间词汇“西北郊”与“北郊”作为多个空间关系语义参数类的子类。使用 Ontology 的优点在于：一方面，当语义参数类似 GUM-Space 进行扩展时，对空间词汇与其已有的关联不产生影响；另一方面，空间词汇与空间语义参数类间形成的本体结构有助于在词汇词典内部的相似度及相关度计算。

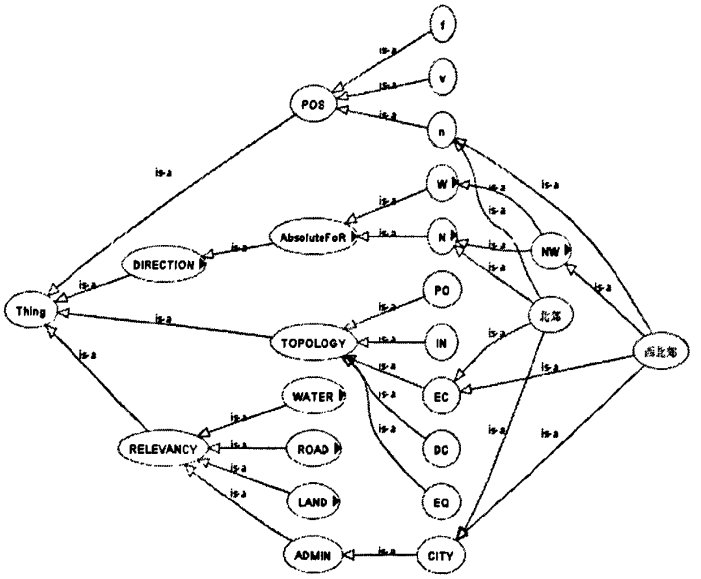


图 4.9 空间词汇“西北郊”与“北郊”的 Ontology 表达

在 GATE 平台中，采用 OntoGazatteer 实现空间词汇词典。其使用轻量级的 OWLIM（OWL In Memory）实现 Ontology 结构及其功能，而沿用传统的词典列表（Gazetteer）管理关键字词。在中图 4.10，上图为空间语义参数类的本体结构编辑，下图为某一语义参数类所有的空间词汇的文本编辑。这种将本体网状结构与词汇文本序列结构解耦的处理，兼容了传统的词汇管理方式，也为逐步采用 Ontology 进行大规模词典组织提供的基础。

图 4.10 OntoGazeteer 编辑界面

4.5.2 JAPE 正则文法模板引擎

在自然语言空间查询研究中多采用上下文无关文法进行查询语句的文法分析。在受限情况下，使用自上而下或自下而上的分析器构建句法树，进而获取查询语义，可以取得较好的性能。而面对自由文本，目前中文句法分析尚未达到实用。因此本文使用正则文法引擎作为抽取模板的具体实现。JAPE（a Java Annotation Pattern Engine）是 CPSL（Common Pattern Specification Language）的一个分支版本，针对标注提供了基于正则表达式的有限状态转换。JAPE 的每条规则由左侧（LHS,left-hand-side）和右侧（RHS,right-hand-side）两部分组成，左侧是一个包含正则表达式操作符（如*，?，+）的标注模板，右侧部分包含了对标注集合的操作代码，与左侧部分匹配上的标注集合将执行右侧的操作。如下例 JAPE 规则所示，LHS 实现了 BNF 范式 4-9 的正则表达，RHS 则是对匹配上的标注集合进行操作的 java 代码。

SpatialRelation::=<PLACE>(POS_v)<PLACE>(SIGNAL)+ （4-9）

```
Phase:    RLINK_Base
Input:  PLACE Segment SIGNAL
Options: control = appelt debug = true
/*exp: 阿里地区 位于 西藏自治区 西部*/
Rule: BaseRLINK_1
Priority:300
(
  ({PLACE}):target
  ({Segment.pos == v})
  ({PLACE}):source
  ({SIGNAL})+:signal
):tag-->
{ gate.AnnotationSet tagAS = (gate.AnnotationSet)bindings.get("tag");
  gate.AnnotationSet targetAS = (gate.AnnotationSet)bindings.get("target");
  gate.AnnotationSet sourceAS = (gate.AnnotationSet)bindings.get("source");
  gate.AnnotationSet signalAS = (gate.AnnotationSet)bindings.get("signal");
  ......
}
```

可见，作为正则文法，JAPE 虽然不能回溯，但由于 GATE 标注模型是建立在图（Graph）而不是序列（String）基础上，因此 JAPE 能表达比基于字符串的正则文法更为复杂的模式。图 4.11 是上例 JAPE 规则 LHS 部分在 GATE 标注模型上的匹配路径（箭头表示）。

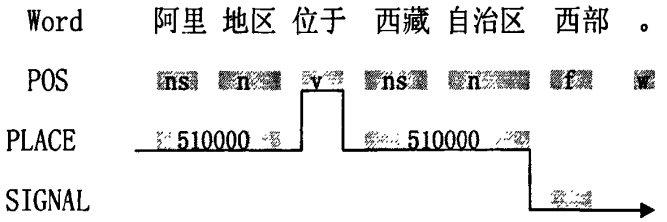


图 4.11 JAPE 规则的匹配路径

JAPE 的模板匹配有两种策略：1）Brill/All 方式，在文档的每个节点（Node）使用所有模板尝试匹配，匹配成功的模板即输出；2）Appelt 方式，在文档的任一节点，在匹配成功的模板中取最长的输出。一般情况下使用 Appelt 方式。

JAPE 与 GATE 的关联不仅表现在对标注模型的绑定上，GATE 为其提供了与平台结合的调试工具 Jape Debugger（图 4.12），并将 JAPE 规则看作一种特殊的语料资源。

图 4.12 Jape Debugger 编辑界面

4.5.3 空间方位关系抽取规则

在构建空间词汇词典的基础上，使用 JAPE 实现人工归纳的抽取模板，即可进行空间方位关系的抽取。其中空间词汇语义参数中 POS 与 Relevancy 在 LHS 中用于模板的约束，Topology 与 Direction 在 RHS 代码中设置具体的 LINK/RLINK 类型。对于 LHS 中的抽取模板，人工进行的归纳大多基于规整的空间方位关系描述实例，而实际语料中存在着大量省略、焦点转移以及组合的情况。因此，在抽取模板的基础上，需根据空间方位关系描述特征扩展匹配算法。

例 4-3

[岷山]是[岷江]、[涪江]、[嘉陵江]上源[白龙江]和[黄河]支流[白河]、[黑河]的分水源地。

如例 4-3，可将整句归纳为一条 JAPE 规则（其中匹配成功后的标注省略）：

```
(
    {PLACE}{Segment.pos==v}
        ({PLACE}{Segment.pos==w})+{PLACE}{SIGNAL_n}{PLACE}
        {Segment.pos==c}
        {PLACE}{SIGNAL_n}({PLACE}{Segment.pos==w})+{PLACE}
    {Segment.pos==d}
    {SIGNAL_n}
    {Segment.pos==w}
)
```

显而易见，该条规则的适用性很低。本文借鉴上下文无关文法的思路，使用分层匹配的策略进行空间方位关系的抽取。将抽取模板分成整体与局部两类，一类是有完整 Target/Source 的整体模板，如式 4-10；一类是只包括 Target/Source 之一的局部模板，如式 4-11。其中前一类模板又分为含空间域模板（式 4-12）与不含空间域模板。

SpatialRelation::= <PLACE>(SIGNAL_v)<PLACE>　　（4-10）

SpatialRelationPart::=<SIGNAL_v><PLACE>　　（4-11）

SpatialRelationWithDomain::=<PLACE><SIGNAL_v><Domain>

Domain ::= (Segment)+ (4-12)

含空间域的模板用以处理空间方位关系描述的组合与焦点转移情况，局部模板用以处理空间方位关系描述的省略情况。在此基础上，将空间方位关系的抽取分成两步：首先使用 Brill/All 方式进行 JAPE 模板匹配，获取所有可能的匹配作为中间结果；接着对获取的中间结果由外到内，由左到右逐层抽取空间方位关系。递归算法伪码如下：

(1) 设 Span = 整句的文本跨度；
(2) 在 Span 范围内由外向内，由左向右搜索中间结果；
(3) 如果没有，本次递归结束；
(4) 如果中间结果 SE 是含未知空间域的模板，
　　则 SE 入 SE 栈，待定 Domain 入 D 栈；设 Span=SE 的文本跨度，转(2)递归；
(5) 如果中间结果 SE 是局部模板
　　则取前驱 preSE 或后驱 postSE 的 Target 或 Source 设为缺失项
(6) 将中间结果 SE 设为最终结果，
　　将 SE 中 Target 或 Source 赋予 D 栈顶层；如果 D 栈空，则对应 SE 出栈；
　　转(3)继续搜索；

对于例 4-3 空间方位关系描述的组合情况，可归纳出三条作为整体模板的 JAPE 规则，其中 A）为含空间域的模板，三条规则都具有可重用性：

```
A) ({PLACE}{Segment.pos==v}
        ({Segment})+
        {Segment.pos==d}{SIGNAL_n}
   {Segment.pos==w}))
B) (({PLACE}{Segment.pos==w})+{SIGNAL_n}{PLACE})
C) ({PLACE}{SIGNAL_n}({PLACE}{Segment.pos==w})+)
```

使用 Brill/All 方式匹配并修正后，获取三组中间匹配结果（空间表达式 SE）：

（A） EQ(“岷山”,EC&({$Domain_i$|i=1,2,3…}))

（B） IN(“白龙江”,“岷江”) IN(“白龙江”,“涪江”)
　　IN(“白龙江”,“嘉陵江”)

（C） IN(“白河”,“黄河”) IN(“黑河”,“黄河”)

取文本跨度最长且存在未知空间域的 A）入栈，在 Domain 文本跨度内自左向右搜索到 B）组中间结果，由于其空间表达式不含未知空间域或缺项，将其设置为最终结果，且将“白龙江”赋予 Domain1。继续在 Domain 文本跨度内搜索到 C)组中间结果，同理将其设置为最终结果，且将“白河”、“黑河”赋予 Domain2，Domain3。完成 Domain 文本跨度的搜索，A）空间表达式完整，出栈。至此，空间方位关系组合描述各项完整，匹配结束。并可参照表 3.2 进行归一处理。

4.6 实验与分析

选用标注的《中国大百科全书（地理分册）》空间方位关系语料进行基于规则的空间方位关系抽取实验。实验分三个部分：1）基于 BootStrapping 空间词汇的获取。设置适当的参数，选取词性、地理命名实体类型编码以及两者的合并作为上下文相似度计算的特征项，比较不同词性的空间词汇作为种子词汇，以及不同特征向量对空间词汇自动获取性能的影响。2）基于概念相似度与相关度计算的空间词汇语义参数设置。应用语义型词典进行相似度及相关度计算，进行空间词汇的语义参数赋值。3）基于规则空间方位关系抽取方法评测。通过实例模板泛化辅助人工归纳训练语料中的抽取模板，在抽取模板及空间词汇词典支持下，评测基于规则的空间方位关系方法的性能。

4.6.1 实验一：空间词汇的获取

选用仅经过分词处理、标注地理命名实体的《中国大百科全书（地理分册）》语料进行基于 BootStrapping 的空间词汇迭代获取实验。首先，选取语料中频度最高的空间词汇“位于[v]”进行迭代，以确定相应的实验参数。经实验，设置模板区分阈值 PatternDistictThreshold 为 0.4，空间词汇上下文相似度阈值 ContextSimiThreshold 为 0.9，前后上下文相似度比例为 1。并将上下文相似度计算的文本跨度限制在句子内，选取窗口大小为 4 个文本单位。特征项分为以下三类：词性（POS），地理命名实体类型编码（PLACEType）以及两者的合并（POS&PLACEType）。选取三种空间词汇类型中较为典型且频度较高的方位词“北部[f]”（频度 20）、空间谓词“流经[v]”（频度 22）、名词“省辖市[n]”（频度 16），将其分别及全部作为初始种子词进行空间词汇的迭代获取实验。此外，对词性不为动词（v）、方位词（f）、名词（n）以及频度高于 100 的候选词汇进行过滤。实验获取的空间词汇是否有效可由熟语料中人工标注的 SIGNAL 确定，由于 SINGAL 的文本跨度常与分词获取的 Segment 不同，而在人工标注过程中仍有部分 SINGAL 未标注，因此对空间词汇是否有效最终由人工判断。实验结果见表 4.4。

表 4.4 空间词汇迭代获取的实验结果

种子词	特征项	迭代次数	获取模板数	获取空间词汇数	获取空间词汇示例	准确率（%）
北部	POS	5	7	32/169	别名[n] 主峰[n] 风景区[n] 支流[n] 驻[v] 相通[v] 南北[f] 中心[f]	18.93

	PLACEType	3	34	96/1023	自然保护区[n] 省会[n] 横贯[v] 濒临[v] 毗邻[vn] 东北[f] 南部[f]	9.38
	POS& PLACEType	4	13	29/110	中心[n] 分水岭[n] 纵横[nz] 分布[v] 穿过[v] 内[f] 中部[f]	26.36
流经	POS	4	8	21/153	驻地[n] 下游[n] 相望[v] 分布[v] 西北[f] 内[f] 西侧[f]	13.73
	PLACEType	3	34	142/4321	要隘[n] 中央[n] 素有[v] 纵贯[v] 濒临[vd] 直辖[v] 南北[f] 左侧[f]	3.29
	POS& PLACEType	4	11	34/201	遗址[n] 通道[n] 中心[n] 枢纽[n] 称[v] 建成[v] 北麓[f] 内[f]	16.92
省辖市	POS	3	8	37/322	良港[nr] 驿道[n] 中心[n] 垂直[vn] 云集[v] 中下游[f] 东南[f]	11.49
	PLACEType	3	35	221/3433	中心[n] 古城[n] 直辖市[n] 迂回[v] 开通[v] 南部[f] 南方[f]	6.44
	POS& PLACEType	4	14	52/313	分水岭[n] 交界处[n] 置[v] 属[v] 邻接[v] 东[f] 西南部[f]	16.61
北部流经省辖市	POS	3	6	42/182	支流[n] 驻地[n] 分水岭[n] 始建[v] 密布[v] 内[f] 北部[f]	23.08
	PLACEType	3	33	144/952	边界[n] 关口[n] 北坡[nr] 称为[v] 相距[v] 素有[v] 东南[f] 中下游[f]	15.13
	POS& PLACEType	4	13	40/112	分水岭[n] 中心[n] 名胜古迹[n] 流经[v] 分布[v] 称[v] 西[f] 附近[f]	35.71

