**文章编号：**

# 基于语义解析的中文 GIS 自然语言接口实现研究*

周俊生[1]， 曲 维光[1]，许菊红[1]，龙毅[2]，朱耀邦[1]

（1.南京师范大学 计算机科学与技术学院，江苏 南京 210023；

2. 南京师范大学地理科学学院，虚拟地理环境教育部重点实验室，江苏 南京 210023）

**摘要：**本文对基于语义解析的中文地理信息系统（GIS）自然语言接口实现技术与方法进行了探索性的研究。首先，我们针对一个具体 GIS 应用领域设计和开发了一种函数式的形式化意义表示语言 GISQL 和一个中文语义解析标注语料库；然后，我们通过引入混合树作为隐变量用于构造输入句子与输出表示结构之间的对应关系，提出了一种基于含隐变量的感知器模型的语义解析算法。在开发的中文语义解析标注语料库上的实验结果显示，本文提出的语义解析算法的 F1 值达到了 90.67%，明显优于 baseline 系统。更重要的是，本文的研究证明了基于语义解析方法实现中文 GIS 的自然语言接口是一种有效可行的途径。

**关键词：**地理信息系统；自然语言接口；语义解析

**中图分类号：**TP391 **文献标识码：**A

# Implementing NLIs to GISs Using Semantic Parsing

Junsheng Zhou[1], Weiguang Qu[1], Juhong Xu[1], Yi Long[2], Yaobang Zhu[1]

(1. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu, 210023, China;

2. Key Laboratory of Virtual Geographic Environment of Ministry of Education, School of Geographic Science，Nanjing Normal University, Nanjing, Jiangsu 210023, China)

Abstract: Natural Language Interfaces (NLIs) to Geographical Information Systems (GISs) have not received a lot of attention in computational linguistics, in spite of the potential values of such systems for users of GISs. This paper presents a pilot study of implementing Chinese NLIs to GISs based on semantic parsing. First, we design a formal meaning representation language (MRL) related to a specific GIS application and develop a corresponding corpus. Second, we translate the natural language questions into GIS queries in MRL using semantic parsing. In particular, we propose a semantic parsing approach based on a latent structural perceptron with hybrid tree. Our evaluation results on the developed corpus show that the proposed methods significantly outperform the baseline approaches, and more importantly, demonstrate that it is feasible to build such NLIs to GISs using semantic parsing.

**Key words:** Geographical Information Systems; Natural Language Interfaces; Semantic Parsing.

## 1 引言

随着地理信息系统（GIS）应用的普及，中文GIS应用越来越面向公众服务，如位置信息服务、车载地图导航及旅游景点介绍等。人们可以通过GIS系统查询一些与日常生活息息相关的信息，比如“107国道穿越哪几个县”、“查询金陵饭店附近500米范围内的超市”等。但如果在传统的基于窗口、菜单和对话框等形式的GIS条件界面上执行这些GIS操作时，经常需要在不同的图层设置条件和输入信息，比较繁琐与低效。因此，如果在GIS中合理运用自然语言接口实现人机间的通信交互，更符合人们的认知习惯和语言习惯，更有助于GIS的应用普及。近些年来，许多研究者在中文GIS的自然语言接口技术上展开了一系列的研究[1-4]，但是目前的研究主要还是基于文法规则或模式匹配的方法。显然，这种基于规则匹配的方法很难解决中文表达的灵活性问题。

* **收稿日期：** **定稿日期：**

**基金项目：**国家自然科学基金资助项目(61073119, 61272221，41171350)；江苏省社科基金资助项目(12YYA002)。

另一方面，近些年来语义解析（semantic parsing）已成为自然语言处理领域的一个研究热点问题。语义解析的目标是将自然语言形式的句子转换成一种完全形式化的意义表示MR（Meaning Representation）[5]。由于意义表示语言MRL（Meaning Representation Language）是一种无歧义的形式化语言，因而，基于一种形式化的MRL给出的自然语言句子的意义表示可以被计算机直接处理和自动推理。在过去的十来年中，研究者们提出了多种基于统计学习模型的语义解析方法。例如，Wong (2006) 提出了一种基于统计机器翻译技术的语义解析算法WASP[6]，Lu (2008) 提出了一种基于生成式模型（generative model）的语义解析方法[7]。Kwiatkowski (2010)则提出了基于组合范畴文法CCG（Combinatory Categorial Grammar）以及高阶合一方法的语义解析方法等[8]。

