

基于语义知识的汉语句法分析

王 进

(知识产权出版社 北京 10008)

kingsten_88@hotmail.com

摘要 语义知识是句法分析的必要基础, 缺乏语义知识, 自动句法分析将始终处于浅层的结构分析。基于概念依存的关系, 建立一套语义知识形式化体系, 使得词语内涵的描述和语法结构的描述得到统一, 让各级语言单位之间信息计算成为可能。文章根据语义知识体系改进了传统 chart 算法的形式规则, 增强了形式规则排除歧义的能力。

关键词 语义知识, 原型系统, 特征结构, chart 算法。

A Semantic-Based Syntactic Parsing for Chinese

Wang Jin

(Intellectual Property Publishing House, Beijing, 10008, China)

kingsten_88@hotmail.com

Abstract: Semantic knowledge system is essential for automatic syntactic parsing. If lacking of semantic knowledge, automatic syntactic parsing will always works on shallow syntactic structure far from semantic calculation. For achieving the purpose of semantic calculation between syntactic units of all levels, it is important to make the unification of expressions about the senses of words and the characteristics of syntactic units through establishing a set of formal systems of semantic knowledge basing on the dependency rule between concepts. This article improves the syntactic rules of traditional chart parsing approach according to a truly original semantic knowledge system, which greatly depreciates the ambiguity of syntactic rules.

Key words: Semantic Knowledge, Prototype System, Feature Structure, and Chart Parsing

自然语言处理技术, 特别是自动句法分析, 其优劣好坏终究取决于对语言知识的掌握程度。从语言处理的层次来看, 语言知识可以分为语言符号、语法结构、语义范畴以及语境语用四个层级的知识, 这四个层级的语言知识自下而上表现了语言不同层面的组成元件。但是, 在自然语言处理中, 这四个层级的语言知识始终融为一体, 不能完全分隔开来, 系统在设计语言知识表示方法的时候, 必须要考虑到各层级语言知识的综合要求。

长期以来, 句法分析系统所依赖的语言知识多集中在语言符号和语法结构两个层面上, 由于缺乏对语义方面的描述, 难以进行语义信息的计算, 也就难以达到自动句法分析的应用目标。自动句法分析的目标不仅是解析出句子的语法结构, 更重要的是, 要确定出各个字词的真实含义, 以及字词之间的语义关系。对字词含义的选择和语义关系的推断, 都是需要深层次语义知识作为基础的。利用语法知识或者一些统计信息进行句法分析, 虽然能够建立起可能成立的语法结构, 但是, 要获取较为准确的语义关系, 还必须通过语义知识加以验证。

一个句法分析系统必须全面构建语言信息，才能为句法分析奠定“理解”的基础。本文出于对语义知识的理解，建立了一套语义知识的描述系统，可以描述任何复杂的语义范畴，并为句法分析设计了一套形式规则描述语言，该规则具有综合描述能力，易于扩充、表达自然。

1 词汇语义知识

语义知识涉及内容很广，包括研究词语关系的词汇语义学和研究语法结构的形式语义学。词汇语义学，主要研究词语内涵的表示，以及词语内涵之间的各种关系。形式语义学主要以语法结构为研究对象，用形式化手段描述各种语言现象，并加以逻辑演算。词汇语义学是语言知识的基础，具有非常重要的作用，本文所介绍的语义知识主要是指词汇语义知识，而形式语义将在句法分析中加以介绍。

1.1 词语内涵的表示

词语内涵是对词语凸显的语义特征进行分类的结果，以不同的特征分类，可以得出不同的特征分类体系。一般来讲，词语内涵分类(Category)是从本质特征(Essence)和从属特征(Attribute)两个方面进行分类，本质特征是唯一的，而从属特征数量不限，本质特征和从属特征之间是附属关系，这种关系可以形式化为：

$$C = \langle A, \dots \rangle E \quad (1)$$

无论本质特征，还是从属特征，都是相对的，也就是说，不能孤立地为某一个符号指派意义，而是通过相互区别做出的分类结果。因此，这就涉及到对本质特征和从属特征各自分类标准的讨论。