由实验结果可知：

（1）采用不同词性的空间词汇作为种子词汇以及选取不同的特征项进行上下文计算，通过 BootStrapping 迭代获取新的有效空间词汇，都涵盖了空间词汇的三种词性类型。其中，方位词作为典型的空间词汇，使用其作为种子词汇进行迭代的准确率最高。

（2）由于地名辞典类语料的领域特性，地理命名实体较为密集，因此当单纯选用 PLACE 作为特征项时，特征向量结构简单，迭代次数相对较少，而获取的模板与空间词汇较多，同时准确率较低。相比较而言，选用 POS 作为特征项时，特征向量较为丰富，获取的有效空间词汇较多。而当选用 POS&PLACEType 构建更为丰富的特征向量时，迭代获取的准确率最高。

（3）当使用三种词性类型的空间词汇同时作为种子词汇时，由于初始种子模板的约束，使迭代获取的准确率最高。

4.6.2 实验二：空间词汇的语义参数设置

（1）方位词的语义参数设置

方位词语义参数 POS 为 f，Direction 根据表 2.3 空间方向关系类型对应的“前后左右”、“东南西北”等封闭词字面进行设置。由于“北麓”等方位词扩展置于

名词类空间词汇讨论，因此方位词的 Relevancy 设置为空。实验集中在方位词 Topology 参数的设置。

文献[93]首先手工将知网 2005 版中的方位词分为 61 个绝对方位词和 133 个相对方位词，通过使用空间词汇“里”、“外”和“附近”作为内部概念类、外部概念类和边缘概念类的初始聚类中心，针对 133 个相对方位词进行层次聚类，效果较为理想。本文使用上述三个典型空间词汇，添加词汇“周围”，使用同义词词林获取同义概念，设置相似度范围为词群，并人工筛选出方位词。实验结果见表 4.5。

表 4.5 方位词的 Topology 语义参数设置

方位词 / Topology	绝对方向方位词	相对方向方位词	
	实验语料统计	知网聚类	同义词词林聚类
IN	东部，西部，南部，北部，东南部，东北部，西南部，西北部	里：里头，中，内，里边，里面，当中，中间，内里，内部，内侧，际，中心，之中，之间，当间，中部	里：内，中，里边，里头，以内，之内，内中，其中，此中，中间，其间，里面，内部，间，个中，之中
DC	东面，西面，南面，北面，东南面，东北面，西南面，西北面，之东，之西，之南，之北，东方，西方，南方，北方，东北方，西南方，西北方，东南方，东边，西边，南边，北边，东南边，东北边，西南边，西北边	外：外头，之外，外边，外面，以外，外侧，外边	外：外面，外边，外头，外界，以外，之外，外围，外侧
EC	东头，西头，南头，北头，东端，西端，南端，北端，东侧，西侧，南侧，北侧，以东，以西，以南，以北，东南角，东北角，西南角，西北角	附近：跟前，以近，旁，旁边，边上，以远，周围，一边，一面，两边，两头，两旁，四面，四外，四周，边，侧，四围，四周围	附近：就近，邻近，邻，近处，一带，内外，近旁，不远处 周围：四周，四围，方圆，四邻，四下里，郊，四郊，四下，四周围，四旁
拓扑类型不定	东，西，南，北，东西，南北，东西向，南北向，东南，东北，西南，西北，东南偏东，东南偏南，东北偏东，东北偏北，西南偏西，西南偏南，西北偏西，西北偏北	上，下，前，后，上头，下头，前头，后头，左，右，之下，之前，之后，上边，下边，前边，后边，左边，右边，上面，下面，前面，后面，左面，右面，上下，前后，左右，里外，底下，面前，背后，对面，这儿，这里，那儿，那里，下部，前部，后部，中部，左侧，右侧，左上方，左下方，右上方，右下方，左上角，左下角，右上角，右下角，前后左右，头，尾，面，顶	

可见，使用知网或同义词词林进行概念聚类，可对辐轴方向（D 类）和泛方向（E 类）方位词的 Topology 语义参数进行几乎相同的设置，辐轴方向集中在

DC 和 IN，泛方向集中在 EC，表明这两类方位词具有明确的拓扑描述功能。对于相对方向中的 A 类和 C 类方位词，以及绝对方向方位词（B 类），可将“前后左右上下”或“东南西北”作为初始聚类中心，使用知网或同义词词林进行聚类获取 Direction 语义参数。但与仅需通过字面进行设置的方法相比，意义已不大。

依据地理空间方位关系描述的领域性，本文通过对语料中空间词汇指示的 LINK 关系的频度统计设置绝对方向方位词的 Topology 参数。如语料中 20 例空间词汇“北部”，在指示方向为 N 的 RLINK 关系的同时，都指示拓扑为 IN 的 LINK 关系。由于 RLINK 的参照框架为 INTRINSIC 与 RELATIVE 的实例过少，因此，对相对方向中剩余的 A 类和 C 类方位词不进行设置。由实验可见，在地理方位关系描述中，“部”、“头”、“以”等文字对以指示方向为主的方位词的拓扑类型的指示具有一定作用。

（2）空间谓词的语义参数设置

空间谓词语义参数 POS 为 v，Direction 根据包含的相应方位词部分设置，为简便，示例使用“北”代表 A 类、C 类和 B 类方位词（下同）。实验集中在 Topology 与 Relevancy 参数的设置。

针对实验语料中空间谓词，将空间词汇“相距”（DC）、“邻接/会合/靠近/环绕”（EC）、“经过”（PO）、“隶属”（IN）、“别名/简称”（EQ）作为初始聚类中心，进行 Topology 类型聚类。首先，在知网中查询空间谓词与空间方位关系相关的概念义项，通过相似度计算划分类型；如果知网中没有相关义项，则使用同义词词林进行相似度计算；最后根据语料中具体的标注类型，对错误的划分及未划分的空间谓词进行人工修正。在上述处理流程中，同步使用概念相关度计算获取空间谓词的 Relevancy 参数。实验结果示例见表 4.6。

表 4.6 空间谓词的语义参数设置示例

Topology / Relevancy	DC	EC	PO	IN	EQ	拓扑类型不定
Water		源出，发源，濒临，滨临，相汇，（自北）汇入，（自北）流入，汇合，交汇，注入，流至	横渡，流经，流贯，流过			流向，汇向
Road		北延，北达，通达，直达，北至，接轨	纵贯，横贯，贯穿			通往，并行，北沿
Land		北逾，北起	横卧	蜿蜒，绵延		
Admin		接壤，毗连		隶属，隶，属，直辖，驻，管辖，划归，原属		

相关类型不定	相离，相隔，北隔，相距，相间	相切，相连，邻接，靠近，临近，北邻，北临，北连，通，紧依，链接，衔接，为界，交界，分界，交会，相会，会合，围绕，环绕，背依	相交，经过，切穿，斩切，穿越，穿过	包含，被包含，包括，座落，构成，组成，分布于，集中在，地处	相等，简称，别名，又称，又名，得名，改名，语意为，古名为，原名，别称	朝向

可见，空间谓词的描述多样性主要集中在 EC/PO/IN 类型中，这与自然语言空间关系研究中对空间谓词的选择相一致。这三种类型的中文空间谓词更是与地理命名实体类型密切相关。而 DC 与 EQ 类型的空间谓词概念较为单一。此外，存在拓扑关系不定的指向性空间谓词，其功能类似与方位介词，多与方位词同时使用。

（3）名词类空间词汇的语义参数设置

名词类空间词汇的语义参数 POS 为 n，Direction 根据包含的相应方位词设置。实验集中在 Topology 与 Relevancy 参数的设置。从部分-整体关系的角度，名词类空间词汇侧重描述作为整体地理命名实体和其组成部分的地理命名实体间的方位关系，强调边缘与内部的区分，一般没有外部概念。针对实验语料中的名词类空间词汇，将“边境/周围”（EC）、“境内/中心”（IN）作为初始聚类中心，进行 Topology 类型聚类，处理流程同空间谓词的聚类。实验结果示例见表 4.7。

表 4.7 名词类空间词汇的语义参数设置示例

Topology / Relevancy	EC	IN
Water	源，源头，发源地，北源，入海口，入河口，交汇处，河畔，岸，河岸，海岸，湖岸，江岸，北岸，沿岸，两岸，交汇处，环湖，背山濒湖	上游，上源，上下游，下游，中游，中上游，中下游，支流，分支，北支，主流
Road	终点站，起点站，路边，沿线，路旁	路段，北段，北行线，中间站，中点站，驿站，交通枢纽，中途要站
Land	天然屏障，界山，分水岭，过渡地带，分水源地	北麓，北隅，北支，北坡，最高峰，第一高峰
Admin	城郊，市郊，北郊，郊区，外城，城郊结合部，边境，边境线，国境，省境，边陲	城内，内城，境内，首都，省会，经济中心，北境，北区，辖县，辖市，核心地区，工业中心，著名县份（*）
相关类型不定	边界线，周围，分界点，边界，边缘，交接处	中心部位

可见，名词类空间词汇即是描述某一地理命名实体类型组成部分的词汇，因此几乎都具有 Relevancy 参数。此外，在空间方位关系描述的领域文本中，大量形如“著名省份”、“新兴工业区”这类本身对空间方位关系指示不显著的词汇，也被设置为 Topology 为 IN 的空间词汇（参见 3.1.2 节）。

综上，使用词汇语义型词典进行空间词汇语义参数的设置，具有一定的效果。但由于词典是通用的而非针对空间方位关系领域，在设置过程中还需较多的人工修正。因此，需在构建更为完整的、专业化的地理本体知识库的基础上，结合语言学本体进行概念相似度与概念相关度的计算，从而提高空间词汇语义参数设置的自动程度和准确率，构建更为完备的空间词汇词典。

4.6.3 实验三：空间方位关系抽取

对实验语料中的空间方位关系实例模板进行统计，结果为表 4.8。可见在地名辞典类空间方位关系描述文本中，大量实例模板是紧凑而规整的。

表 4.8 空间方位关系实例模板频度

频度	实例模板	空间方位关系描述示例
186	NA|SIGNAL/|NA	[漠河县县府]驻[西林吉镇]
74	NA|SIGNAL/|SIGNAL/	[大柴旦镇]位于[柴达木盆地]中北部
69	NA|NA|NA	[平果县][右江]
33	NA|v/|NA	[环市铁路]连接[城区]

实验首先随机选择 20 篇语料 CorpusA，20 篇语料 CorpusB，根据 CorpusA 与 CorpusB 中的空间词汇构建空间词汇词典；接着，在空间词汇词典的基础上，通过文本序列比及模板泛化，协助人工归纳抽取模板，确保可全部准确识别 CorpusA 中的空间方位关系；然后，使用归纳的抽取模板以及空间词汇词典对 CorpusB 进行抽取实验。实验分为两类，一类在抽取时提供全部空间词汇词典，一类在抽取时仅提供 CorpusA 的空间词汇。每类实验都使用普通与层次匹配两种规则方法，且分为发现任务（找出存在的空间方位关系实例）和识别任务（识别出空间方位关系的类型）。由于语料的领域性以及地理命名实体与抽取模板的约束，评测指标采用识别率（即后文的召回率），即识别出的空间方位关系与正确的空间方位关系的比值。

针对 CorpusA 归纳的抽取模板 BNF 范式如下：

```
SpatialRelation::=
    (<PLACE>(Conj_p|Conj_v)<PLACE>(Signal_n|Signal_f)+)
    |(<PLACE>(Pos_p|Pos_v)<PLACE>(Signal_n|Signal_f))
    |(<PLACE>(Signal_f)+(Signal_v)(Pos_p)+<PLACE>)
    |(Signal_n|((Signal_f)+DistanceExpress)))
    |(<PLACE>(Signal_n)(Pos_p)+<PLACE>)
    |(<PLACE>(Conj_c)<PLACE>(((Pos_b)?(Signal_n))|(Signal_v)))
    |(<PLACE>(Pos_w|Pos_c)<PLACE>(((Pos_b)?(Signal_n))|(Signal_v)))
    …
    <EOF>
DistanceExpress::=
```

(Distance_prefix)+<Pos_m><Pos_q>(Distance_suffix)+

…

实验结果如表 4.9：

表 4.9 空间方位关系抽取实验结果

识别率 / 规则方法	仅提供 CorpusA 空间词汇词典		提供全部空间词汇词典	
	发现任务	识别任务	发现任务	识别任务
普通	69.93%	47.55%	78.32%	67.13%
层次匹配	72.73%	50.35%	83.92%	70.63%

可见，对于特定的空间方位关系描述文本，通过半自动的构建空间词汇词典以及适量的文法规则，可取得较好的抽取效果。其中，空间词汇对空间方位关系抽取中识别任务的影响最大，发现任务次之。此外，本文提出的层次匹配策略对性能的提升具有一定的作用。

4.7 本章小结

本章通过构建空间词汇词典，结合关系模板及匹配规则，研究基于规则的空间方位关系抽取方法。首先使用 BootStrapping 迭代获取空间词汇，利用词汇语义型词典对空间词汇进行语义参数的设置；然后通过文本序列比对及模板聚类进行空间方位关系实例模板的泛化，结合 ANNIC 辅助人工归纳抽取模板；最后，使用 OntoGazetteer 构建空间词汇词典，JAPE 正则文法引擎编纂抽取模板，在此基础上通过扩展匹配算法实现基于规则的空间方位关系抽取方法。

第5章 基于支持向量机的空间方位关系抽取方法

5.1 方法的提出

基于规则的空间方位关系抽取方法中，在标注语料的基础上，主要的人工工作包括以下几个部分：空间词汇自动获取后的人工判断；空间词汇语义参数的修正以及词典的构建与维护；泛化获得的抽取模板的选取及匹配规则的编写。基于规则的关系抽取方法的优点在于最终的模板规则具有可解释性，并可不断修正。但同时，其需要较多的人工参与，编写模板规则的周期长，应用成本高；另一方面，当抽取系统被应用于新的领域时，需要对模板规则进行修改。