因此，本文将采用语义解析方法对中文 GIS 自然语言接口实现技术展开探索性的研究。为了能够采用有监督学习的中文语义解析算法实现中文 GIS 自然语言接口，我们首先选择一个 GIS 具体应用领域设计了一种形式化意义表示语言，并开发了一个相应的语义解析标注语料库；然后，我们设计了一种有效的语义解析算法，实现了 GIS 操作的自然语言输入到形式化意义表示形式的转换。在所开发的语料库上进行的十折交叉验证实验结果显示，本文所采用的语义解析算法的 F1 值达到了 90.67%，性能明显优于 baseline 系统。

## 2 形式化意义表示语言的设计及语料库的开发

为了将自然语言的句子在转化成一种计算机可理解和执行的形式化表示，首先需要定义一种形式化的意义表示语言。具体的，我们以南京市地图信息查询作为应用领域，设计了一种函数式的形式化意义表示语言 GISQL；在此基础上，我们进一步开发了一个相应的中文语义解析标注语料库。

### 2.1 形式化意义表示语言 GISQL 的设计

GISQL 是一种函数式的意义表示语言，之所以选择函数式的形式语言表示而没有选择更加普遍使用的 SQL 语言是因为函数式语言能够提供一种更加易于实现映射的组合形式将自然语言句子映射到复杂的意义表示形式。意义表示语言中的基本元素与 GIS 数据库对象的一些术语之间存在一定的对应关系。这些基本元素包含非终结符和函数（或谓词）。在 GIS 数据库中，存在很多的实体类型，例如学校、超市、银行、景点、娱乐场所等，所以对于不同的实体类型定义不同的非终结符是不切实际的。因此我们引入了一个非终结符“ENTITYNAME”代表各种不同类型的实体，包括地名、单位名、街道名、行政区名等。但在每一次引用时它指代的实体是确定和唯一的，例如“夫子庙”、“文苑路”、“玄武区”等地理命名实体。此外，在自然语言的表达中有一些实体名具有不确定性，比如 “苏果超市”、“银行”等并不能代表一个特定位置的超市和银行。为此，我们引入了另外一种非终结符“ENTITYTYPENAME”代表不确定的实体类型。在 GISQL 文法中，我们共设计了 10 种不同的非终结符，如表 1 所示。

基于以上的非终结符集合的设计，我们进一步为 GISQL 文法设计和构造了一个函数（或谓词）集合，共包含 54 个不同函数，表 2 中给出了 GISQL 中的部分函数实例及其相应的意义。GISQL 中的函数和 GIS 系统本身提供的函数并不具有直接的对应关系（本文实验中使用的 GIS 系统是 ArcGis）。简单的说，GISQL 中的单个函数可能涉及到 GIS 中多个函数的嵌套调用。例如，GISQL 中的函数的 contain(Entity, EntityTypeName)函数是首先由 GIS 中的 QueryEntity(List<EntityStruct> targetObject, ISpatialFilter pSpatialFilter, refList<int> featuresID) 查找所有 EntityTypeName 类型的全部实体；其次对 EntityTypeName 中的每一个元素调用 GIS 函数 contain(entity1,entity2)判断是否为真，若为真则保存相应实体，若为假则进行下一次判断，最后返回满足所有条件的实体，从而实现了 GISQL 中的 contain(Entity, EntityTypeName)函数功能。在 GISQL 文法中的函数可以有多种解释，例如一个函数返回值可以是一个实体集合，一个实体的属性，实体之间的某种空间关系，或者是返回一个 GIS

操作的中间结果集。

表 1. GISQL 中的非终结符

| 非终结符 | 相应的意义 |
|---|---|
| QUERY | 开始符号 |
| ENTITY | 实体集合 |
| ENTITYNAME | 特定实体名称 |
| ENTITYTYPENAME | 特定实体类型名称 |
| ORIENTATION | 某一方向 |
| PATH | 路径集合 |
| POSITION | 某一具体实体位置 |
| NUM | 数字集合 |
| BOOLEAN | 二元值：“是”“否” |
| REGION | GIS 函数返回的区域 |

表 2. GISQL 中的一些函数实例描述

| 函数 | 意义 |
|---|---|
| adjacent(Entity, EntityTypeName ) | 与实体 entity 相邻的 entitytypename 类型的实体集合 |
| area_greater(EntityTypeName, Num ) | 面积大于 num 的类型为 entitytypename 的实体集 |
| closest(Entity, EntityTypeName) | 离 Entity 距离最近的 Entitytype 类型的实体 |
| contain(Region, EntityTypeName) | 区域 Region 内包含的 entitytypename 类型的实体集 |
| largest_one(density(Entity)) | 实体集合 Entity 中人口密度最大的一个实体 |
| contain_fewest(EntityTypeName1, EntityTypeName2) | 包含类型 EntityTypeName1 实体数量最少的 EntityTypeName2 类型的实体 |
| path(Entity1, Entity2 ) | 查找实体 Entity1 到实体 Entity2 之间的路径 |
| search(Entity) | 查找指定实体 Entity 的位置 |
| buffer(Entity, Num ) | 返回距离实体 Entity 的 Num 米以内的区域 |
| direction(Entity, Orientation) | 实体 Entity 的某个方位 Orientation 的区域 |