本质特征是外延对象在认知世界中的存在形式，事物的存在形式就是与相关事物的作用关系，因此，本质特征的分类标准就是作用关系。比如，“吃”与“喝”，二者与客体的作用关系均是“吸纳”，所以本质特征是相同的，不同的是“吸纳”对象所表现的从属特征。又如，“吃”与“看”，二者与客体的作用关系均是“吸纳”，但是，“看”和“吃”的“吸纳”方式有所不同，由于“看”的吸纳对象是信息，所以，“看”并没有对信息有直接改造作用，相反，是信息对“看”的主体产生了直接刺激作用，因此，“吃”与“看”的本质特征是相反的。

体词是不存在作用关系的，其本质特征是从作用关系中的角色来分的，即时间、处所、人、物、事体、性质、状态等。无论体词，还是谓词，范畴划分的标准必须上下一致，永不变轨。

从属特征是词语内涵存在的各种“偶现”和“非偶现”的表现，包括对象本身的性质、状态和与外界发生的联系。比如，“一个被扔弃的破瓶子”，该词的从属特征有“一个”、“被扔弃”、“破”、“容器”四个方面，其中，前三个特征是偶现的，而最后一个是“非偶现”的。词语内涵的从属特征是根据词语含义挖掘出来的，不是从现实事物中归纳出来的，因为我们要描述的是词汇，而不是一般存在实例。我们立足于词汇的内涵寻找从属特征，就是避免陷入经验主义和唯理主义的纷争。从属特征本身没有分类体系，而是基于本质特征分类体系的描述。

下面就本质特征的分类系统（也称为原型系统）做一个介绍：

根据作用关系的异同，词语内涵大致划分三类：0元、1元、2元。0元就是本体描述类，它不存在作用关系，1元是性状描述类，它有作用关系，但处于隐性状态，2元是关系描述类，它具有显性的作用关系；其中，2元显性作用关系又分为表示主从关系的“支配描述”

和表示注解关系的“关联描述”，根据作用方式的异同，可以依次继续划分下去。分类体系的上层框架如下：



按照作用关系进行分类，实际上是寻找认知世界中各种微观“作用力”。每个“作用力”表示了最基本的语义关系。因此，对复杂的语义分析，就是深入解构并还原成这些最基本的作用关系，我们可以将这些作用关系形式化为如下三个基本表达式：

0元 (f) (2)

1元 (a:f) (3)

2元 (a, f, b) (4)

其中，f表示作用关系，a、b表示作用对象，a为作用关系的主体，b为作用关系的客体。作用对象为一定范畴意义的变元，在特定语境中，作用对象可能为一个实例。

通过上述的(1)(2)(3)(4)基本式子，我们可以描述任何复杂的语义范畴，而不仅仅限于词汇。下面列举一些词语内涵形式化的例子（其中，p、s、o 分别代表中心作用关系及其主体和客体，a、b 等代表作用对象变元，数字代表作用关系编号）。

| | 词例 | 伪范畴表达式 | 范畴表达式 |
|-----|------|--------------------------------|---|
| 名词 | 花生 | ⟨⟨(食物)⟩植物生成体⟩ | ⟨⟨(110)⟩ ¹⁰²¹ ⟩ |
| | 好人 | ⟨⟨(p:品行好)⟩人⟩ | ⟨⟨(p:2453)⟩ ⁰⁰⁰⁵ ⟩ |
| | 唐装 | ⟨⟨(艺术风格名 唐朝风格), (衣物)⟩人工物⟩ | ⟨⟨(2033 唐朝风格), (2672)⟩ ⁰⁰¹² ⟩ |
| | 红色 | ⟨(颜色 (255, 0, 0))⟩ | ⟨0231 (255, 0, 0)⟩ |
| 动词 | 吃饭 | (a, 吃, ⟨⟨(食物)⟩ ^b ⟩) | (a, 2015, ⟨⟨(110)⟩ ^b ⟩) |
| | 开花 | (花: 出现) | (0136: 2011) |
| | 消灭 | (a, 使之消逝, b) | (a, 2231, b) |
| | 击打 | (a, 碰击, b) | (a, 2481, b) |
| 形容词 | 美丽 | (a: 优美性 10) | (a: 3210 10) |
| | 孩子气的 | (a: 城府 2) | (a: 3011 2) |
| | 可憎的 | (a, 属性有, (b, 厌恶, s)) | (a, 2731, (b, 2612, s)) |
| | 导电的 | (a, 属性有, (s, 传导, 电)) | (a, 2731, (s, 2835, 0112)) |
| 副 | 碰巧 | (a: 巧合性 10) | (a: 3203 10) |