基于机器学习的关系抽取方法是目前应用比较广泛的方法。ACE 评测中，采用机器学习方法进行的实体关系抽取效果要明显高于基于规则的方法。不同于使用规则抽取的方法，当使用机器学习方法进行关系抽取时，不再依赖大规模的知识库及专家系统，需要的是提供可进行机器学习的实体关系标注语料，即训练数据。一般将这些训练数据构造成特征向量的形式，通过诸如支持向量机[157]（Support Vector Machine，SVM）、Winnow[158]等基于感知机（Perception）的学习算法构造分类器。然后，通过这些分类器来判断测试数据中是否存在所需的实体关系。

如同 BootStrapping 方法中计算上下文相似度使用的向量空间模型，可将序列文本中某一类型的现象，通过构建特征向量进行数值化的表达。一个实例可以转化为特征向量 X，其中 xi 为 n 维特征向量 X 的第 i 个元素。而基于特征向量的机器学习算法就是在一组训练数据(X1,y1),(X2,y2),...,(Xn,yn)，其中对于二元分类问题 yi ∈{-1, +1}，学习一个分类函数 f，使得给定新的特征向量 X'，f 能够将其正确的分类，即 f(X')= y'。分类函数 f 可看作一个由权值向量 ω 决定的超平面，其能够将标号为-1 和标号为+1 的数据严格分开。而该权值向量的求解过程由各种机器学习算法完成。

基于特征向量的机器学习在建立学习模型时，需要考虑“过配”（overfitting）与“低配”（underfitting）问题。如果构造模型有太多参数，会导致模型过于依赖训练数据，而不能较好地对其它特征不完全相同的实例进行分类，这即为“过配”。相反，如果模型过于泛化，则学习的结果无法很好的反映训练数据蕴含的规则约束，即为“低配”。

5.2 统计学习理论与支持向量机

5.2.1 统计学习理论

传统统计学可认为是在样本数目趋于无穷大时的渐近理论，是以经验风险最小化（Empirical risk minimization,ERM）作为目标的统计方法。与之相比，统计学习理论（Statistical Learning Theory，SLT）是一种专门研究小样本情况下机器学习规律的理论，其以结构风险最小化理论（Structural risk minimization,SRM）为基础[159] 。ERM 归纳学习原则通过最小化学习机在训练数据集上的经验误差作为选取输出函数的条件，其理论基础来自大数定律，即经验风险和期望风险之间存在双边一致收敛公式。但当大数定律不满足时，ERM 归纳学习的泛化性能无法得到保证。而 SRM 通过同时缩小经验误差和函数的复杂性作为选取函数的条件，其合理性来自于风险泛函的界公式，该不等式至少以概率 1- η 成立：

$$R(f) \le R_{emp}(f) + \sqrt{\frac{h(\ln(2l/h)+1) - \ln(\eta/4)}{l}} \qquad (5\text{-}1)$$

其中 R(f)为函数 f 的期望风险，Remp(f)为函数 f 的经验风险，l 为样本量，而 h 是 VC 维（Vapnik Chervonenkis dimension），用以衡量函数类的容量，也就是函数结构的复杂性。公式 5-1 可简化为：

$$R(f) \le R_{emp}(f) + \Phi(l/h) \qquad (5\text{-}2)$$

其中 Φ(l/h)称为置信区间，是与函数类结构相关的一个量化指标。函数容量项 h 是函数类 F 的一个特性，并不是单个函数 f 的特性。因此，误差界不能仅仅通过 f 优化。让集合 S 代表函数类 F(D, θ)，$\theta \in \Lambda$ 的结构。一系列 S_k={ F(D, θ), $\theta \in \Lambda_k$}构成了一个嵌套结构，而每个集合 Sk 的 VC 维 h_k 形成一个序。

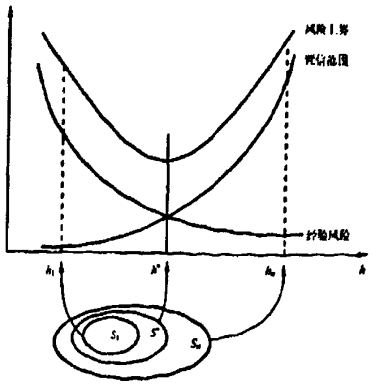


图 5.1 风险的界是经验风险与置信区间之和

这样，风险泛函取决于经验分析和置信区间（图 5.1）。如果选取的 VC 维

小的函数类，将导致大的经验风险；如果选择 VC 维大的函数类，虽然经验风险会变小，但是置信区间将会增大。学习机最小的期望风险来自于经验风险和置信区间的折衷[159]。1989 年，Vapnik 与 Chervonenkis 发现了经验风险最小化归纳原则和最大似然方法一致性的充要条件，完成了对经验风险最小化归纳原理的分析：如果数据服从某个（固定但未知）分布，要使学习机的实际输出与理想输出之间的偏差尽可能小，学习机应当遵循结构风险最小化原则，而不是经验风险最小化原则。也就是说应当使风险泛函的上限最小化[160]。

目前，统计学习理论仍是小样本统计与预测的主要理论。其既有严格的理论基础，又能较好的解决小样本、非线性、高维度和局部极小等问题。

5.2.2 支持向量机

支持向量机（Support Vector Machine，SVM）是在统计学习理论基础上出现的一种机器学习方法，其来源于二值分类问题中最优超平面具有最好推广能力的思想，兼顾了训练误差和泛化能力。SVM 算法利用了 Mercer 核、凸二次规划、稀疏解和松弛变量等技术。

超平面作为一个决策曲面，将正样本与负样本隔离开来。在线性可分的情况下，构造超平面就是使公式 5-1 右边的经验误差为 0，并使置信区间最小。这样构造的分类超平面具有可控的统计特性。对于样本集$\{(x_i,y_i)|i=1,\ldots,l\} \in R^n \times \{-1,1\}$,可在某个内积空间 H 上，构造分类超平面：

$$\langle w,x\rangle + b = 0, w \in H, b \in R \tag{5-3}$$

w 是一个可调的权向量且垂直于超平面，b 是偏值。对于可分的样本 x，总可以调整 w 和 b，使得$\forall x$存在：

$$\langle w,x\rangle + b \geq 1, y = 1; \langle w,x\rangle + b \leq -1, y = -1 \tag{5-4}$$

而<w,x>+b=1 和<w,x>+b=-1 即为两个标准超平面（canonical hyperplane）。而落在其上的样本点称为支持向量（Support vector,SV），支持向量机由此得名。如图 5.2 所示，SVM 具有很好的几何解释。

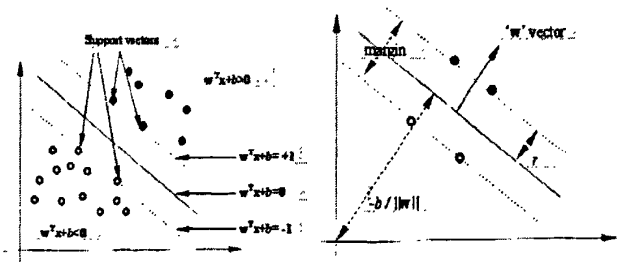


图 5.2 SVM 分类的几何描述

支持向量到分类超平面的距离 r=1/‖w‖，而分类间隔（Margin）ρ 为：

$$\rho = 2r = \frac{2}{\|w\|} \tag{5-5}$$

SVM 中最简单清晰的模型即为最大间隔分类器，其通过具有分开正负样本的最大间隔的超平面（即最优超平面）来优化学习机的泛化误差界。而更为复杂的 SVM 大都基于此思想。要获取最大的分类间隔，即最小化‖w‖，也就是‖w‖^2=w^{T}w。因此，线性可分的 SVM 的优化问题即为最小化泛函：

$$\Phi(w) = \frac{1}{2}\langle w, w\rangle \tag{5-6}$$

其约束条件为公式 5-4，公式中 1/2 仅为计算上的方便。该问题称为 SVM 的主问题，是一个凸二次优化问题，一般通过构造 Lagrange 乘子的方法转成对偶问题解决，本文不再展开。

对于无法使公式 5-1 右边的经验误差为 0 的情况，也就是不可能构造一个不具有分类误差的超平面。这种情况下，可以构造一个超平面，使其对整个训练集平均的分类误差的概率最小。为此，引入了一组非负的标量变量$\{\xi_i \mid \forall i\}$到超平面，该变量称为松弛变量（Slack variable），用于度量一个数据点对模式可分的理想条件下的偏离程度（如图 5.3）。

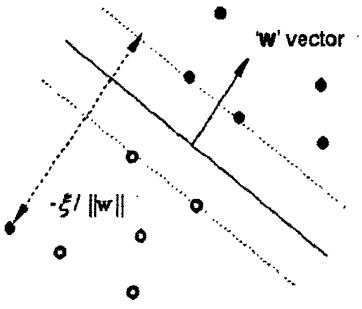


图 5.3 引入松弛变量的超平面

5.2.3 非线性分类与核函数

对于数据内在关系是非线性的情况，SVM 首先将输入向量经非线性变换映射到一个高维（或者无限维）的特征空间，在变换后的高维空间获得一种线性关系，使得 SVM 具有非线性分类的能力[161]。此外，Cover 定理也指出数据在映射到高维空间中更容易线性划分。

如公式 5-7 为一个非线性映射的示例：

$$\Phi: R^2 \to R^3$$

$$(x_1, x_2) \to (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{5-7}$$

其几何意义如图 5-4，在 2 维空间非线性的情况，在 3 维空间线性可分。

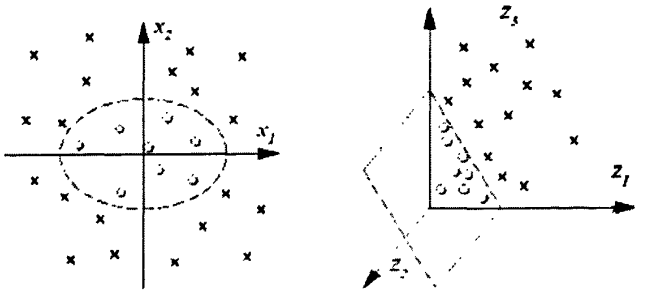


图 5.4 非线性映射

将其一般化，设该映射为Φ，其将样本空间 R^n 映射到某个高维的特征空间 H，该映射记为：

$$\Phi: R^n \to H \tag{5-8}$$

如公式 5-4 所示，在对高维空间 H 中的样本进行操作时，SVM 训练及测试过程都只采用样本的内积形式，即$<\Phi(x_i),\Phi(x_j)>$。因此，如果存在

$$k(x,x') = \langle \Phi(x), \Phi(x') \rangle \tag{5-9}$$

则可以通过核函数 $k(x,x')$ 间接的实现映射 $x \to \Phi(x)$，获得在高维特征空间中的线性特性。因此，核函数的意义体现在两个方面：（1）实现了低维空间 R 到高维空间 H 的隐式转换；（2）极大的简化了计算，即不需要知道映射Φ的具体实现，计算只发生在核函数上。而且一般来说，R 的维度远远小于 H，通过核函数，使高维特征空间中的内积计算变得容易实现。一个函数能成为核函数的充要条件有如下的 Mercer 定理给出：

对于任意对称函数 $k(x,y)$，它是某个特征空间内积运算的充分必要条件是，对于任意不恒等于零的函数 $\varphi(x)$ 且 $\int \varphi^2(x)dx < \infty$，均满足：

$$\iint k(x,y)\varphi(x)\varphi(y)dxdy \geq 0 \tag{5-10}$$

常见的核函数有：

高斯核（Gaussian RBF）：

$$k(x,x') = \exp\left(-\frac{\|x-x'\|}{2\sigma}\right), \sigma > 0 \tag{5-11}$$

拉普拉斯核（Laplacian RBF）：

$$k(x,x') = \exp\left(-\frac{|x-x'|}{\sigma}\right), \sigma > 0 \tag{5-12}$$

多项式核（polynomial）：

$$k(x,x') = (\langle x,x'\rangle + c)^d, d \in N, c \geq 0 \quad (5\text{-}13)$$

Sigmoidal:

$$k(x,x') = \tanh(a\langle x,x'\rangle + b) \quad (5\text{-}14)$$

5.2.4 多分类方法

SVM 本质上是一种二值分类的机器学习方法。但在实际应用中，更多的是多分类的情况，如 LINK 关系五种类型的区分。使用 SVM 进行多分类主要有两种思路：一种为整体方法，即从整体上构造一个复杂的约束优化；另一种为分解方法，即将多分类问题转换为二值分类问题。

整体方法是从统计学习理论出发，把多个决策函数联合成一个整体决策函数，然后对该整体决策函数引入结构风险。虽然整体方法具有较好的推广能力，但如果仍采用二次规划求解，算法复杂度将随分类的类别呈指数增长。

分解方法把多分类问题分解为二值分类问题处理，通常有 one-against-others 和 one-against-another 两种。One-against-others 在多分类问题分解时，将某一类型的实例看作二值分类学习中的正例，而将其它类型的实例都看作二值分类学习中的反例，最终通过区分“此类型”与“非此类型”来识别该类型，进而实现多分类。相反，one-against-another 则首先对多个类型进行两两组合，完成多类问题的分解，然后直接对每个组合的实例进行二值分类的学习，例如对某个组合(C1,C2)，将属于 C1 的实例看作 C1 类型的正例，而将属于 C2 的实例看作 C1 类型的反例，反之亦然。这样直接在每对组合间构建最优超平面而实现多分类。

分解方法对每个二值分类问题各自引入结构风险，该方法虽然简单，推广能力较整体方法稍差，但由于其算法复杂性随分类类别呈线性增加，因此具有较好的实用性。

5.3 基于支持向量机的空间方位关系抽取

5.3.1 空间方位关系的实例

从机器学习的角度，实例（Instance）是指一个具体的待学习或待识别的语言现象，在本文中即为一个空间方位关系的描述。这个实例的属性（Attribute），即为这个实例的上下文特征，可通过特征向量得以形式化。而每个实例都有一个类型（Class）对应，本文中即为一种具体的空间方位关系类型。

如图 5.5，对于句子“[青龙镇]位于[上海市]西部[青浦县]东北境。”而言，可认为有两组存在空间方位关系的实例，分别是“青浦县”/“青龙镇”的类型为 IN 的 LINK 关系、类型为 NE 的 RLINK 关系；“上海市”/“青浦县”的类型