形式文法是由一系列产生式组成的，定义了非终结符和函数集合后，就可定义形式化意义表示语言中的产生式。对于每个非终结符都可以定义一个或多个产生式，而每一个句子意义表示均是由多个产生式组合而成，并且一个特定的产生式组合能确定唯一的 MR 解析树。图 1 给出了一个自然语言查询实例和其相应的意义表示以及对应的 MR 解析树。

(a) 自然语言查询实例： 查询在人口密度最小的行政区内所有苏果超市的面积和是多大?

(b) 形式化意义表示：answer(sum(area(contain(smallest_one(density(queryentity('行政区'))),'苏果超市'))))

(c) 意义表示（MR）解析树：

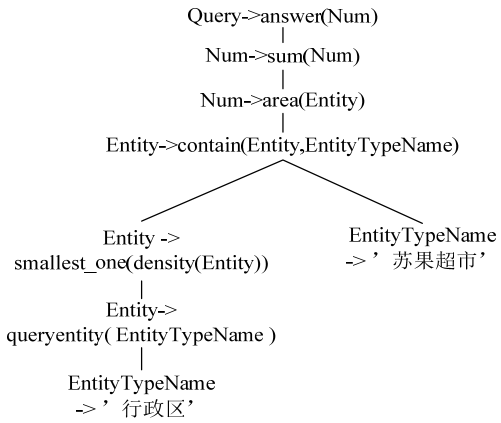



图 1. 一个自然语言的查询实例及其相应的的 MR 解析树

## 2.2 中文语义解析标注语料库的开发

为了建立基于有监督学习的中文语义解析器和实验测试的需要，我们在 GISQL 文法设计的基础上开发了一个中文语义解析标注语料库。为此，我们需要收集大量关于南京市地图查询的中文自然语言查询实例。为了使收集的查询问题实例更接近于人们在实际生活中可能提出的真实查询问题，我们在组织学生收集具体的中文自然语言查询实例之前，首先全面分析和考虑了涉及南京市地图查询的所有可能的问题类型，并设计了一个实际查询问题的类型方案。具体的，我们依据可能的查询目标将所有可能的真实查询问题共分为七种类型，如表 3 所示。其中每种类型下可包含大量不同的查询实例表达，而且一些类似的问题也可以根据不同的句式添加不同的实例表达，例如可以根据人们的表达习惯，将查询动词、语气词、查询目标三者之间的位置互换等。针对这七种查询问题的类型，我们共收集了 1110 条自然语言实例。这些自然语言查询实例表达都是非常常见和灵活的自然语言查询问句，有些比较口语化，如包含词语的缺失、词序的灵活变动等。

表 3. 自然语言查询的问题类型与相应实例

| 类　型 | 实　例 |
|---|---|
| 查询指定实体的地理位置 | 请问 中山陵 坐落 在 什么 位置 ？ |
| 查询满足某种空间条件的实体（或实体集合） | 请 查询 有 哪个 行政区 与 秦淮区 和 鼓楼区 都 相邻 ？ |
| 查询满足某种空间条件的实体的数量 | 请 查询 南师大 东南 方向 的 宾馆 的 个数 ？ |
| 查询指定实体的某种属性值 | 请问 灵谷寺公园 的 占地 面积 有 多 大 ？ |
| 查询满足条件的路径信息 | 请 查找 南京站 到 夫子庙 的 最短 路径 ？ |
| 查询属性值满足某种条件的实体或实体集合 | 请 查询 在 雨花台区 内 有 多少 个 景点 的 面积 比 石头城公园 大 ？ |
| 查询两个或多个实体之间的空间关系 | 南京大学 在 南京师范大学 的 附近 吗 ？ |

对于收集的这 1110 个自然语言实例，我们根据 GISQL 文法对每个实例的意义表示形式都进行人工标注和校对，从而构成了 1110 个自然语言句子/意义表示（NL/MR）对的语料库。其中，自然语言句子的平均长度为 16.38 个字，意义表示的平均长度为 7.72。