| | | | |
|----|-------|--------------|---------------|
| 词 | 及时 | (a: 及时性 10) | (a: 3281 10) |
| | 早晚 | (a: 时长评估 5) | (a: 3229 5) |
| 介词 | 向 | (a, 趋向关系, b) | (a, 2415, b) |
| | 关于 | (a, 内容涉及, b) | (a, 2437, b) |
| 连词 | 因为…所以 | (a, 因果联系, b) | (a, 2510, b) |
| | 如果…那么 | (a, 条件假设, b) | (a, 2582, b) |

原型系统分类体系，目的是定义各种基本的作用关系和元概念范畴，为语义范畴的描写提供基本参照体系。相比其他知识库，该语义描写方法有如下优点：一、利用基本式子描述，表达式书写灵活，不仅可以描述词语，还可以描述短语，不同语言单位可以融合计算；二、统一了各词类范畴的分类，简化了语义比较算法；三、范畴代码化，保持语义知识的“中立性”，可以作为各个自然语种意义交换的中介。

1.2 词语内涵之间的关系

语言符号系统中存在大量的语义关系，比如，近义关系、反义关系、上下位关系、整局关系、致使关系、蕴含关系等等，这些关系在逻辑推理中是必不可少的语言知识。严格讲，这些关系存在于元概念之间，而不存在于词语内涵之间，因为词语内涵是本质特征和从属特征的综合描述，其意义不是单纯的，语义关系也是多方面的。

从元概念讲，这些关系是原型系统的一部分，因此，原型系统应该具有表示这些关系的能力。从原型系统的分类原则看，异级分类保持了上下位关系，同级分类保持了近义、反义关系，但是，目前原型系统还缺乏其他语义关系的描述。鉴于句法分析的实用性，下面介绍两种主要的语义关系：主述关系和因果关系。

主述关系是主题与陈述内容的关系，这种关系反映在语法上，就是广义的主谓结构（述宾结构、偏正结构从原型上也可以看作主谓结构）。一般来讲，主题部分是一个话题的起点，而陈述部分是与主题相关的性质、状态和内容。句子的基本结构就是主题和陈述的二元结构。在元概念上，主述关系就是作用关系和作用对象的关系，即“什么作用关系下有什么作用对象”。

因果关系是作用关系及其作用下其主客体所发生的变化之间的关系，一般来讲，作用关系的主客体由于受到改造或者刺激会有所反应，比如，“敲打”可能让客体在形体或者肉体上产生两种变化，一类是“破碎”，一类是“伤痛”。这种变化反映在语法上，就构成了广义的述补结构（兼语结构、联动结构也可以看作述补结构）。句子中的述补结构总是两个谓词中心，前一个谓词表原因，后一个谓词表结果（或者目的），二者之间有着直接或者间接的联系，通过述补结构表达出事物之间的前因后果。事实上，因果关系反映了客观世界“作用”与“变化”的基本规律，比如：“隐现变化”、“动静变化”、“离合变化”、“增减变化”、“性质变化”、“时空顺序变化”等。在原型系统的分类标准中，以“作用关系”为划分原则的，其中之一就包括“作用”的结果。

在原型系统的节点上，增加以上的关系描述，使得原型系统具备一定的推理能力。比如，在节点“敲打”上有了“破碎”和“伤痛”的结果描述，那么，针对所有本质特征为“敲打”和“破碎”或者“伤痛”的词语都能建立起搭配关系，这有利于句法分析对述补