为 IN 的 LINK 关系、类型为 W 的 RLINK 关系。S1、S2、S3 分别为空间方位关系实例中两个地理命名实体之前、中间和之后的上下文。在以句子的文本跨度为上下文界限时，S1、S3 可能为空；而对于地址中的相邻地址要素而言，S2 为空。

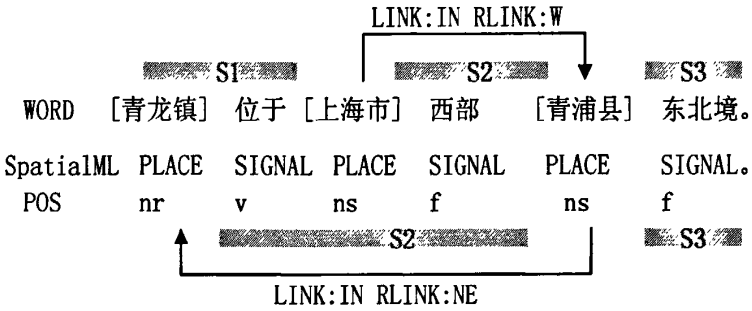


图 5.5 空间方位关系实例

空间方位关系抽取可看作对于一定上下文中出现的两个地理命名实体所构成的实例赋予一个具体的空间方位关系类型，公式表达如下：

$$\text{Instance}(S_1, place_1, S_2, place_2, S_3) \rightarrow C \qquad (5\text{-}15)$$

其中 $place_1$ 和 $place_2$ 表示两个地理命名实体，上下文 S1、S2、S3 的特征项类型及窗口大小根据具体的学习情况设定，C 为空间方位关系类型。设 $X=\{x_i|i=1...n\}$表示所有实例组成的集合，n 是语料中实例的总数。集合 X 中有 m 个实例分别被标注为 $y_i(y_i \in \{c_j|j=1...R\})$，$c_j$ 表示空间方位关系类型，R 为类型总数。这 m 个实例称为训练实例，而集合 X 中剩余的 l(l=n-m)个实例未标注，这部分实例称为测试实例。空间方位关系抽取就是通过对 m 个已标注的训练实例的学习，对 l 个未标注的测试实例进行类型分类。

在图 5-5 中，如果将无空间方位关系也作为类型的一种，则“青龙镇”/“上海市”、“上海市”/“青龙镇”可看作类型为 NULL 的空间方位关系实例（依据标注规范，在标注及抽取阶段不进行空间推理）。如果将忽略 Source/Target 指向的空间方位关系也作为类型的一种，则“青龙镇”/“青浦县”也可看作类型为 T_IN 的 LINK 关系实例、类型为 T_NE 的 RLINK 关系实例（T 表示 Source/Target 转置的空间方位关系）。“青浦县”/“上海市”类推。

5.3.2 空间方位关系的特征向量

根据空间方位关系实例的上下文而构建的特征向量，需要满足区分性和代表性。在通用的关系抽取中，可选的特征项包括词形、词性、关系涉及的命名实体及其包含关系、语义型词典提供的概念信息等[80]。在基于 Winnow 算法的汉语空间关系射体的识别中，针对其描述特点，选用的特征项除词性和概念信息外，还

可包括以下特征：（1）动词特征，将待处理的词语与其所在的句子中每个动词的距离作为一类特征；（2）“把被”特征，将处理的词语与其所在子句中的“把”字或“被”字的距离作为一类特征；（3）空间表达式特征，将待处理的词语与空间表达式的距离作为一类特征[95]。而在选定特征项的基础上，还需考虑上下文窗口的大小。特征向量维度的增加可以提高抽取性能。相应的，抽取颗粒度越细，越需要构造丰富合适的特征向量。针对空间方位关系的描述特点以及标注语料的情况，本文拟采用词性、地理命名实体词形及类型、空间词汇词形及语义参数作为特征项构建特征向量：

（1）词性（POS）

一个空间方位关系实例上下文中词汇的词性，是最基本的特征项。词性作为特征项整体上虽然较为稀疏，但在空间方位关系抽取这一特定领域，方位词、方位介词及动词的区分作用还是能较好的体现出来。词性特征可选择 Source 前后窗口的词性串、Target 前后窗口的词性串、Source 与 Target 间的词性串。

（2）地理命名实体（PLACE）

在空间词汇的 BootStrapping 获取中，将词性与地理命名实体合并作为特征项，合并后的特征项更能反映空间方位关系描述的特点。该合并特征可选择 Source 前后窗口的特征串、Target 前后窗口的特征串、Source 与 Target 间的特征串。此外，根据概念相关性原则，一个空间方位关系实例涉及的两个地理命名实体的词形或类型，与空间词汇语义参数中的 relevancy 具有对应关系。因此，可将其作为特征项。

（3）空间词汇（SpatialWord）

空间词汇对空间方位关系起着重要的指示作用。除了空间词汇的词形外，通过语义型词典获取的语义参数能进一步的区分空间方位关系的类型。空间词汇特征项可选择 Source 前后窗口的空间词汇词形、Target 前后窗口的空间词汇词形、Source 与 Target 间空间词汇的词形、以及空间词汇语义参数中 topology、direction、relevancy 参数。

对于图 5.5 中“青浦县”/“青龙镇”这一空间方位关系实例而言，如果取上述特征项构建特征向量，较为丰富的示例如下：

EC
_POS_S1_NA _POS_S2_v/ns/f _POS_S3_f
_POSPLACE_S1_NA _POSPLACE_S2_v/place/f _POSPLACE_S3_f
_PLACE_Former_青龙镇 _PLACE_Later_青浦县
_PLACEType_Former_510000 PLACEType_Later_510000
_SpatialWord_S1_NA _SpatialWord_S2_位于/西部 _S3_SpatialWord_东北境
_SpaitalWord_S1_linktype_NA _SpatialWord_S1_direction_NA
_SpatialWord_S2_linktype_IN/IN|DC _SpatialWord_S2_direction_NA/W

_SpatialWord_S3_linktype_EC _SpatialWord_S3_direction_NE

其中，NA 表示为空的特征值。该特征向量的构建未考虑 Source/Target 的指向，地理命名实体对以前/后（Former/Later）来区分，且较为简洁的将 S1、S2、S3 中的特征项各自作为一个整体构建特征向量。在具体应用中，如果考虑指向性以及特征项按窗口距离分开设置，则特征向量将相对复杂。从示例可见，S1、S2 中的空间词汇及其语义参数对本实例的类型判断起干扰作用。由于空间词汇的这种强指示作用，对空间词汇特征的窗口范围需根据语料情况进行选择。

5.3.3 空间方位关系抽取流程

基于 SVM 的空间方位关系抽取流程如图 5.6，包括以下 7 个步骤：1）针对第三章标注的空间方位关系语料，将其根据抽取的不同任务生成相应实例，包括学习所需的正例和反例；2）根据不同的评测方法，将实例分为训练实例和测试实例，也就是训练语料和测试语料；3）将训练实例和测试实例根据特征项及上下文窗口的设置，映射为特征向量；4）调用 SVM 的具体实现及相应参数对训练实例进行训练；5）获取训练结果，即 SVM 的分类模型；6）使用分类模型对已映射为特征向量的测试实例进行分类，获取分类结果；7）将测试实例的分类结果与实际分类情况比较，进行抽取性能的评测。

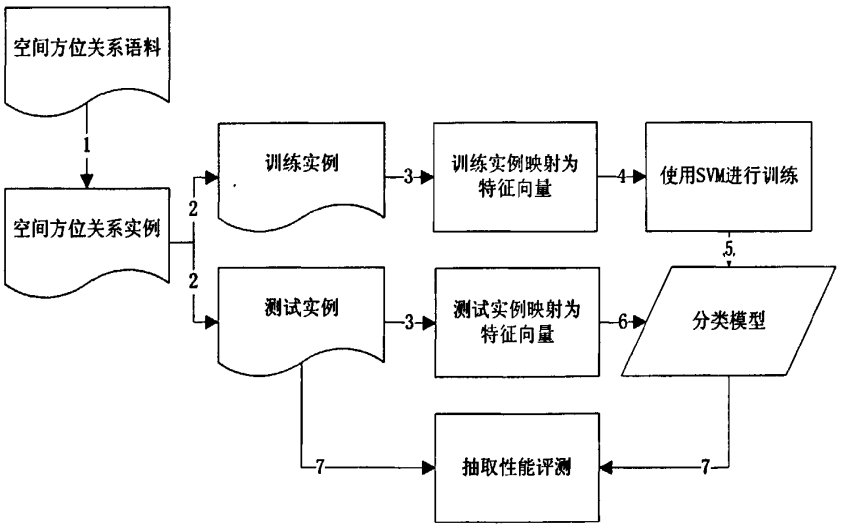


图 5.6 基于 SVM 的空间方位关系抽取流程

5.4 实验与分析

5.4.1 评测指标

对抽取性能的评测是通过对已人工正确标注的测试标注（Key Annotation）与分类模型获取的抽取标注（Response Annotation）进行量化对比而获取。两类

标注间存在以下六种关系：

（1）Coextensive（同现）：指两个标注具有同样的文本跨度，也就是 StartNode 和 EndNode 相同。相当于 Allen 一维拓扑[162]中的“相等”。

（2）Overlap（叠置）：指两个标注在文本跨度上具有叠置部分。这种关系除了 Allen 一维拓扑中的“相交”，还包括“内相接”和“包含”。

（3）Compatible（一致）：指两个标注在同现的情况下，抽取标注的标注特征包含测试标注的标注特征。

（4）Partially Compatible（部分一致）：指两个标注在叠置的情况下，抽取标注的标注特征包含测试标注的标注特性。

（5）Missing（未识别）：指不存在与测试标注同现或叠置的抽取标注，或是在同现或叠置情况下，抽取标注的标注特征中不存在测试标注的标注特性。

（6）Spurious（误识别）：指不存在与抽取标注同现或叠置的测试标注，或是在同现或叠置情况下，测试标注的标注特性中不存在抽取标注的标注特性。

实验采用信息抽取领域常用的 F-Score[163]作为抽取性能的评测指标。该指标由 Precision（准确率）、Recall（召回率）和 F-Measure（F 值）三部分组成。计算参数包括：1）Correct（正确标注个数）：抽取标注与测试标注一致的个数；2）Partilly Correct（部分正确标注个数）：抽取标注与测试标注部分一致的个数；3）Missing（未标注个数）：未识别的测试标注；4）False Positive（错误标注个数）：误识别的抽取标注。计算公式如下：

$$\text{Strict Precision}=\frac{Correct}{Correct+PartillyCorrect+FalsePositive} \tag{5-16}$$

$$\text{Strict Recall}=\frac{Correct}{Correct+PartillyCorrect+Missing} \tag{5-17}$$

$$\text{Lenient Precision}=\frac{Correct+\alpha PartillyCorrect}{Correct+\alpha PartillyCorrect+FalsePositive} \tag{5-18}$$

$$\text{Lenient Recall}=\frac{Correct+\alpha PartillyCorrect}{Correct+\alpha PartillyCorrect+Missing} \tag{5-19}$$

$$\text{F-Measure}=\frac{(\beta^2+1)\times Precision\times Recall}{\beta^2\times Precision+Recall} \tag{5-20}$$

其中，Precision 与 Recall 都存在严格（Strict）（公式 5-16）（公式 5-17）与宽松（Lenient）（公式 5-18）（公式 5-19）之分，前者不考虑 Partilly Correct 的贡献，后者根据 α 的权重设置而考虑 Partilly Correct 的贡献。α 可取 0-1 间的值，在命名实体、组块等连续文本的识别中通常取值为 0.5。β 是 Recall 和 Precision 的相对权重，一般取值为 1。

在本文对于空间方位关系抽取的评测中，主要关注两个地理命名实体间的空

间方位关系是否能正确的被发现及识别，对于关系实例的文本跨度并不严格。如图 5.5 中，“青浦县”/“青龙镇”这一 LINK 关系，其严格的文本跨度为整句。在规则抽取中，由于空间词汇在 S3 中，因此匹配后的文本跨度为整句。而在 SVM 抽取中，由于在构建候选实例时文本跨度固定，即前 PLACE 始到后 PLACE 末，因此识别后在标注特征正确的情况下，因文本跨度不为整句而视为 Partially Compatible。因此，针对空间方位关系抽取性能的评测，本文取 Lenient Precision 和 Lenient Recall 计算 F-Measure 指标，并取 α 、β 值为 1。

由于基于 SVM 的空间方位关系抽取本质上是通过对所有待定实例的分类进行关系识别，在待定实例确定的情况下，总会赋予一个类型。因此需使用 F 值对准确率和召回率综合后进行评测。而在基于规则的抽取中，由于在地理命名实体、空间词汇、以及抽取模板与匹配规则的多重约束下进行空间方位关系的抽取，False Positive 的数量很少，因此对基于规则的抽取评测使用的是识别率（即召回率）。

除了评测指标外，还需选择划分训练语料与测试语料的方法。目前最常用的语料划分方法分为两类：一类是 k 分组交叉验证（k-fold cross-validation），这种评测方法首先随机的将语料等分为 k 组，然后将每一组作为测试语料，而其它组作为训练语料；一类称为 hold-out 验证，这类评测方法是由系统随机的选择一些语料作为测试语料，而其它的则作为训练语料。根据标注语料的情况，本文选用后一种评测方法。

5.4.2 实验过程与分析

选用的《中国大百科全书（地理分册）》空间方位关系语料进行基于 SVM 的空间方位关系抽取实验。将标注好的语料随机按 3：2 划分为训练语料与测试语料。其中训练语料共有 LINK 关系 1506 例，RLINK 关系 380 例。其中 IN 关系 913 例，PO 关系 275 例，EC 关系 155 例，EQ 关系 111 例，DC 关系 52 例。测试语料共有 LINK 关系 930 例，RLINK 关系 135 例。其中 IN 关系 573 例，PO 关系 182 例，EC 关系 74 例，EQ 关系 78 例，DC 关系 23 例。

根据语料的情况，抽取实验分为 LINK 关系抽取和 RLINK 关系抽取两个部分，且采用无 Source/Target 指向性的空间方位关系实例。其中 LINK 关系抽取分为两个任务：（1）发现任务，即判断两个地理命名实体间是否存在 LINK 关系；（2）识别关系，即判断两个已存在 LINK 关系的地理命名实体间属于哪一种 LINK 类型。而由于 RLINK 关系类型较多而导致训练实例较少，因此（3）RLINK 关系的抽取仅为发现任务，即判断两个地理命名实体间是否存在 RLINK 关系。所有任务都从词性（POS）特征或词性与地理命名实体的合并（POS&PLACE）特征开始，逐步增加关系实例涉及的地理命名实体特征项和空间词汇特征项。从