在英文语义解析研究中，目前广泛使用的一个实验语料库是 GEOQUERY[9]，它是随 Turbo Prolog 2.0 一起发布的一个小的数据集，共包含 880 条关于美国简单地理信息的自然语言查询实例（例如，“美国最高的山是哪个？”、“有哪些河流经德克萨斯州？”等），并对这些实例采用了一种逻辑查询语言进行了标注。相对于 GEOQUERY，本文研究的实际 GIS 应用领域更复杂，因而设计的 GISQL 文法也更复杂，包含了更多数量的函数和产生式；而

且，我们开发的语料库规模也更大。

## 3 基于隐变量感知器的语义解析实现算法

语义解析任务是将自然语言句子 $x$ 转换成形式化的意义表示 $y$，其中，输入 $x$ 是词的序列，输出 $y$ 是由形式化意义表示文法中的产生式构成的 MR 树。显然，判别式的结构化学习模型非常适合于求解语义解析任务。但是，句子 $x$ 中的词和 MR 树 $y$ 中的结点之间并不存在直接的对应关系。为了解决这个问题，一种有效的方法是通过引入隐变量 $h$ 构造输入句子和输出的意义表示之间的对应性[11]。假设给定输入句子 $x$，输出的 MR 树 $y$ 和隐变量 $h$ 的联合特征向量，用 $F(x,h,y)$ 表示，$w$ 表示一组相对应的参数。判别式结构化预测模型 $f$ 用于返回输出的得分最高的意义表示 $y$，同时最大化隐变量 $h$[10]，如公式（1）所示：

$$f(x) = \arg\max_{h,y}(w \cdot F(x,h,y)) \qquad (1)$$

应用隐变量结构化预测模型解决语义解析问题将面临三个方面的挑战[11]：1）如何引入一个合适的隐变量对输入和输出之间的对应关系进行建模；2）如何设计一个有效的学习算法用于直接优化最大化问题的模型参数 $w$；3）在庞大的树结构搜索空间中，如何设计一个有效的解码算法以获得最优输出。

### 3.1 隐变量的结构

我们引入混合树（hybrid tree）作为隐变量构造输入句子和输出的意义表示树（MR-tree）的对应关系，因为它提供了一个自然的结构表示自然语言句子中的词语和意义表示文法中的产生式的相关性[7]。混合树是由自然语言词语作为叶子结点和文法中的产生式作为内部结点的树。图2中给出了在图1中所示的实例对应的一棵混合树。

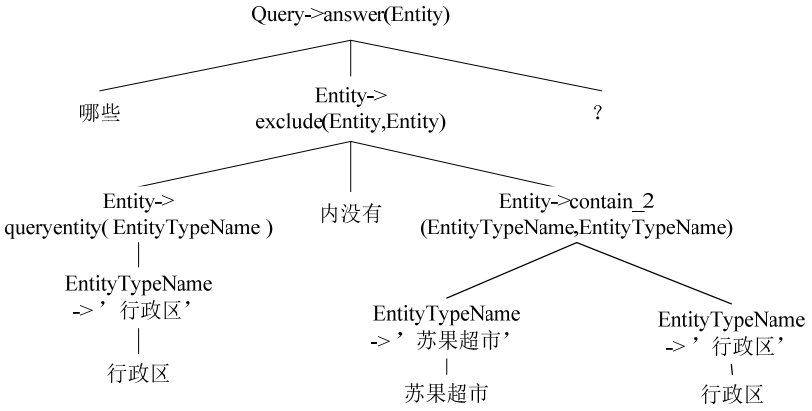



图 2. 混合树实例

对于每一对输入句子$x$和对应的输出MR树$y$，可能存在多个不同的推导能够建立输入输出对（$x$, $y$）之间的对应关系，而其中的每一个推导构成了一棵混合树。对每一棵混合树通过保留其中的产生式中间结点和地理实体终结符可派生出唯一的一棵MR树或一种形式化意义表示。因此，混合树结构非常适合在判别式结构化模型中充当隐变量结构。

### 3.2 参数学习算法

基于效率和收敛性的考虑[12]，我们采用隐变量感知器算法学习判别式模型的语义解析器。类似于结构化感知器[13]，隐变量感知器算法也是一种通过迭代训练集的在线学习算法，图3中描述了语义解析任务中的隐变量感知器算法。此算法主要通过学习预测混合树来帮助解决解析任务，在算法中存在以下两种解码任务：

$$h' = \arg\max_{h,y}(w \cdot F(x_i, y, h)) \quad (2)$$

$$h^* = \arg\max_{h}(w \cdot F(x_i, y_i, h)) \quad (3)$$

其中，$h^*$表示与实例对$(x_i, y_i)$对应的混合树。对训练实例对$(x_i, y_i)$我们可以通过应用一种约束的隐结构解码器来预测混合树$h^*$。约束解码器是指解码搜索过程中仅使用正确解析树$y_i$中的MR产生式作为候选MR产生式集合去搜索得分最高混合树，且此混合树涵盖了句子$x_i$中的所有词语。而混合树$h'$则可以通过一种非约束的普通解码器进行预测，并且从混合树$h'$中可直接提取预测输出$y'$，该操作用运算式*Proj*(*h*)表示。受MIRA在线学习算法的启发[14]，本文采用最大间隔原则更新参数向量$w$。