结构的辨识。当然，原型系统所描述的因果关系是直接作用的致使关系，对于间接作用下的因果关系没有体现，比如，“他打哭了小孩”，其中“打”和“哭”不是直接的作用关系，而是“打”-“伤痛”-“哭”这样的逻辑关系，因此，还需要通过若干因果关系进行递推才能获得，这又需要更多的因果关系知识来推理。

2 基于语义知识的汉语句法分析

要实现语法结构自动解析，除了掌握好词语的语义范畴，还需要有一套规则对句子结构进行推理，将搭配得当的词语结合起来，这就是形式语义学要研究的内容。下面将结合词汇语义知识，制定适合汉语句法分析的形式规则，并以传统的chart算法加以实现。

chart算法的关键是为特定语种制定一套合理的语法标记和形式规则。语法标记的设计十分巧妙，既要充分反映语法结构的具体特征，又要避免过分具体而使得形式规则数量过于庞大。传统上以短语结构作为语法标记，标记特征单一，但歧义性很普遍，不利于语法结构的建立。如何利用有限的语法标记同时又能覆盖各种语法结构呢？一般来讲，利用特征结构是最好的解决途径。本文就是通过特征结构来描述形式规则，除了短语结构标记外，还增加了短语的语义范畴等信息，这样的规则具有更充分的限制条件，从而降低了规则的歧义性。

2.1 特征化的形式规则

形式规则采用特征结构描述，每个语言特征都表示为“（属性：值）”的形式。形式规则描述语言用BNF范式表示如下：

```
<规则> ::= <语法标记> “→” <特征> “+” <结论>
<语法标记> ::= 主谓 | 述宾 | 偏正 | 述补 | 同位 | 合一 | 并列 | 体词 | 谓词 | 结构助词 | 时态助词
<特征> ::= <特征项> { “&” <特征项> }
<特征项> ::= <对象名> “(” <属性> “:” <值列表> { “;” <属性> “:” <值列表> } “)”
<对象名> ::= <序列号> “-” <路径>
<序列号> ::= n
<路径> ::= <分支方向> { “-” <分支方向> }
<分支方向> ::= top | lson | rson
<属性> ::= lex | pos | sem | stdcat | cluster | present | relation | ref
<值列表> ::= “{” <值> { “|” <值> } “}”
<结论> ::= <语法单元>
<语法单元> ::= “(” <单元结构> “SPACE” <对象列表> “)”
<单元结构> ::= <语法标记> “-” <单元中心>
<单元中心> ::= <序列号>
<对象列表> ::= <对象> { “SPACE” <对象> }
<对象> ::= <语法标记> | <衍生词> | <语法单元>
<衍生词> ::= <词语义项> | <语义范畴表达式>
一条形式规则由<特征>和<结论>两部分构成，<特征>部分描述了规则的限制条件，而<
```

结论>部分描述了生成单元的语法结构。〈对象名〉是〈序列号〉和〈路径〉的综合表示，是为了对生成单元底层对象的定位。该描述语言所用属性是：

| 属性 | 名称 | 值域 | 知识层级 |
|----------|---------|--|------|
| lex | 字面文本 | 文字符号 | 语言符号 |
| pos | 词性或者子词性 | 所有语法词类及子词类 | 语法结构 |
| cluster | 关系簇 | 词语义项集合名称 | 语法结构 |
| present | 存现情况 | 存在或者不存在 | 语法结构 |
| sem | 语义范畴 | 语义范畴表达式 | 语义搭配 |
| stdcat | 标准范畴 | 语义范畴表达式集合 | 语义搭配 |
| relation | 语义关系 | 近义/反义/上位/下位/整体/局部/原因/结果/主体/客体(相对于另一个对象名) | 语义搭配 |
| ref | 指引对象 | 另一个对象名 | 语境语用 |