而评测各特征项对抽取性能的影响。

实验选用 LibSVM[164]作为 SVM 的具体实现。通过封闭测试实验，选用多项式核（Polynomial kernal）和“one-against-another”的多分类方法。其他参数使用 LibSVM 默认设置。

5.4.2.1 实验准备：窗口设置

针对 LINK/RLINK 关系，在不考虑 Source/Target 指向性的情况下，分别使用 POS 和 POS&PLACE 作为语言单元，统计前 PLACE 前的 S1、前 PLACE 和后 PLACE 间的 S2、后 PLACE 后的 S3 的平均距离。对 S1、S3 统计时分两种情况，S1 表示以 LINK/RLINK 标注文本跨度的开始为前界，S1+表示以 LINK/RLINK 标注所在句子的文本跨度开始为前界；S3 表示以 LINK/RLINK 标注文本跨度的结束为后界，S3+表示以 LINK/RLINK 标注所在句子的文本跨度结束为后界。S2’表示在前 PLACE 和后 PLACE 存在有指向性的一对多情况下，取文本跨度最小的一组进行统计，即去除省略的情况。统计结果见表 5.1 和表 5.2：

表 5.1 LINK 关系上下文窗口距离统计

	S1+	S1	S2	S2’	S3	S3+
POS	2.96	0.33	8.52	3.03	3.28	11..39
POS&PLACE	1.66	0.17	4.44	2.93	1.78	7.14

表 5.2 RLINK 关系上下文窗口距离统计

	S1+	S1	S2	S2’	S3	S3+
POS	2.08	0.19	7.32	2.17	1.84	12.99
POS&PLACE	1.13	0.08	3.63	1.85	1.11	8.06

由统计结果可见，以 LINK/RLINK 标注的文本跨度为界，大部分情况下 S1 距离为 0，S3 也较小。实际描述中，S1+及 S3+中增加的特征对空间方位关系的影响不大。S2 从以 POS 作为文本距离单位到采用 POS&PLACE 时的显著减小，反映了省略及焦点转移的情况较为普遍，且在除去省略后，紧凑的空间方位关系中 S2’的文本距离较小。据此，构建 LINK/RLINK 关系特征向量时的窗口设置见表 5.3。

表 5.3 LINK/RLINK 关系特征上下文窗口设置

	LINK			RLINK		
	S1	S2	S3	S1	S2	S3
POS	1	3/9	3	1	7/2	2
POS&PLACE	1	3/4	2	1	4/2	1

其中，S1 与 S3 的起始位置分别为前 PLACE 的开始节点与后 PLACE 的结

束节点。针对普遍存在的省略情况，S2 窗口设置为 m/n，表示当 S2 的文本距离小于 n 时，取以前 PLACE 的结束节点后 m 个语言单元构建特征向量；当 S2 的文本距离大于 n 时，取以后 PLACE 的开始节点前 m 个语言单元构建特征向量。从而克服由于省略带来的特征稀疏。

将前 PLACE 和后 PLACE 为基准，空间词汇作为语言单元，对 LINK/RLINK 关系中空间词汇位置的统计见表 5.4 和表 5.5。

表 5.4 LINK 关系空间词汇位置统计

	-3	-2	-1	1	2	3
前 PLACE（相关）	0	2	24	389	111	45
前 PLACE（无关）	52	60	100	133	134	127
后 PLACE（相关）	73	147	353	227	85	23
后 PLACE（无关）	97	131	93	128	123	102

表 5.5 RLINK 关系空间词汇位置统计

	-3	-2	-1	1	2	3
前 PLACE（相关）	0	0	0	114	36	15
前 PLACE（无关）	5	12	14	54	41	24
后 PLACE（相关）	18	50	134	81	11	6
后 PLACE（无关）	26	67	40	78	61	37

由统计结果可见，相关的空间词汇集中在前 PLACE 的后 1 个文本距离以及后 PLACE 的前后各 1 个文本距离。因此，构建空间词汇特征项时，取前 PLACE 的后 1 个 SIGNAL 和后 PLACE 的前后各 1 个 SIGNAL。

5.4.2.2 实验一：LINK 关系抽取

使用上一步获得的窗口参数构建特征向量，进行 LINK 关系的抽取实验。其中 LINK 关系发现任务分别在两种范围的实例中进行。一种情况是针对篇章中所有的句子，对每句中两两地理命名实体构成无指向性的实例，存在 LINK 关系的为正例，不存在的为反例；另一种情况是在包含 LINK 关系的句子中，两两地理命名实体构成无指向性的实例。LINK 关系发现任务的实验结果见表 5.6。

表 5.6 LINK 关系发现任务实验结果

实例范围 / 特征项	篇章中的句子			含 LINK 关系的句子		
	准确率 P	召回率 R	F 值	准确率 P	召回率 R	F 值
POS	62.77	19.28	29.50	66.76	65.03	65.89
+PLACEType	59.85	26.80	37.02	67.87	65.32	66.57
+SIGNAL	65.02	47.39	54.82	65.31	60.40	62.76
POS&PLACE	57.14	28.76	38.26	63.66	63.29	63.48

+PLACEType	54.74	24.51	33.86	60.10	64.45	62.20
+SIGNAL	59.43	41.18	48.65	62.05	59.54	60.77

由实验结果可见：（1）针对以篇章中的句子作为实例范围的情况，初始 POS&PLACE 作为特征项的性能要优于 POS。随着各自特征项的增加，抽取性能都有显著的提高，SIGNAL 最为明显。而最终 POS+PLACEType+SIGNAL 的性能要优于 POS&PLACE+PLACEType+SIGNAL。（2）针对以含 LINK 关系的句子作为实例范围的情况，由于正反例相对平衡，因此在初始使用 POS 或 POS&PLACE 作为特征项时，即获得相对（1）而言较好的抽取性能。但之后特征项的增加并未明显提高抽取性能，SIGNAL 的添加反而稍稍降低了性能。

LINK 关系识别任务的实验结果见表 5.7。

表 5.7 LINK 关系识别任务实验结果

类型 / 特征项	DC			EC			PO			IN			EQ		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
POS	0.0	0.0	0.0	0.0	0.0	0.0	20.69	10.53	13.95	67.48	90.53	77.33	45.00	29.03	35.29
+PLACEType	0.0	0.0	0.0	12.50	2.56	4.26	50.00	26.31	34.48	69.81	91.36	79.14	42.31	35.48	38.60
+SIGNAL	80.00	33.33	47.09	40.00	15.38	22.22	56.66	29.82	39.08	73.37	93.00	82.03	70.83	54.84	61.81
POS&PLACE	0.0	0.0	0.0	0.0	0.0	0.0	23.08	5.26	8.57	65.01	91.77	76.11	37.5	19.35	25.53
+PLACEType	0.0	0.0	0.0	21.43	7.69	11.32	56.52	22.81	32.50	69.16	91.36	78.72	36.36	25.81	30.19
+SIGNAL	0.0	0.0	0.0	46.67	17.95	25.93	51.51	29.82	37.78	73.55	93.83	82.46	65.22	48.39	55.56

由实验结果可见：（1）五类 LINK 关系实例不平衡，IN 实例最多，而 DC 实例很少。因此，IN 识别性能相对较好，而 DC 则较差。（2）PLACEType 与 SIGNAL 特征项的增加，能显著提高识别任务的性能。其中 PLACEType 的添加对 EC/PO 类型的识别性能提高较明显，说明这两种 LINK 类型涉及的地理命名实体类型特征较为显著。而 SIGNAL 的添加对 DC/EC/EQ 类型的识别性能提高较明显，说明这三种 LINK 类型相应的空间词汇较为集中。

5.4.2.3 实验二：RLINK 关系抽取

使用上文获得的窗口参数构建特征向量，进行 RLINK 关系的抽取实验。其中 RLINK 关系发现任务也分别在两种范围的实例中进行。RLINK 关系发现任务的实验结果见表 5.8。

表 5.8 RLINK 关系发现任务实验结果

实例范围 / 特征项	篇章中的句子			含 RLINK 关系的句子		
	准确率 P	召回率 R	F 值	准确率 P	召回率 R	F 值
POS	0.0	0.0	0.0	56.10	37.10	44.66
+PLACEType	0.0	0.0	0.0	45.95	27.42	34.34
+SIGNAL	31.58	23.53	26.97	55.55	40.32	46.73

POS&PLACE	0.0	0.0	0.0	50.00	30.65	38.00
+PLACEType	0.0	0.0	0.0	43.18	30.65	35.85
+SIGNAL	22.86	15.69	18.60	39.53	27.42	32.38

由实验结果可见：（1）针对以篇章中的句子作为实例范围的情况，由于正反实例的不平衡，POS+PLACEType 与 POS&PLACE+PLACEType 的 F 值都为 0，SIGNAL 的添加使性能显著提高。而最终 POS+PLACEType+SIGNAL 的性能优于 POS&PLACE+PLACEType+SIGNAL。（2）针对以含 RLINK 关系的句子作为实例范围的情况，由于正反例相对平衡，因此在初始使用 POS 或 POS&PLACE 作为特征项时，即获得相对（1）而言较好的抽取性能。但之后特征项的增加并未明显提高性能。

综上可见：（1）针对 RLINK/LINK 发现任务，鉴于现有的语料量，在已确定存在空间方位关系的句子中，基于支持向量机的抽取方法具有较好的发现性能，且不依赖于空间词汇的识别，具有一定的实用意义。此外，加入空间词汇后，在篇章中对空间方位关系的发现也具有一定泛化能力。(2)对于 LINK 识别任务，对已确定存在 LINK 关系的地理命名实体对，在空间词汇的约束下，基于支持向量机的抽取方法具有较好的识别性能。由于目前语料中实例的不平衡，各 LINK 类型的识别性能不同。

与基于规则的方法相比，基于支持向量机的方法对空间词汇的依赖较规则方法弱一些。因此，当前者在未登录空间词汇的情况下完全失灵时，后者则能依靠上下文中的其他特征项进行空间方位关系的抽取。

5.5 本章小结

本章选用关系抽取中性能最佳的支持向量机进行基于机器学习的空间方位关系抽取。首先引入了统计学习理论与支持向量机方法，分析了其适用于空间方位关系抽取的两个特性：结构风险最小化原则与使用核函数实现到高维特征空间的非线性映射。然后论述了空间方位关系的实例化方法，特征向量构建及抽取流程。最后通过发现任务及识别任务实验，分析基于支持向量机的空间方位关系抽取方法的适用性。

第6章 空间方位关系抽取的应用

在地理信息检索、基于位置服务、自然语言空间查询等应用领域，空间方位关系抽取的引入，使对文本空间信息的挖掘更为深入，应用模式得以扩展。本章以文本驱动的地理知识库构建以及空间方位关系的图形重建为例，简要的论述空间方位关系抽取的应用场景。

6.1 应用一：地理知识库构建

6.1.1 文本驱动的地理知识库构建

地理知识库，尤其是地理本体的研究是当前 GIS 的研究重点[165, 166]。主要的研究包括地理本体的形式化表达、地理本体的语义互操作、地理空间信息的语义网络服务、语义地理信息系统的应用等方面[150]。SFB_TB8 是一个融合了 19 个子课题，旨在从推理、行为和交互三方面对空间认知的机理及其应用进行全面研究的在研课题[167]。其中 I1-OntoSpace 是研究空间本体（Spatial Ontology）交互的子课题。从应用的视角，I1-OntoSpace 课题组认为空间本体本质上具有三个方面的功能：（1）用于空间定性推理，这是本体表达的基本用途；（2）用于不同空间演算的互操作，这是在 GIS 数据互操作已通过业界规范日臻完善的基础上的进一步发展；（3）自然语言中空间描述（Spatial Expressions）的本体表达[168]。不同于侧重前两点的研究，I1-OntoSpace 将与空间认知密切相关的空间本体研究和自然语言紧密联系在一起，提出了由“通用本体-空间本体-语言本体-交互本体”四步组成的地理空间知识库构建方法。前文的 GUM-Space 即是其语言本体阶段（D3）采用的面向自然语言空间表达的语言学本体，而空间本体阶段（D2）则借鉴了面向认知的空间本体 DOLCE（a Descriptive Ontology for Linguistic and Cognitive Engineering）[169]。

地理知识库的构建需文本驱动，还来源于数据与知识工程目前面临的三个挑战[170]：（1）统计方法与逻辑方法的融合；（2）从小规模、面向局部应用领域转向大规模复杂系统；（3）自然科学及工程应用与人文学科的融合。具体针对地理知识库而言，目前的特征是：多借助于专家构建逻辑结构、小规模面向特定领域、重数理的时空模型而轻地学的文本描述。而文本作为广义形式化的信息载体，其可为地理知识库的构建提供更为宽泛的信息来源。首先，针对领域文本，通过自然语言处理技术，可自动半自动的获取地理知识结构；其次，在网络的尺度下，可使用统计的方法构建大规模地理知识库；最后的目标则是达到时空模型的可计算性和文本的可描述性间的平衡。

空间本体的语言特性和地理知识工程的发展都要求由文本驱动来进行地理知识库的构建。这部分研究可归属于文本本体学习（Ontology Learning from Text）的研究范畴。而在文本本体学习提出的学习层次中，恰好也以地理领域为例（图 6.1）[170]。