**Input:** Training data $S = \{(x_i, y_i)\}_{i=1}^{N}$

$w^1 = 0$
for t = 1 to T do
  for i = 1 to N do
    $h' = \arg\max_{h,y}(w \cdot F(x_i, y, h))$
    $y' = Proj(h')$
    if $y' \neq y_i$ then
      $h^* = \arg\max_{h}(w \cdot F(x_i, y_i, h))$
      $w^{t+1}$ *update* $w^t$ *according to* $(x_i, h^*, h')$
    else
      $w^{t+1} = w^t$
  end for
end for

**Output**: parameter vectors *w*

图3. 基于隐变量感知器的语义解析训练算法

### 3.3 特征模板的设计

在基于含隐变量的结构化感知器的判别式学习模型中，特征模板的设计非常重要。在混合树中，结点或者对应于自然语言（NL）词，或者对应于一个MR产生式，每个NL词和子MR产生式都是由它的直接父MR产生式产生的。换句话说，混合树中的所有NL词和子MR产生式都连接到他们的父MR产生式。为了能全面地描述混合树的结构特性，我们共设计了四种类型特征：

1）词特征（Word features）；

2）产生式特征（Production features）；

3）词和产生式的混合特征（Mixture features）；

4）混合模式特征（hybrid pattern features）。

表4中给出了所有类型的特征模板定义。其中，前三种类型特征用于获取父MR产生式和它所有孩子结点之间的相关性。最后一种特征描述由父产生式结点向下延伸的混合模式，具体的说，对于混合树中一个给定的MR产生式结点，混合模式是指该结点下的自然语言的

词序列和其各个子MR产生式结点之间组合的形式。为简化解码过程，在文法GISQL中我们已约定每个MR产生式的右边最多有两个子语义范畴，即含有两个子MR产生式。

表 4. 特征模板

| | |
|---|---|
| Word features | *par*+*w* |
| | *par*+*isConstant*(*w*) |
| | *predicate* (*par*)+*w* |
| | $par+w_{-1}+w$ |
| | $predicate(par)+w_{-1}+w$ |
| Production features | *par*+*p* |
| Mixture features | *p*+*w* / *w*+*p* |
| | *Par*+*p*+*w* / *Par*+*w*+*p* |
| Hybrid pattern features | *par* + *rule* |

其中，*w* 表示自然语言中的词，$w_{-1}$ 表示词 *w* 左边的第一个词，*p* 表示子 MR 产生式，*par* 表示与一个 NL 词或者一个子 MR 产生式直接相关的父 MR 产生式，*rule* 表示一个混合模式；*isConstant*(*w*)用于检查 *w* 是否是已知常量，例如地理命名实体等；*predicate*(*p*)表示从 MR 产生式 *p* 中提取出函数（或谓词）。

**Input**: Sentence *s* with *n* words, a set of candidate MR productions.
Algorithm:
for *len*=1 to *n* do
for *begin*=0 to *n* - *len* do
*end* = *begin* + *len*
for *m* as the next MR production in the set of candidate MR productions do
if *m* has no nonterminal in RHS then
if(*len*=1)
calculate directly the score of the subtree rooted by *m* and covering the only word *s*[*begin*].
else
calculate the score of the subtree rooted by *m* and covering the words *s*[*begin*..*end*] by combining the subtree covering previous (*len*-1) words and word *s*[*end*].
else if *m* has only one nonterminal in RHS then
calculate the highest-scored subtree rooted by *m* and covering the words *s*[*begin*..*end*] by decomposing the subtree according to the value of *len* and corresponding unary hybrid patterns.
else if *m* has two nonterminals in RHS then
calculate the highest-scored subtree rooted by *m* and covering the words *s*[*begin*..*end*] by decomposing the subtree according to the value of *len* and different segmentation of corresponding binary hybrid patterns.
end for
end for
end for
pick out the highest-scored hybrid tree rooted by a start MR production and covering all words.
**Output**: the optimal hybrid tree corresponding to the sentence