〈属性〉中有两个集合型属性：关系簇和标准范畴。关系簇是将特定事件所涉及的事物按照特定角色建立在一起，比如，以“考试”事件为例，其中所涉及的事物有“考生”、“考卷”、“考场”、“考题”、“考试科目”、“考分”、“监考人”、“考试纪律”等。将这些事物集中放在一个“考试”的关系簇中，让每个事物各自扮演不同的角色。关系簇知识有利于分析语篇的中心主题，根据同一个事件域的事物的共现情况，可以快速理解所讲的内容。

标准范畴是指对一些相似的概念建立一个大的范畴集合，方便语义知识的描写。例如，可以建立一个标准范畴表达“处所”，实际上，“处所”这个范畴可能具体包括“地名”、“方位”、“建筑”、“机构”、“社团”等等。标准范畴将这些范畴集结在一起，形成一个比较固定、比较常用的范畴，便于我们书写语义搭配的知识。

从〈属性〉的种类来看，属性可以涵盖语言符号、语法结构、语义范畴、语境语用等多个层次的知识，尤其是利用了语义范畴表达式来表示特征，增强了形式规则的描写能力。例如，如果要描述这样的语言现象：“如果前一个中心词为时间、处所、或者表示时间、处所的方位词，后一个中心词范畴为0200（操作行为），则两个中心词构成状谓结构”，那么用上述描述语言可以写成下面的规则：

偏正→0-top(pos: {s|t|f:s|f:t}) & 1-top(sem: {(a, 0200, b)})+(偏正-1 体词 谓词)

该规则有两个特征，前一个特征关于语法词性，后一个特征涉及语义范畴，从两种不同层级的知识描述了这个语言现象。

2.2 句法算法实现

一般规则型句法分析系统都是基于上下文无关文法及有限状态自动机模型的，chart算法就是这种系统的典型代表。chart算法通过扫描、归约、预测三个环节，逐层发现各级语言单位，最终形成覆盖全句的语言单位，该算法的优点是利用了预测和回溯的手段，可以发现规则内任何合理的语法结构，在信息存储和算法流程上，也都是比较简约的。

本文所述的形式规则完全适用于chart算法，只是需要在扫描和归约时，对规则加强特征信息的验证，验证过程如下：

- a) 设从 agenda 中弹出一条边，为〈P1, P2, L〉；

- b) 又设从活动边中获得一条边，为 $\langle P_0, P_1, A \rightarrow \alpha \cdot L \beta \rangle$ （从形式规则中获得的活动边可看作 $\langle P_1, P_1, A \rightarrow \cdot L \beta \rangle$ ）；
- c) 将L的中心以及 α 的中心带入 $A \rightarrow \alpha \cdot L \beta$ 规则中，进行特征验证，如果合格，则增加一条活动边，为 $\langle P_0, P_2, A \rightarrow \alpha L \cdot \beta \rangle$ ；
- d) 如果活动边的点移动到规则的最右端，则该活动边转变为非活动边，建立非活动边的单元中心，然后将该边压入 agenda。

每一条非活动边都必须确定中心，因为每一个语法结构的特征都是通过单元中心体现出来的，这个中心可能有三种情况：1) 选取生成单元之一；2) 通过现有生成单元进行衍生；3) 动态插入中心；这些信息都要在形式规则的 \langle 语法单元 \rangle 中被描述出来。一旦建立中心，那么，非活动边不再是一个简单的语法标记，而是一个有语义范畴的“概念”。

2.3 实例分析

设分析词语序列“今天星期天”，该序列由两个时间词组成。但是两个时间词组成的序列也有三种关系：1) 合一关系，如“今天上午”；2) 并列关系，如“今天明天”；3) 同位关系，如“今天星期天”。如果形式规则对两个时间词不加以语义限制，那么这三种关系都可能成立，结果将生成两种可能有误的结构。为了排除这些歧义，形式规则就需要一定的限制条件，从语义范畴上增加一些特征信息，以便区分各种形式规则。假设设计了如下三条形式规则：

1) 合一关系：合一 $\rightarrow 0\text{-top}(\text{pos:}\{t\}) \& 1\text{-top}(\text{pos:}\{t\}) \& 1\text{-top}(\text{relation:}\{\text{局部:}\{0\text{-top}\}\}) + (\text{合一-1 体词 体词})$