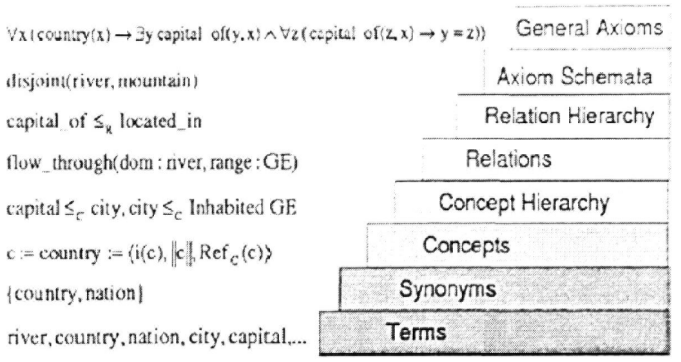


图 6.1 本体学习层次（Ontology learning layer cake）

其中，地理命名实体的抽取可看作是地理概念（Concepts）的类实例（Class Individe）的获取，而空间方位关系的抽取可看作是空间关系（Relations）的关系实例（Relation Individe）的获取。一般认为，文本本体学习的层次是与自然语言处理的层次一致的，如术语（Terms）学习在词汇层面，而公理（Axiom）学习在篇章层面。由于英文与德文自然语言处理技术的相对成熟，仅在词汇层面，即能从术语到同义词（Synonyms），进而获取概念（Concepts）和概念层次（Concept Hierarchy）。而针对行业规范描述等受限中文文本，即使是采用相对简单的自然语言处理技术，也能完成公理和规则学习任务[171]。

6.1.2 应用示例

在自然语言处理项目中使用 Ontology 作为标注集，组织和管理格式化信息是当前的一大趋势。在语义网（Semantic Web）的应用架构下，通过为非结构化的网络文本提供本体语义的支撑，极大的提升了信息的共享与检索性能。GATE 将 JAPE 模板引擎扩展为 Ontology-Awared JAPE Transducer，在 RHS 中使用集成的 OWLIM 实现对 Ontology 的组织和管理。由于抽取的空间方位关系实例可看作是一个三元组（LINK/RLINK，Source，Target），则在地理命名实体类与空间方位关系属性类的约束下，每个空间方位关系实例可直接映射为形如 Class Individe-Relation Individe-Class Individe 的本体形式。而 OAT 则实现了标注的这种本体形式与语料文本间的关联，图 6.2 为 OAT 的编辑界面。

图 6.2 GATE Viewing-Ontology-Based Annotation

在本体技术与平台的支撑下，抽取获得的空间方位关系可以充分利用 Ontology 在信息融合、数据共享、逻辑推理上的优势。通过 Protégé平台，由多个语料文档获取的空间方位关系集合得以融合，并可通过外部插件进行数据的可视化（图 6.3）。

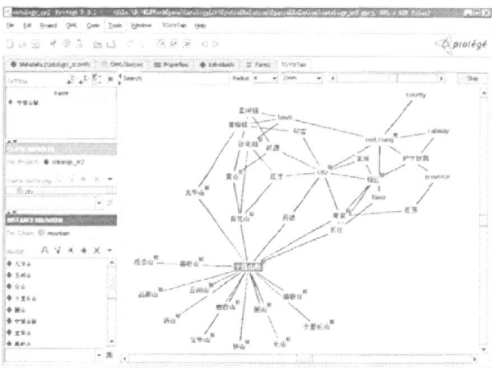


图 6.3 空间方位关系 OWL 可视化

6.2 应用二：空间方位关系的场景重建

严格意义上，空间方位关系的场景重建涉及空间词汇的 GIS 语义映射，可认为是空间意象图式的显性化过程。本文将空间谓词语义映射这类解析阶段的研究内容进行简化，仅针对实验语料中特定的 2D 场景重建用例进行展示。示例包括：(1) 针对句子级的静态地物路径描述，通过空间方位关系抽取构建节点模型，进行路径重建；(2) 针对篇章级的模糊地物描述，首先对各句子级的边界描述进行边界重建，进而通过对边界的空间运算进行地物的图形重建。

6.2.1 应用示例一：地物的路径重建

实验语料中地物路径描述的例句如下：

[无定河]发源于[陕西省定边县白于山]北坡，上源称[红柳河]，东北流至[内蒙古乌审旗巴图湾]，折向东流复入[陕西]境，经[横山县]至[榆林县]的[鱼河堡]折向东南，至[清涧县]入[黄河]。

表 6.1 为根据“无定河”路径描述抽取的空间方位关系，已经过人工修正。其中仅包括 Target/Source 其一为“无定河”的关系，并按文本位移顺序进行排序。其中“无定河”与“黄河”在 EC 关系的基础上进行 Intersect 操作，进而与“清涧县”的 IN 关系，进行归一后取与“清涧县”的 PO 关系。

表 6.1“无定河”例句的空间方位关系抽取

文本位移顺序	Source/Target	LINK	RLINK
1	S 陕西省定边县白于山	PO	N
2	T 红柳河	IN	
3	T 内蒙古乌审旗巴图湾	EC	NE
4	T 陕西	PO	E
5	T 横山县	IN	
6	T 鱼河堡	EC	
7	T 清涧县	PO	

静态地物的路径描述与导航路径的描述类似，都是由按顺序排列的一系列地物与动作组成。其中，动作分为持续性动作与非持续性动作两类，后者又分为状态转移动作和并发动作[57]。地物的路径重建较导航路径而言，由于尺度较大，对描述使用的空间词汇的图形映射相对简单。将与“无定河”存在空间方位关系的地理命名实体转化为路径节点，并按其出现的文本位移顺序定义序号。当地图上这些节点都以确定的点表达时，将尺度与描述对象相当或较大的去除，按序号使用贝塞尔曲线连接即可（图 6.4）。其中“陕西”的点信息未使用。而当节点以线/面的表达时，需根据空间方位关系的信息转换为点表达。例如，（1）、（3）的情况可将 LINK/RLINK 使用细节方向关系表达模型处理；（2）的情况可将“红柳河”的线纳入路径中；（4）、（5）、（6）的情况则取面的中心。

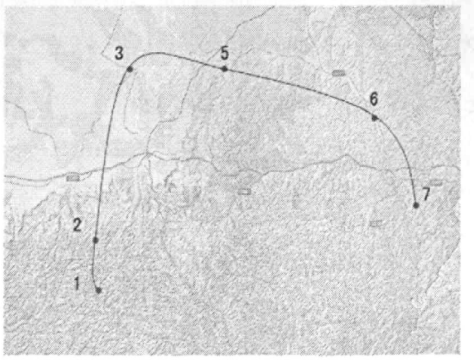


图 6.4“无定河”图形重建结果

6.2.2 应用示例二：模糊地物的图形重建

以下为《中国大百科全书（地理分册）》语料中宁镇山脉的词条解释，由于抽取方法未处理指代消解，实际语料给每句添加了省略的主语“宁镇山脉”：

宁镇山脉　南京、镇江间低山丘陵的总称，略呈东西向向北突出的弧形山脉，耸峙于长江南岸。西起江宁县淳化镇青龙山，经句容县、丹徒县、镇江市、丹阳县境，东止武进县孟河镇黄山。北侧沿江山岭有幕府山、栖霞山、龙潭擂鼓台、五洲山、圌山等，其中以栖霞山为南京市郊著名旅游地。排列于中间的山岭有钟山、宝华山、十里长山、黄山等，其中南京市东郊钟山最高，海拔 448 米；宝华山有保存完好的北亚热带植被，已辟为自然保护区。南侧山岭有青龙山、汤山、仑山、观音山、高骊山等。汤山山麓有汤山温泉，是著名疗养地。

表 6.2 为根据“宁镇山脉”词条解释抽取的空间方位关系，已经过人工修正。其中仅包括 Target/Source 其一为“宁镇山脉”的关系，如“栖霞山”与“南京”的 EC 关系不包括在内，本体可视化即为图 6.3。

表 6.2“宁镇山脉”词条解释的空间方位关系抽取

（A）[宁镇山脉]是[南京]、[镇江]间低山丘陵的总称，略呈东西向向北突出的弧形山脉，耸峙于[长江]南岸。								
S 南京	EC		S 镇江	EC		S 长江	EC	S
（B）[宁镇山脉]西起[江宁县淳化镇青龙山]，经[句容县]、[丹徒县]、[镇江市]、[丹阳县]境，东止[武进县孟河镇黄山]。								
T 江宁县淳化镇青龙山	EC	W	S 句容县	PO		S 丹徒县	PO	
T 武进县孟河镇黄山	EC	E	S 丹阳县	PO		S 镇江市	PO	
（C）[宁镇山脉]北侧沿江山岭有[幕府山]、[栖霞山]、[龙潭擂鼓台]、[五洲山]、[圌山]等，其中以[栖霞山]为[南京市]郊著名旅游地。								
T 幕府山	EC	N	T 栖霞山	EC	N	T 龙潭擂鼓台	EC	N
T 五洲山	EC	N	T 圌山	EC	N			
（D）[宁镇山脉]排列于中间的山岭有[钟山]、[宝华山]、[十里长山]、[黄山]等，其中[南京市]东郊[钟山]最高，海拔 448 米；[宝华山]有保存完好的北亚热带植被，已辟为自然保护区。								
T 钟山	IN		T 宝华山	IN		T 十里长山	IN	
T 黄山	IN							
（E）[宁镇山脉]南侧山岭有[青龙山]、[汤山]、[仑山]、[观音山]、[高骊山]等。								
T 青龙山	EC	S	T 汤山	EC	S	T 仑山	EC	S
T 观音山	EC	S						

首先对每个句子中的边界描述进行边界重建（图 6.5）。为简化，将与“宁镇山脉”存在空间方位关系的地理命名实体的图形表达默认为线或线上的点。与上例中路径重建不同，边界重建获得的图形不是线，而是由线及其一侧或两侧缓冲形成面。例如，对于句子（A）、（C）、（E）的情况，根据 LINK 类型 EC 与一致的 RLINK 类型，边界重建形成由相关的地理命名实体构成的线及其一侧形成

的面。不同于路径重建中要求严格按照描述顺序，在边界重建由点构线的过程中，其顺序可将地物点纵向或横向投影后确定。而在由线构面的过程中，需根据 Source/Target 的不同确定缓冲方向，如（C）中“北侧”为向南缓冲，而（E）中“南侧”为向北缓冲。对于句子（A）、（D）的情况，由相关地理命名实体构成的线严格上不是边界，而是被模糊地物所包含的部分，因此两侧作缓冲（图中未绘出）。在边界重建过程中，缓冲区的宽度取值一般取边界长度。

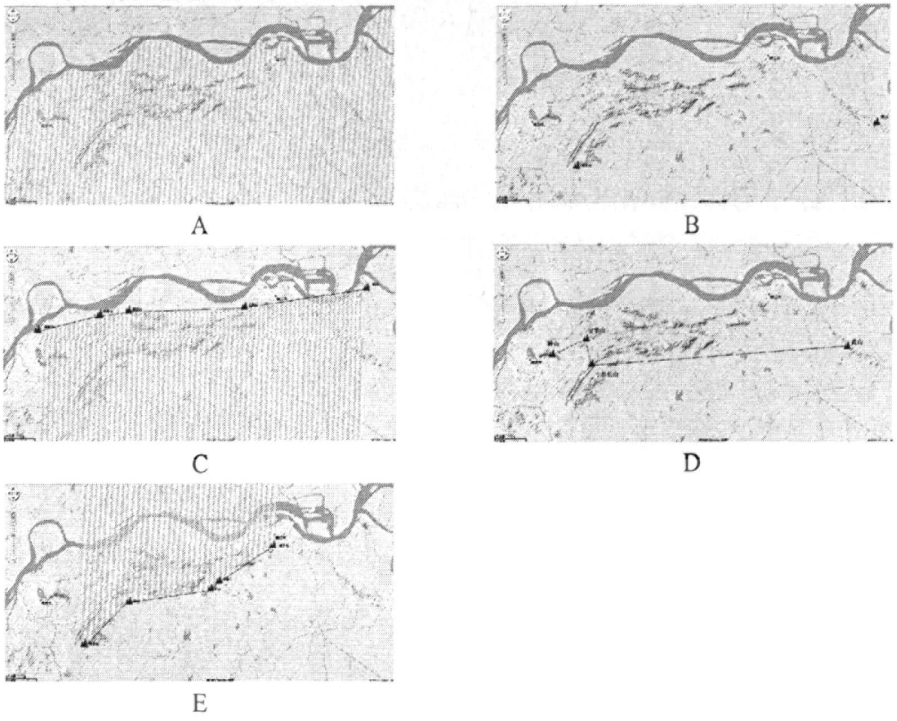


图 6.5“宁镇山脉”图形重建中间过程

在对每个句子进行边界构建的基础上，可简单的将缓冲面进行求交集获取模糊地物的图形重建。而实际处理中，情况较为复杂。如图 6.6，在“黄山”部分的图形处理上，需根据置信度高的（C）、（D）交集获得的图形进行调整。

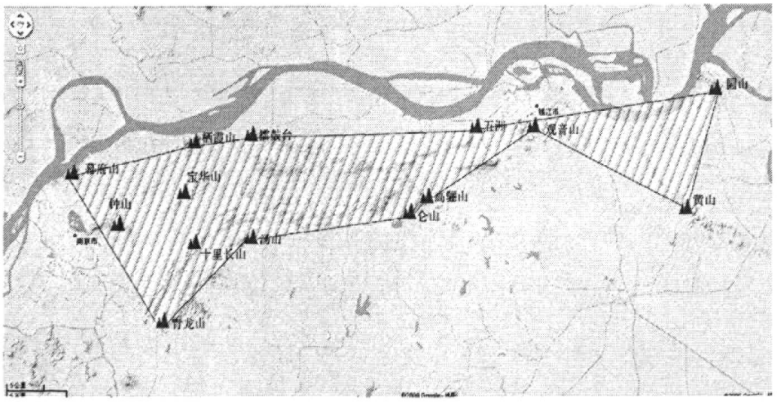


图 6.6“宁镇山脉”图形重建结果

6.3 本章小结

本章简要分析了空间方位关系抽取在地理知识库构建和场景重建中的应用模式，通过文本驱动的地理知识库构建以及空间方位关系的图形重建演示了相关的应用场景，验证了本文空间方位关系表达以及抽取方法的有效性。

第7章 结论与展望

7.1 研究结论

本文的研究是国家“863 计划”课题“基于中文文本的 GIS 空间操作挖掘关键技术研究”（2007AA12Z221）和国家自然基金项目“面向 GIS 的文本空间关系解析机制研究”（40971231）的组成部分，作为自然语言空间关系“描述-识别-解析”整体框架中“描述-识别”阶段的内容。对整个研究框架而言，自动识别过程起着承上启下的作用。

本文研究分为四个部分：（1）面向信息抽取的中文文本空间方位关系表达；（2）中文文本空间方位关系的语料标注与分析；（3）基于规则的中文文本空间方位关系抽取方法；（4）基于支持向量机的中文文本空间方位关系抽取方法。主要研究结论与成果如下：