图 4. 语义解析中的动态规划解码算法

### 3.4 解码算法的设计

解码算法的目标是根据模型参数找到分值最高的混合树。由于前述的所有特征模板均具有局部性，因此我们设计了一种动态规划解码算法有效地产生最优混合树。

在动态规划的解码算法中，首先让每一个子问题对应于混合树中以某个MR产生式为根的子树，该子树派生自然语言句子中的部分词；然后，根据每个根MR产生式涵盖的词的个数以及根MR产生式相关的所有可能混合模式来分解子问题；最后，依照自底向上的次序求解所有子问题。但是，由于算法中可能的混合模式数量多达21个，从而导致动态规划中的递归表达非常复杂，图4中仅给出了算法的简要轮廓描述。该动态规划算法的时间复杂度为$O(n^2T^2)$，其中$n$为句子的长度，$T$为候选MR产生式的个数。

### 3.5 候选 MR 产生式集合的提取

由于解码算法的时间复杂度不仅依赖于句子的长度，而且还与候选MR产生式集合的大小有关。因此，为了在测试阶段能进一步提高解码的效率和准确率，我们提出了一个基于向量空间模型的MR产生式排序方法来提取相关的MR产生式用于解码，而不是简单地使用所有可能的MR产生式作为候选集合。

类似于文档排序方法[15]，我们利用向量空间模型将相关 MR 产生式的提取问题转换为 MR 产生式排序问题。但是与文档排序问题不同的是，将每个可能的 MR 产生式表示成一个向量是非常困难的。对于训练数据集中的每个实例，它的正确 MR 树均是给定的，而每个训练实例的正确 MR 树中一般都包含多个不同的 MR 产生式，如何建立各个 MR 产生式与自然语言句子中一个词或多个词之间可能存在关联性呢？为了解决这个问题，我们首先设计了一个简单有效的方式来构建每个 MR 产生式的向量。

第一步，对于每一个训练实例通过从其自然语言句子中抽取所有一元、二元、三元词汇串的方式建立一个相应的向量表示；接下来，为了给出每个 MR 产生式的向量表示，我们对包含该 MR 产生式的所有训练实例的向量进行求和，用此和向量作为该 MR 产生式的对应的向量表示。对每个 MR 产生式的向量表示均按此方法计算获取。采用这种计算方法的基本理由是：因为与某个 MR 产生式密切相关的一些词或短语可能会多次出现在其 MR 树中包含该 MR 产生式的训练实例句子中。因此，对包含相同 MR 产生式的实例向量进行相加求和可以导致在该 MR 产生式对应的和向量中与这些词或短语对应的项会具有较高的频度值。

其次，为每个 MR 产生式构建向量表示的另一个重要问题是 MR 产生式向量中每一项的权重如何设置？如果简单按照上述求和方式直接构建每个 MR 产生式向量将会导致在和向量中必然存在很多噪音，为此我们采用一种修改的 tf-idf 权重方案，即通过计算相对词频值来替换传统的词频，因为相对词频值可以更好地反应向量中的各个特征项对于一个 MR 产生式的重要性。

在测试时，对于一个给定的测试自然语言实例，首先按上述方法构造一个向量表示，然后根据余弦相似度计算提取前 $n$ 个相似度最高的 MR 产生式作为该测试实例的相关 MR 产生式集合。其中，$n$ 的值可由句子中包含词的个数确定。

## 4 相关工作比较

在过去的十来年中，研究者们提出了多种基于有监督学习的语义解析模型与算法。Wong (2006) 提出了一种基于统计机器翻译技术的语义解析算法 WASP [6]。该算法从成对的标注训练语料中学习同步上下文无关文法 SCFG 形式的转换规则来捕捉自然语言句子与意义表示之间的关系。Wong (2007) 进一步将 WASP 扩展到处理 $\lambda$ 演算意义表示形式，提出了一种语义解析算法 $\lambda$-WASP[16]。Li (2013) 通过对统计机器翻译领域中经典的同步文法学习算法 GHKM 进行了扩展[17]，用于从成对的自然语言句子与逻辑形式的标注数据集中学习归纳 $\lambda$-SCFG 的规则集，更好地建立了自然语言句子与逻辑形式的对应关系。然而，这些基于 SCFG 规则的语义解析算法主要是在采用基于 $\lambda$ 演算的逻辑形式的意义表示类型的语义解析问题中表现了较好的性能，而本文主要聚焦于函数式（variable-free）的意义表示类型。

Lu (2008) 提出了一种基于生成式模型的语义解析方法[7]，该方法首先定义了一种混合树结构，然后提出一种的生成式模型对自然语言句子和其意义表示关系进行联合建模，在利用生成式模型输出 n-best 结果的基础上，进一步采用一个判别式模型并引入各种非局部特征对 n-best 结果进行重排序。本文中的语义解析算法也借鉴了混合树的结构，但我们将混合树视为一种隐变量，设计了一种有效的判别式学习模型直接实现了语义解析过程，避免了生成式模型中需要引入各种独立性假设的不足。该方法既具有判别式模型能够方便地嵌入各种灵活的特征组合表示的优点，又自然地将解码算法集成在训练与推导阶段。

