

# 一种汉语句子语义表达模型及其实现策略

黄 海      姚 天 昉

(上海交通大学计算机科学与工程系, 上海 200030)

**摘 要:** 本文介绍了作者运用句子语义描述语言 FAS (Functor Argument Structure) 建立的汉语句子语义表达模型 C-FAS, 并介绍了其实现策略。C-FAS 主要抓住了汉语句子中谓语动词与主宾语之间、限定词与被限定词之间等关系, 并以统一的形式描述为函词和变元关系。最后介绍了从 GPSG 描述的句法结构向 C-FAS 结构的转换实现策略。

**关键字:** 语义表达模型, 语义分析

## A Model for the Representation of Chinese Sentence Semantics and its Realization Strategy

Huang Hai      Yao Tianfang

( Shanghai Jiao - Tong University , Shanghai 200030 )

**Abstract :** In this paper a model for the representation of semantics of Chinese sentence (C-FAS) is discussed, which is created by us using the sentence representation language FAS. The relations between the subject and object, the determiner and the determined etc; are mainly described. An introduction of the strategy of realization which transfers the syntactic structure in GPSG grammar to the C-FAS semantic structure is also presented.

**Key Words :** semantic model, semantic analysis

### 1 引 言

随着语言学和计算语言学的发展, 汉语语义方面的研究取得了显著的成果。无论对词汇的语义, 还是句子篇章等语法单位一级的语义分析研究工作都取得了一定的进展。本文尝试建立一个汉语句子的语义表达模型并讨论其实现策略。它基于GPSG语法分析汉语句法的结构表示, 使用一种自然语言的句子语义表达结构FAS并根据汉语本身的特点, 来建立汉语句子的语义表达模型。在此过程中, 建立GPSG句法结构与FAS语义结构之间的对应关系, 解决从句法分析所得的GPSG句法结构到FAS结构之间的转换是本文讨论的重点。

## 2 汉语句子的句法结构

笔者采用广义短语结构文法 ( GPSG ) 来进行句法分析, 是因为它具有许多特点。GPSG主张句子只有一个语法平面, 即表层结构。GPSG用特征-值对来描述语法, 坚持严格的形式化。GPSG使得每一个句法结构都和一个语义结构相匹配, 句法分析和语义分析紧密地结合在一起, 相互作用, 相互制约。运用GPSG对汉语进行句法分析, 保留了上述特点, 便于使用蒙太格文法风格的语义解释。

同时, 也必须注意到, 汉语有着不同于印欧语系语言的特点: 汉语缺乏严格意义上的形态变化。同一词作为不同的语法成分出现时, 词的形态保持不变。汉语中语法成分的结合没有形式标识, 显得比较松散自由, 往往只要两个成分在意义上相配合便可组合。汉语中词的语法功能较复杂, 词类与相应的句中语法成分不能一一对应。

汉语的句子, 除了通常的主谓句外, 还存在大量的非主谓句及一些特殊句型 and 结构, 如“是...的”句, 存现句, 主谓谓句, “是”字句, “有”字句, “把”字句, “被”字句等。

为了描述汉语句法结构, 对GPSG进行了部分修改, 主要是:

(1) 扩充 CONTROL 特征集为:

$$\text{CONTROL} = \{ \text{S\_agr}, \text{Do\_agr}, \text{Io\_agr}, \text{SLASH} \}$$

其中新引进 S\_agr 表示主语, Do\_agr 表示直接宾语, Io\_agr 表示间接宾语。

(2) 针对汉语特殊句型引入相应的规则, 例如:

1) “是...的”句:

$$\text{VP} \Rightarrow \text{V}[13], \text{VP/NP}, \text{Aux}[\text{de}] \quad \text{VP} \Rightarrow \text{V}[13], \text{S/NP}, \text{Aux}[\text{de}] \dots$$

2) 存现句:

$$\text{VP}[\text{INV}] \Rightarrow \text{V}[1], \text{NP}[\text{S\_agr}] \quad \text{VP}[\text{INV}] \Rightarrow \text{V}[10], \text{NP}[\text{S\_agr}]$$
$$\text{VP}[\text{INV}] \Rightarrow \text{V}[11], \text{NP}[\text{Do\_agr}]$$
$$\text{S} \Rightarrow \text{LocalP}, \text{VP}[\text{INV}] \quad \text{S} \Rightarrow \text{TimeP}, \text{VP}[\text{INV}] \dots$$