1）提出了中文文本空间方位关系表达的两个层次：A）从自然语言到语义量化公式的映射，即解析；B）从自然语言到形式语言的映射，即识别。其中，受语境约束的术语在这两个空间方位关系表达研究层次中起着纽带的作用。

2）为处理图形语义与文本语义间的术语歧义，根据原型理论，以 RCC8 定义的类型图形作为原型，借鉴定性空间表达中的“概念邻近图”思想，对由文本描述触发的拓扑关系意象图式与 RCC8 类型术语进行映射，得以有效指导文本空间方位关系的分类判断。

3）在对地理辞典类语料的空间方位关系标注过程中，针对领域特征进行了处理：A）扩展了地理命名实体组合情况的标注，并设计了归一规则。B）扩展了地理空间描述中拓扑关系判断的规范。

4）通过构建空间词汇词典，结合关系模板及匹配规则，研究基于规则的空间方位关系抽取方法。实验表明，针对在地名辞典类语料，使用不同词性类型的空间词汇同时作为种子词，选择丰富的特征向量，通过 BootStrapping 迭代方法获取的新增空间词汇的准确率最高；使用词汇语义型词典中概念相似度、相关度计算，对空间词汇语义参数的设置具有一定的效果，但由于其通用性，还需较多的人工修正；针对特定的空间方位关系描述文本，通过半自动的构建空间词汇知识库以及适量的文法规则，可取得较好的抽取效果。

5）选用关系抽取中性能最佳的支持向量机进行基于机器学习的空间方位关系抽取方法研究。实验表明，基于支持向量机的抽取方法具有较好的发现性能，且不依赖于空间词汇的识别，具有一定的实用意义。

针对目前 GIS 自然语言空间关系研究对象局限在选定的空间词汇以及有限文法的空间关系描述，本文通过引入了中文信息处理领域的研究范式及技术，对“描述-识别”阶段中文文本空间方位关系的抽取方法进行了系统的研究。主要创新有以下两个方面：

1）面向信息抽取的空间方位关系表达。首先，通过文本与地图两类符号系统在空间关系表达上的对比，提出了中文文本空间方位关系表达的两个层次，以及受语境约束的术语在空间方位关系表达研究中的纽带作用。然后，从空间方位关系表达式、分类、类型术语判断三个角度分析了面向信息抽取的空间方位关系表达。提出基于意象图式的术语判断方法，即能有效的指导对文本空间方位关系的分类判断，也为抽取后 GIS 空间关系的解析提供基础。

2）将中文信息处理领域的研究范式与成熟技术引入 GIS 自然语言空间关系的研究。首先，基于标注集与标注规范进行空间方位关系语料库的构建与分析；其次，采用自然语言处理中“理性主义”的研究范式，研究基于规则的空间方位关系抽取方法；最后，采用自然语言处理中“经验主义”的研究范式，研究基于机器学习的空间方位关系抽取方法。

7.2 进一步工作

本文作为“描述-识别”阶段的研究，下一步的工作即是“识别-解析”阶段的研究，以及空间方位关系抽取的应用研究，这里不再展开。仅就本文研究而言，还有诸多问题有待深入：

1）如同知网中概念网络的构建一样，本文在空间方位关系表达，尤其是拓扑类型术语判断规范扩展上，有较多自我心证的成分。虽在语料标注过程中得以证实其有效性，但对于空间关系图形认知与文本认知深层次的问题还需要进一步的研究。尤其需要更多的参考认知心理学、认知语言学、符号学领域的成果。

2）地理空间信息语料库建设才刚刚起步，标注集的设计与标注规范的制定需对地名命名实体、空间方位关系、空间过程、空间属性等进行统一的考量。而标注工程本身在技术支撑，人员组织，语料管理等方面也是巨大的挑战。

3）中文文本空间方位关系抽取作为领域应用，如何有效的融合中文信息处理现有的成熟技术，诸如核方法的引入等，有待进一步的深入。

参考文献

[1] Goodchild M F. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web2.0[J]. International Journal of Spatial Data Infrastructures Research, 2007, 2: 24-32.

[2] 崔希亮. 空间关系的类型学研究[J]. 汉语学习, 2002, (1): 1-8.

[3] 方经民. 汉语空间方位参照的认知结构[J]. 世界汉语教学, 1999, 50(4): 32-38.

[4] 张金桥. 汉语空间关系复杂句心理表征项目互换效应[J]. 暨南大学华文学院学报, 2004, (04): 44-49.

[5] 刘宁生. 汉语怎样表达物体的空间关系[J]. 中国语文, 1994, (3): 169-179.

[6] 张克定. 空间关系及其语言表达的认知语言学阐释[J]. 河南大学学报(社会科学版), 2008, 48(1): 1-8.

[7] Lloyd R. Spatial Cognition-Geographic Environments [M]. Kluwer Academic Publishers, 1997.

[8] 蓝纯. 认知语言学与隐喻研究[M]. 外语教学与研究出版社, 2005.

[9] Levinson S C. Language and Space[J]. 1996, 25: 353-382.

[10] 刘丽虹. 空间词汇运用对空间认知与时间认知的影响[D].博士. 2005.

[11] 文炼. 处所，时间和方位[M]. 上海教育出版社, 1984.

[12] 李静熹. 现代汉语方位词研究[D].博士. 上海师范大学, 2000.

[13] 廖秋忠. 空间方位词和方位参考点[J]. 中国语文, 1989, 1: 9-16.

[14] 朴珉秀. 现代汉语方位词“前、后、上、下”研究[D].博士. 复旦大学, 2005.

[15] 方经民. 现代汉语方位参照聚合类型[J]. 语言研究, 1987, (4): 32-38.

[16] 郭锐. 方位词“前”、“后”、“左”、“右”的参照策略[J]. 中国语言学论丛, 2004, (3): 1-30.

[17] 陶氏河宁. 现代汉语方位词“东、西、南、北”的语义分析[J]. 云南师范大学学报, 2006, 4(5): 44-48.

[18] 王家耀. 空间信息系统原理[M]. 北京: 科学出版社, 2001.

[19] 郭仁忠. 空间分析（第二版）[M]. 高等教育出版社, 2001. 192-201.

[20] Egenhofer M. Pre-Processing Queries with Sptial Constraints[J]. PE&RS, 1994, 60(6): 783-970.

[21] Mark D. Spatial Representation: A Cognitive View[A]. Geographical information systems: principles and applications. 2nd[M]. London: Longman Scientific, 1999. 81-89.

[22] 杜世宏,秦其明,王桥. 空间关系及其应用[J]. 地学前沿(中国地质大学), 2006, 13(3): 69-80.

[23] 闫浩文. 空间方向关系理论研究[M]. 成都地图出版社: 成都, 2003.

[24] Egenhofer M J, Franzosa R D. Point-Set Topological Spatial Relations[J]. International Journal of Geographical Information System, 1991, 5(2): 161-174.

[25] Egenhofer M J. A Formal Design of Binary Topological Relationships[Z]. New York: Springer-Verlag, 1989.

[26] Egenhofer M J. Reasoning About Binary Topological Relations[Z]. New York: Springer-Verlag, 1991.

[27] Egenhofer M J. Categorizing Binary Topological Relationships Between Regions, Lines, and Points in Geographic Database[R]. Technical Report, Department of Surveying Engineer, University of Maine, 1991.

[28] 陈军. Voronoi动态空间数据模型[M]. 北京: 测绘出版社, 2002.

[29] Clementini E, Felice P D. Approximate Topological Relations[J]. International Journal of Approximate reasoning, 1997, 16(20): 172-204.

[30] Clarke B L. Individuals and Points[J]. Notre Dame Journal of Formal Logic, 1985, 26: 61-67.

[31] Clarke B L. A Calculus of Individuals Based On "Connection"[J]. Notre Dame Journal of Formal Logic, 1981, 22: 204-218.

[32] Cohn A G, Bennett B, Gooday Y J. Qualitative Spatial Representation and Reasoning with the Region Conncetion Calcu-

lus[J]. GeoInformatica, 1997, 1(1): 1-44.

[33] Cohn A G, Gotts N M. The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries[A]. Burrough P A, Frank A U. Geographic Objects with Indeterminate Boundaries [M]. Morgan Kaufmann, 1996. 230-241.

[34] Oriented Point Relation Algebra(Opram, a Relative Orientation Algebra with Adjustable Granularity). http://www.sfbtr8.spatial-cognition.de/project/r3/opra/index.html

[35] Dipole Relation Algebra(Dra). http://www.sfbtr8.spatial-cognition.de/project/r3/QualitativeCalculi/DipoleCalculus/DipoleCalculus.html

[36] Doublecross Calculus.http://www.sfbtr8.spatial-cognition.de/project/r3/QualitativeCalculi/Doublecross/Doublecross.html

[37] Ternary Point Configuration Calculus(Tpcc). http://www.sfbtr8.spatial-cognition.de/project/r3/QualitativeCalculi/TPCC/index.html

[38] 郭庆胜,杜晓初,闫卫阳. 地理空间推理[M]. 北京: 科学出版社, 2006. 114-115.

[39] Goyal R. Similarity Assessment for Cardinal Directions Between Extended Spatial Objects[D].Maine: University of Maine, 2000.

[40] 杜世宏,王桥,杨一鹏. 一种定性细节方向关系的表达模型[J]. 中国图象图形学报, 2004, 9(12).

[41] 杜世宏,王桥,李治江. GIS中自然语言空间关系定义[J]. 武汉大学学报(信息科学版), 2005, 30(6): 533-538.

[42] 廖楚江,杜清运. GIS空间关系描述模型研究综述[J]. 2004, 29(004): 79-82.

[43] Hong J. Qualitative Distance and Direction Reasoning in Geographic Space[D].Ph.D. University of Maine, 1994.

[44] Papadias D, Delis V. Relations-Based Similarity[Z]. Lasvegas: ACM Press, 1997.

[45] Bruns T, Egenhofer M. Similarity of Spatial Scenes[C]. 1996. 31-42.

[46] Mark D M, Frank A U. Ncgia Initiative2 "Languages of Spatial Relations" Closing Report[R]. NCGIA, 1992.

[47] Mark D M, Egenhofer M J. Calibrating the Meanings of Spatial Predicates From Natural Language: Line-Region Relations[J]. 1994.

[48] Egenhofer M J, Shariff A. Metric Details for Natural-Language Spatial Relations[J]. 1998, 16(4): 295-321.

[49] Shariff A, Egenhofer M J, Mark D M. Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms[J]. 1998, 12(3): 215-245.

[50] Nedas K A, Egenhofer M J. Metric Details of Topological Line-Line Relations[J]. International Journal of Geographical Information Systems, 2007, 21(1): 21-48.

[51] Xu J. Formalizing Natural-Language Spatial Relations Between Linear Objects with Topological and Metric Properties[J]. 2008, 21(4): 377-395.

[52] Mark D M, Egenhofer M J. Topology of Prototypical Spatial Relations Between Lines and Regions in English and Spanish[J]. 1995, : 245-254.

[53] Mark D M, Comas D, Egenhofer M J, et al. Evaluating and Refining Computational Models of Spatial Relations through Cross-Linguistic Human-Subjects Testing[J]. 1995.

[54] 许珺,张晶,司望利,等. 线状物体空间关系的自然语言理解的双语比较[J]. 遥感学报, 2008, 12(2): 362-369.

[55] 马林兵. 基于受限自然语言的空间信息查询研究与应用[D].博士. 武汉: 武汉大学, 2003. 155.

[56] 马林兵,曹小曙. 空间关系的动态性和模糊性描述[J]. 地理与地理信息学科学, 2006, 22(6): 1-4.

[57] 刘瑜,高勇,林报嘉,等. 基于受限汉语的GIS路径重建研究[J]. 遥感学报, 2004, 8(4): 323-330.

[58] 自然空间查询语言解译机制研究[D].博士. 解放军信息工程大学, 2009. 131.

[59] 邓敏,李成名,刘文宝. 利用拓扑和度量相结合的方法描述面目标间的空间关系[J]. 测绘学报, 2002, 32(2): 164-169.

[60] 陈学工,张弛伟,张文艺,等. 度量参数与空间关系描述的研究[J]. 计算机技术与发展, 2007, 17(5): 187-190.

[61] 张连蓬,刘国林,江涛,等. 受限自然语言查询在GIS中的应用[J]. 测绘学院报, 2002, 19(4): 283-286.

[62] 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.

[63] 吴静,蔡砥,王铮. 地理信息系统中自然语言查询的分词处理与应用[J]. 地球信息科学, 2005, 7(3): 67-71.

[64] 付胜博. 基于自然语言的空间数据检索研究[D].硕士. 西安: 西北工业大学, 2007.

[65] 徐爱萍,边馥苓. 基于语义查询树的GIS中文查询语句向SQL的转换[J]. 武汉大学学报（信息科学版）, 2006, 31(10): 924-927.

[66] Gaizauskas R, Wilks Y. Information Extraction: Beyond Document Retrieval[J]. Journal of Documentation, 1997.

[67] Ace.http://www.nist.gov/speech/tests/ace/index.htm

[68] Cardie C. Empirical Methods in Information Extraction[J]. AI Magazine, 1997, 18(4): 65-78.

[69] Appelt D D. Introduction to Information Extraction[J]. AI Commun, 1999, 12(3): 161-172.

[70] Technology N I O S. In Proceedings of the 6Th Message Understanding Conference(Muc-7)[C]. 1998.

[71] Ace (Automatic Content Extraction) English Annotation Guidelines for Relations[S].

[72] 邓擘,樊孝忠,杨立公. 用语义模式提取实体关系的方法[J]. 计算机工程, 2007, 33(10): 212-214.

[73] Riloff E, Jones R. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping[Z]. 1999.

[74] 姜吉发,王树西. 一种自举的二元关系和二元关系模式获取方法[J]. 中文信息学报, 2005, 19(2): 71-77.

[75] 张素香,李蕾,秦颖,等. 基于Boot Strapping的中文实体关系自动生成[J]. 微电子学与计算机, 2006, 23(12): 15-18.

[76] 陈晓颖,胡熠,陆汝占. 实体关系模板的获取技术[J]. 计算机工程, 2007, 33(21): 199-201.

[77] 陈文亮,朱靖波,姚天顺,等. 基于Bootstrapping的领域词汇自动获取[Z]. 2003.