近年来，基于组合范畴文法 CCG（Combinatory Categorial Grammar）的英文语义解析研究受到了较多的关注[18]。CCG 作为一种能够耦合语法和语义关系的有效语言文法形式，能够对各种语言现象进行描述与建模[19]。但采用基于 CCG 的语义解析方法时，如何获取一个好的、有效的词典是一个非常困难的问题。Kwiatkowski (2010) 则通过使用高阶合一（higher-order unification）的方法定义了一个与训练数据一致的包含所有文法的假设空间，实现了词项的自动生成，从而避免了人工设计规则模板的复杂性[8]。

## 5 实验结果与分析

基于我们开发的包含 1110 条实例的中文语义解析标注语料库，采用我们提出的含隐变量的感知器模型的语义解析算法进行了十折交叉验证实验，并计算其微平均（micro-averaged）结果。实验的评价指标采用了传统的准确率（precision）、召回率（recall）和 F1 值。其中，对于每个测试实例预测正确性的判定方法是：当预测产生的 MR 树与对该实例标注的正确 MR 树完全一致时，才认为该实例的测试输出是正确的。

### 5.1 候选 MR 产生式集合提取方法的有效性验证

为提高测试阶段的效率与准确率,我们提出了一种基于排序方法的候选 MR 产生式集合抽取方法,为了验证该方法的有效性,我们进行了两组十折交叉验证对比实验，实验结果如表 5 所示。表中第一行（LP）表示采用隐变量感知器模型进行训练，在测试时使用所有的 MR 产生式作为候选产生式集合；而第二行（LP+EXT）表示采用同样的隐变量感知器模型 LP 进行训练，但在测试时对每个每个测试实例分别使用排序访法抽取一个更小的候选 MR 产生式集合后进行解码。从表中的实验结果可以看出，通过基于排序法实现更小 MR 产生式候选集合的抽取明显改进了时间效率，将总的测试时间缩短了将近 2/3。同时，语义解析的准确率也得到了显著的提高，F1 值提高了 3.2%，获得了 25.5%的错误减少率。

表 5. 增加候选 MR 产生式集合提取方法的实验结果对比

| | Time(秒) | Precision(%) | Recall(%) | F1(%) |
|---|---|---|---|---|
| LP | 126 | 87.47 | 87.47 | 87.47 |
| LP+EXT | 43 | **90.71** | **90.63** | **90.67** |

### 5.2 不同语义解析算法的实验比较

为了能够验证我们的方法在中文 GIS 自然语言接口实现中的有效性，我们也实现了两个 baseline 系统。我们选择 Lu (2008) 提出的产生式模型并结合重排序的后处理过程[7]，以及 Kwiatkowski (2010) 提出的基于 CCG 文法和采用高阶合一方法自动构造词典的的语义解析模型[8]构造了两个 baseline 系统，分别记为 baseline-1 和 baseline-2。因为这两种方法是目前英文语义解析研究中性能领先的基于有监督学习模型，而且它们也不需要任何额外的语法先验知识，因而这两种方法和我们的方法具有直接的可比较性。

表 6 中的实验结果显示，在 F1 值上，我们的系统比 baseline-1 系统获得了 4.11%的提高，相对于目前在英文语义解析任务中具有最佳解析性能的 baseline-2 系统也高出了 1.77%。同时注意到，我们系统的召回率和准确率几乎相等。这意味着对于几乎所有的测试实例，我们

的系统都能解析出一个意义表示树结果。一个可能的原因是因为我们的方法是基于判别式结构化预测模型，它能够很好的集成各种有效的特征组合，因而对一些训练数据中未见的 MR 产生式具有一定的平滑作用。

表 6. 不同方法的实验结果对比

| | Precision(%) | Recall(%) | F1(%) |
|---|---|---|---|
| Baseline-1 | 88.82 | 84.41 | 86.56 |
| Baseline-2 | 91.08 | 86.84 | 88.90 |
| Our | **90.71** | **90.63** | **90.67** |

## 6 结束语

本文针对基于语义解析的中文 GIS 自然语言接口实现技术与方法进行了探索性的研究。我们选择南京市地图查询作为具体的实际应用领域，首先设计了一个形式化意义表示语言 GISQL，并在基础上开发了一个相应的中文语义解析标注语料库。据我们所知，这也是第一个中文语义解析语料库。然后，我们提出了一种基于含隐变量的感知器模型的语义解析算法。在开发的中文语义解析标注语料库上的实验结果显示，该算法的 F1 值达到了 90.67%，明显优于两个 baseline 系统。更重要的是，本文的研究结果证明了基于语义解析方法实现中文 GIS 的自然语言接口是一种有效可行的途径。