2) 并列关系：并列 $\rightarrow 0\text{-top}(\text{pos:}\{t\}) \& 1\text{-top}(\text{pos:}\{t\}) \& 1\text{-top}(\text{relation:}\{\text{近义:}\{0\text{-top}\}\}) + (\text{并列-01 体词 体词})$

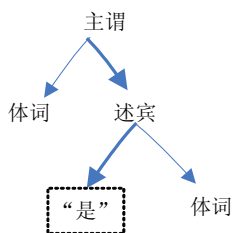
3) 同位关系：同位 $\rightarrow 0\text{-top}(\text{sem:}\{(0222)\}) \& 1\text{-top}(\text{sem:}\{(0231)\}) + (\text{同位-0 体词 体词})$ ，其中“0222”为“相对时间”的范畴标记；“0231”是“周期性时间”的范畴标记。

这些规则都增加了词汇语义范畴信息，通过特征比较运算，就能更加明确体词和体词之间所能构成的语法结构。

如果以上词语序列是一个句子，则还需要另外设计一条形式规则，以便获取句子的中心谓语：

主谓 $\rightarrow 0\text{-top}(\text{sem:}\{(0222)\}) \& 1\text{-top}(\text{sem:}\{(0231)\}) + (\text{主谓-1 体词 (述宾-0 (501) 体词)})$ ，其中“501”为判断词“是”的义项编号。

该形式规则的 \langle 结论 \rangle 部分显示了目标语法结构（见下图），该语法结构增加了一个词语序列中没有的词语“是”，这是整个语法结构的中心。由于语言表述中常常会遇到中心词省略的情况，在进行语义分析时，有必要补充出来，保证语义的完整性和结构的一致性。



这个规则意味着，对于形式规则的描述，不仅要增加一些特征信息限制规则的歧义性，而且要对各个组成的成分结构化，这是因为中文的词法功能和句法功能并不能一一对应，比如，作为一个谓词，既可能做定语、状语、补语；也可能做主语、谓语、宾语。

另外，词语在具体的表述中，还会出现“范畴变迁”的过程，比如，中文句子“我吃大碗”，其中“吃”不再是原来的意义“进餐”，而转变为“使用”这样的意义，在确定单元中心时，就需要加以修正。再有，词语在特定上下文环境中，还会出现“词干分离”的现象，比如，中文句子“他滑雪滑丢了钱包”，其中第二个“滑”字的含义实际上是上文“滑雪”的简称，并不是单字“滑”的意思。因此，在形式规则中，需要明确描述出单元中心的位置。

词语一旦应用到句子中，就不能再看作固守原义的“符号”，在特定的应用场景会发生形态或者意义的变化，很多时候，与它原来的意义有很大的差距，这种变化需要通过语义信息在上下文中进行推测，这就是特征结构描述的内容，基于短语结构的语法将无法做到。

3 结束语

语义知识建设是一个繁重的工作，但是，在探索自然语言的生成机制上，我们又绝对不能回避它。语义问题的核心是词汇范畴的表示问题和形式语义的描述问题，只有解决好这些问题，才能深层次描述语言现象，加强形式规则的限制条件，从而降低形式规则的歧义性。

参考文献

- 1 徐烈炯. 生成语法理论. 上海外语教育出版社, 1996.
- 2 冯志伟. 机器翻译研究. 中国对外翻译出版公司, 2004.
- 3 Daniel Jurafsky & James H. Martin. 自然语言处理综论. 电子工业出版社, 2005.
- 4 赵艳芳. 认知语言学概论. 上海: 上海外语教育出版社, 2001.
- 5 王寅. 认知语言学的哲学基础: 体验哲学. 外语教学与研究, 2002.
- 6 俞士汶. 中文概念词典规格说明. 北京大学计算语言学研究所, 2003.
- 7 石毓智. 现代汉语语法系统的建立-动补结构的诞生及其影响. 北京语言文化大学出版社, 2002.
- 8 列维·布留尔. 原始思维. 商务印书馆, 1995.