[78] 车万翔,刘挺,李生. 实体关系自动抽取[Z]. 2005.

[79] Guodong Z, Jian S, Jie Z, et al. Exploring Various Knowledge in Relation Extraction[Z]. Ann Arbor: 2005.

[80] 徐芬,王挺,陈火旺. 基于SVM方法的中文实体关系抽取[J].

[81] 刘路,李弼程,张先飞. 基于正反例训练的SVM命名实体关系抽取[J]. 计算机应用, 2008, 28(6): 1444-1446.

[82] 董静,孙乐,冯元勇,等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-85.

[83] Haussler D. Convolution Kernels On Discrete Structures[D].Califomia: University of Califomia, 1999. 7-10.

[84] Bunescu R C, Mooney R J. Subsequence Kernels for Relation Extraction[J].

[85] Collins M, Duffy N. Convolution Kernels for Natural Language[J]. 2001.

[86] 刘克彬. 基于核函数的命名实体关系抽取技术研究[D].硕士. 上海: 上海交通大学, 2006.

[87] 潘坤. 基于树核函数的命名实体语义关系抽取方法的研究[D].硕士. 苏州: 苏州大学, 2008. 67.

[88] 黄瑞红,孙乐,冯元勇. 基于核方法的中文实体关系抽取研究[J]. 中文信息学报, 2008, 22(5): 102-108.

[89] 周俊生,戴新宇,陈家俊,等. 基于一种新的合成核的中文实体关系自动抽取[J]. 2009.

[90] Coyne B, Sproat R. Wordseye an Automatic Text-to-Scene Conversion System[Z]. Los Angeles, CA USA: 2001.

[91] Reinbergerr M. Automatic Extraction of Spatial Relations[J]. 2005.

[92] 陆汝钤,张松懋. 从故事到动画片——全过程计算机辅助动画自动生成[J]. 自动化学报, 2002, 28(03): 321-348.

[93] 李晗静. 基于自然语言处理的空间概念建模研究[D].博士. 哈尔滨工业大学, 2007. 125.

[94] 李晗静,李生,赵铁军. 汉语中方位参考点恢复研究[J]. 计算机研究与发展, 2007, (02).

[95] 赵纪元,李晗静,赵铁军. 汉语空间关系中射体识别问题的研究与分析[C]. 中国辽宁沈阳: 2006. 1-6.

[96] 乐小虬,杨崇俊,于文洋. 基于空间语义角色的自然语言空间概念提取[J]. 武汉大学学报 信息科学版, 2005, 30(12).

[97] 乐小虬,杨崇俊. 非受限文本中深层空间语义的识别方法[J]. 计算机工程, 2006, 32(4): 36-38.

[98] 刘元凤. 基于文本的地理空间数据挖掘与可视化[J]. 中国科技论文在线, 2008.

[99] 张雪英,闾国年. 自然语言空间关系及其在GIS中的应用研究[J]. 地球信息科学, 2007, 9(6): 77-81.

[100] 刘瑜,龚咏喜,张晶,等. 地理空间中的空间关系表达和推理[J]. 地理与地理信息科学, 2007, 23(5): 1-7.

[101] Herskovits A. Language Spatial Cognition and Vision[A]. Oliver S. Spatial and Temporal Reasoning[M]. Kluwer academic publisher, 1997.

[102] Lakoff G. Women,Fire,and Dangerous Things[M]. Chicago: University of Chicago Press, 1987.

[103] Opengis Abstract Specification, Version 4[S].

[104] 崔巍. 用本体实现地理信息系统语义集成和互操作[D].博士. 武汉大学, 2004. 133.

[105] 徐健,张智雄,吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术, 2008, (8): 18-22.

[106] Metadata for the Adl Feature Type Thesaurus[S].

[107] Getty Thesaurus of Geographic Names Online: Place Type Lookup.
http://www.getty.edu/research/conducting_research/vocabularies/tgn/
[108] Leidner J L. Toponym Resolution: A First Large-Scale Comparative Evaluation[J]. 2006.
[109] Hill. Georeferencing[M].
[110] 唐旭日. 基于汉语文本的中国地名解析系统[J]. 2009, : 12.
[111] Tezuka T, Tanaka K. Landmark Extraction: A Web Mining Approach[Z]. Springer-Verlag, 2005.
[112] Talmy L. Figure and Ground in Complex Sentences[C]. 1978.
[113] Talmy L. How Language Structures Space[Z]. New York: 1983.
[114] 方经民. 论汉语空间区域范畴的性质和类型[J]. 世界汉语教学, 2002, (3): 37-48.
[115] Clausner T C, Croft W. Domains and Image Schemas[J]. Congnitive Linguistics, 1999.
[116] Levinson S C. Space in Lanugage and Cognition: Explorations in Cognitive Diversity[M]. Cambridge: Cambridge University Press, 2003.
[117] Tarquini F, Clementini E. A User Model for Spatial Relations[Z]. F.Tarquini, E.Clementini, 2007.
[118] Barbara K K. The Linguistic, Cognitive and Cultural Variables of Conceptualization of Space[J]. The Construal of Space in Language and Thought., 1996.
[119] Mark T. Using Siscourse Focus, Temporal Focus, and Spatial Focus to Generate Multisentential Text[Z]. Pittsburgh,PA: 1990.
[120] Denis M. La Descrip Tion D' Itineraires: Des Reperes Pourdes Actions (the Descrip Tion of Routes: Landmarks for Actions)[C]. Orsay, France: 1994. 14.
[121] 邬伦,王晓明,高勇. 基于地理认知的GIS数据元模型研究[J]. 遥感学报, 2005, 9(5): 583-588.
[122] Boden M A. The Philosophy of Artificial Intelligence[Z]. 2006.
[123] Sowa J. Ontology, Metadata and Semiotics[A]. Ganter B, Mineau G. Conceptual Structures: Logic, Linguistic and Computational Issues[M]. Springer, 2000. 55-81.
[124] 杜清运. 空间信息的语言学特征及其自动理解机制研究[D].博士. 武汉: 武汉大学, 2001.
[125] Bennett B. What is a Forest? On the Vagueness of Certain Geographic Concepts[J]. Topoi, 2001, 20(2): 189-201.
[126] Zhang H, Liu Q, Zhang K, et al. Statistical Chinese Parser Ictprop[J].
[127] Ruppenhofer J, Ellsworth M, Petruck M R L, et al. Framenetii:Extended Theory and Practice[EB/OL].
[128] Palmer M, Gildea D, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles[J].
[129] Gum.http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/gum/index.htm
[130] Blaylock N, Swain B, Allen J. Tesla:a Tool for Annotating Geospatial Language Corpora.[Z]. Boulder, Coloradl: 2009.
[131] Spatialml: Annotation Scheme for Marking Spatial Expressions in Natural Language (Version 3.0)[J]. 2009.
[132] Mani I, Anderson D, Hitzeman J. A Framework for Inferring Spatial Locations and Relationships From Text[J]. 2007.
[133] Gate. Developing Language Processing Compometns with Gate5[M]. 2009.
[134] Gamallo P, Agustini A, Lopes G P. Selection Restrictions Acquistion From Corpora[Z]. Springer-Verlag: 2001.
[135] Pustejovsky J. The Generative Lexicon[M]. MIT Press, 1995.
[136] Gamallo P, Agustini A, Lopes G P. Using Co-Composition for Acquiring Syntactic and Semantic Subcategorisation[Z]. 2002.
[137] Pinker S. Language Learnability and Language Development[M]. Harvard University Press, 1984.
[138] Yangarber R, Grishman R, Tapanainen P, et al. Automatic Acquisition 0F Domain Knowledge for Information Extraction[Z]. Saarbriicken, Germany: 2000.
[139] Weigang L, Ting L, Sheng L. Bootstrapping for Extracting Relations Form Large Corpora[J]. 2008.
[140] 张素香,李蕾,谭咏梅. 特定领域下关系模板的研究[J]. 北京邮电大学学报, 2006, 29(5): 79-83.
[141] 杨沐昀,刘桐菊,赵铁军. 基于TFIDF的专业词汇获取[C]. 2003.
[142] 贾君枝,董刚. FrameNet_WordNet_VerbNet比较研究[J]. 情报科学, 2007, 25(11): 1682-1686.

[143] Vernet: A Broad-Coverage Comperhensive Verb Lexicon.http://repository.upenn.edu/dissertations/AAI3179808

[144] Introduction to Wordnet: An On-Line Lexical Database.http://courses.media.mit.edu/2002fall/mas962/MAS962/miller.pdf

[145] Xue N, Xia F, Chiou F, et al. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[146] 郝晓燕,刘伟,李茹,等. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报, 2007, 21(5): 96-138.

[147] 知网.http://www.keenage.com

[148] 梅家驹. 同义词词林[M]. 上海: 上海辞书出版社, 1996.

[149] 同义词词林扩展版.http://ir.hit.edu.cn/

[150] 景东升. 基于本体的地理空间信息语义表达和服务研究[D].博士. 中国科学院遥感应用研究所, 2005. 133.

[151] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 计算语言学及中文信息处理, 2002.

[152] 李涓子. 汉语语义排歧方法研究[D].博士. 北京: 清华大学, 1999.

[153] 许云,樊孝忠,张锋. 基于知网的语义相关度计算[J]. 北京理工大学学报, 2005, : 411-414.

[154] 周强,冯松岩. 构建知网关系的网状表示[J]. 中文信息学报, 2000.

[155] Weir B S. Genetic Data Analysis Ii-Methods for Discrete Population Genetic Data[M]. Sunderland: Sinauer Associates Inc. Publishes, 1996.

[156] Lucene[DB/CD].http://lucene.apache.org.

[157] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines(and Other Kernel-Based Learning Methods)[D].2000.

[158] T Z. Regularized Winnow Methods[J]. In Advances in Neural Information Processing System 13, 2001, : 703-709.

[159] Vapnik V N. Statistical Learning Theory[M]. New York: Wiley, 1998.

[160] Vapnik V N, Chervonenkis A J. The Necessary and Sufficient Conditions for Consistency of the Method of Empirical Risk Minimization(in Russion)[A]. Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting[M]. Moscow: Nauka, 1989. 207-249.

[161] Vapnik V N. The Nature of Statistical Learning Theory(中文版：张学工译, 统计学习理论的本质,北京清华大学出版社,2000)[M]. Berlin: Springer, 1995.

[162] Allen J F. Maintaining Knowledge About Temporal Intervals[J]. Communications of the Association of Computing Machinery, 1983, 11(26): 832-843.

[163] Chinchor N. Muc-4 Evaluation Metrics[Z]. 1992.

[164] Libsvm: A Library for Support Vector Machines[DB/CD].http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2008.

[165] 黄茂军. 地理本体的形式化表达机制及其在地图服务中的应用研究[D].博士. 武汉: 武汉大学, 2005.

[166] 胡鹤. 本体方法及其时空推理应用研究[D].博士. 吉林: 吉林大学, 2004.

[167] Sfb_Tb8.http://www.sfbtr8.spatial-cognition.de/

[168] Bateman J, Farrar S. Spatial Ontology Baseline (Deliverable D2)[M]. 2006.

[169] Kassel G. Integration of the Dolce Top-Level Ontology Into the Ontospec Methodology[J]. 2005.

[170] Cimiano P. Ontology Learning and Population From Text(Algorithms, Evaluation and Applications)[M]. Springer, 2006.

[171] 薛晓蕾. 自然语言描述的空间规划结构化解析及模型研究[D].硕士. 南京: 南京师范大学, 2009. 87.

攻读博士期间发表的学术论文及科研成果

主要参加项目：

1. 国家自然科学基金项目“面向 GIS 的文本空间关系解析机制研究”（40971231），排名第三；
2. 国家“863 计划”专题课题“基于中文文本的 GIS 空间操作挖掘关键技术研究”（2007AA12Z221），参与；
3. “南京市数字房产”项目，参与。

主要发表论文：

1. 蒋文明,张雪英,李伯秋.基于条件随机场的中文地址要素识别方法.计算机工程与应用, 2010,46(13):129-131.
2. 蒋文明,盛业华,严岩.基于 FDS 的 RIA WebGIS 研究.微计算机信息, 2007,23(31):178-180.
3. 蒋海琴,闾国年,蒋文明.房产管理信息系统.北京:科学出版社,2007.

致谢

这篇博士论文得以顺利完成，首先要感谢我的导师闾国年教授与盛业华教授。两位导师言传身教，以深厚的学术思想，渊博的知识，严谨的治学态度，高屋建瓴的眼光和勇于开拓创新的精神深深的影响着我，令我受益良多，终生难忘。八年来，我的每一点进步和收获，都渗透着导师们的汗水；克服的每一个困难都离不开导师们的支持和鼓励。师恩如海，我谨向闾老师与盛老师表达我深深的敬意和由衷的感谢。

感谢张雪英教授为我提供了研究平台。从论文选题、构思到最后定稿的各个环节张老师都给予细心指引与教导，其敏锐的学术洞察力、精益求精的工作态度以及侮人不倦的师者风范是我学习的楷模。感谢张宏老师在项目管理与技术研发上对我的谆谆教导，使我打下了坚实的基础。

感谢实验室汤国安教授、黄家柱教授、周卫教授、陈锁忠教授、孙毅中教授、龙毅教授、刘学军教授、朱长青教授、李云梅教授、韦玉春教授、沈捷副教授、张书亮副教授、王永君副教授及李安波、周良辰、郭飞、杨林、杨昕、温永宁、吴明光、张卡、刘晓燕、叶春、周洁雨、周安宁、朱岭和贺德刚等诸位老师在学习和生活中给予的诸多支持与帮助。

感谢小组共事的陈洋、乔延春、刘二年、刘爱利、刘剋、丰江帆、汪钟琪、朱瑶、徐希涛、周玉红、倪峰等。感谢朱少楠、申琪君、李伯秋、李玉森、张春菊、陈雨田等在论文相关的语料标注、规则归纳、模块开发上的帮助。感谢南京房产局的蒋海琴主任，丽水城建测量队的厉旭东队长、赵建华、唐秀娟。

此外，感谢我的父亲蒋家良、母亲高梅娣、岳父向宝祥、岳母钱素英。感谢我的夫人向莉莉，女儿蒋智娴。感谢你们对我巨大的支持。

蒋文明

2010年5月 于南师仙林