在下一步的工作中，我们将扩展形式化意义表示语言 GISQL 和语料库，以覆盖更广泛的 GIS 应用领域与问题，包括地图浏览、数据采集和空间分析等领域；另外，我们将研究基于启发式搜索的结构化学习算法，这样能够引入更多非局部化的特征描述混合树结构，从而会导致更好的语义解析性能。

## 参考文献


[1] 张连蓬，储美华，刘国林，江涛. 车载智能地理信息查询系统及其自然语言接口[J]. 现代测绘, 28(1): 20-23, 2005.

[2] 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), 28 (3): 301-305, 2003.

[3] S. Mador-Haim, Y.Winter, and A. Braun. Controlled language for geographical information system queries[C]//*Proceedings of Inference in Computational Semantics*, 2006.

[4] 余明朗, 明小娜, 龙 毅, 张雪英. GIS 环境下中文命令的规则匹配与语义解析[J]. 地理与地理信息科学， 28(6): 7-12, 2012.

[5] R. J. Kate, Y. W. Wong, and R. J. Mooney. Learning to transform natural to formal languages[C]//*Proceedings of AAAI*, 2005:1062-1068.

[6] Y. W. Wong and R. J. Mooney. Learning for semantic parsing with statistical machine translation[C]//*Proceedings of the HLT-NAACL*, 2006:439-446.

[7] Wei Lu, Hwee Tou Ng, Wee Sun Lee and Luke S. Zettlemoyer. A Generative Model for Parsing Natural Language to Meaning Representations[C]//*Proceedings of EMNLP*, 2008:913-920.

[8] Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater and Mark Steedman. Inducing probabilistic CCG grammars from logical form with higher-order unification[C]//*Proceeding of EMNLP*, 2010: 1223–1233.

[9] John M. Zelle and Raymond J. Mooney. Learning to parse database queries using inductive logic programming[C]//*Proceedings of AAAI*, 1996:1050–1055.

[10] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables[C]//*Proceedings of ICML*, 2009.

[11] Junsheng Zhou, Juhong Xu, Weiguang Qu. Efficient Latent Structural Perceptron with Hybrid Trees for Semantic Parsin[C]// *Proceedings of the IJCAI*, 2013:2246-2252.

[12] Michael Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms[C]//*Proceeding of EMNLP*, 2002.

[13] Xu Sun, Takuya Matsuzaki, Daisuke Okanohara Jun'ichi Tsujii. Latent Variable Perceptron Algorithm for Structured Classification[C]//*Proceedings of IJCAI*, 2009:1236-1242.

[14] Ryan McDonald. Discriminative Training and Spanning Tree Algorithms for Dependency Parsing[D]. University of Pennsylvania, *PhD Thesis*, 2006.

[15] D.L. Lee, H. Chuang and K. Seamons. Document Ranking and the Vector-Space Model[J]. *IEEE Software*, 1997, 14(2): 67-75.

[16] Yuk Wah Wong, Raymond J. Mooney. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus[C]//*Proceedings of ACL*, 2007:203-210.

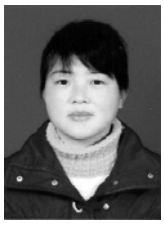
[17] Peng Li, Yang Liu and Maosong Sun. An Extended GHKM Algorithm for Inducing $\lambda$-SCFG[C]//*Proceedings of AAAI*, 2013:605-611.

[18] L. S. Zettlemoyer and M. Collins. Online learning of relaxed CCG grammars for parsing to logical form[C]//*Proceedings of EMNLP-CoNLL*, 2007:678-687.

[19] Mark Steedman. The Syntactic Process[M]. The MIT Press, Cambridge, Mass，2000.

周俊生（1972–），男，博士，副教授，主要研究方向为自然语言处理、语义解析。Email: zhoujs@njnu.edu.cn；

曲维光（1964–），男，博士，教授，主要研究方向为自然语言处理、计算语言学、人工智能。Email: wgqu@njnu.edu.cn；



许菊红（1987-），女，硕士研究生，主要研究方向为自然语言处理。

Email: xujuhong1987@163.com；

龙毅（1968–），男，博士，教授，主要研究方向为地理信息系统。Email: Longyi@njnu.edu.cn；

朱耀邦 (1989–)，男，硕士研究生，主要研究方向为自然语言处理。Email: zhu.yaobang@163.com.