3) 主谓谓句:

$$\text{S}[\text{PRD}] \Rightarrow \text{NP}, \text{S} \quad \dots$$

4) “是”字句:

$$\text{VP} \Rightarrow \text{V}[12], \text{VP} \quad \text{VP} \Rightarrow \text{V}[12], \text{NP}$$
$$\text{VP} \Rightarrow \text{V}[12], \text{AP} \quad \text{VP} \Rightarrow \text{V}[12], \text{TimeP}$$
$$\text{VP} \Rightarrow \text{V}[12], \text{LocalP} \quad \text{S} \Rightarrow \text{VP}, \text{VP}[12]$$
$$\text{S} \Rightarrow \text{S}, \text{VP}[12] \quad \text{S} \Rightarrow \text{NP}, \text{VP}[12] \dots$$

等等。

此外, 对于有连词的并列结构、名词结构以及其他如补语结构等, 也都有相应的句法规则。

针对汉语的特点, 笔者把汉语词类划分为十五个大类, 每个大类下又划分了若干小类。如普通名词大类下又分为表人名词, 处所名词, 方位名词, 时间名词, 专用名词, 集合名词, 抽象名词等七大类。这便于适应汉语的特点来进行GPSG句法分析。

当然这里介绍的仅是对部分汉语句法的分析处理，对于诸如否定句等处理尚未考虑。

### 3 基于句法结构的语义表达模型

对汉语语义的分析，笔者考虑句子中几种主要的关系。包括动作与动作的主、宾语之间的关系，修饰成分与被修饰成分之间的关系，限定成分与被限定成分之间的关系。这些关系可以以一种统一的形式看作为函词与变元之间的关系。语义表达语言 FAS 正是抓住了这一核心问题。

FAS 是一种自然语言句子语义的表达结构，可用于机译系统的句子语义表达。它克服了GPSG文法在表达句法结构时加入语义信息的困难。运用 FAS 结构对汉语句子语义进行描述，首先在句子一级从整体上抓住作为句子核心的动词与主语、宾语之间的作用关系。同时，也表达了组成句子的名词短语、介词短语等的限定与被限定、修饰与被修饰等关系，表达了它们的内部语义结构。

笔者针对汉语的特点，建立了汉语的 FAS 结构 C-FAS。它由三部分组成，包括范畴 (CAT)、基本词汇 (LEX) 和规则集 (RULE)。它的形式描述如下：

定义 1:  $C-FAS = \langle CAT, LEX, RULE \rangle$ ，其中 CAT 是范畴的集合，LEX 是词汇的集合，RULE 是转换规则的集合。

定义 2:  $CAT = \{ \langle h, \{ \langle m, v \rangle \mid m \in hset(h), v \in vdom(m) \cup V_m \} \rangle \mid h \in H \}$  其中: H 是主范畴的集合, M 是特征名的集合, V 是特征值的集合,  $V_m$  是对应特征名  $m (m \in M)$  的特征变量的集合,

函数  $hset : H \rightarrow P(M)$ ，即给每个主范畴分配一个特征值范围

$vdom : M \rightarrow 2^V$

例如:  $CAT = \{ \langle c-fas, \{ \langle lang, c \rangle, \langle op, illoc \rangle \} \rangle, \langle nom, \{ \langle lang, c \rangle, \langle op, pred \rangle, \langle sem, bottom \rangle \} \rangle, \dots \}$

定义 3: LEX 是词汇的集合，例如:  $LEX = \{ \text{我, 学习, 书, 大, 想} \dots \}$

定义 4:  $RULE = \{ c \rightarrow w \text{ 或 } c \rightarrow l \mid c \in CAT, w \in CAT^+, l \in LEX \}$

例如:  $RULE = \{ \langle c-fas, \{ \langle lang, c \rangle, \langle op, illoc \rangle \} \rangle \rightarrow \langle illoc, \{ \langle lang, c \rangle, \langle sem, ass \rangle \} \rangle, \langle clause, \{ \langle lang, c \rangle, \langle op, pred \rangle, \dots \} \rangle ; \dots \}$

建立汉语的 C-FAS 结构, 首先要根据要表达的句子语义信息来确定选取范畴及相应的特征和特征值。C\_FAS 结构主要描述以下信息:

1) 函词一变元关系: 包括句子的谓语动词与主、宾语之间的关系, 以及句子内部短语中介词与介词宾语、限定词与被限定词之间的关系。这些关系通过特征 *op* 及其值来描述。如 *pred* 表示谓语动词与主、宾语之间的关系, *det* 表示限定词(如数量词)与被限定词之间的关系。这些关系中的变元在关系中的作用, 以特征 *role* 及其值来刻画。

2) 句型信息: 句型信息由范畴 *illoc* 的 *sem* 特征表示。它的取值用来区分一般陈述句、疑问句、祈使句以及上节中所述及的其他特殊句型。之所以将这些句型加以区分标识, 是因为汉语句法复杂多变, 不同的句型的语义特点和内部语义关系差别很大。有必要区分以便于从 FAS 结构向其他层次的结构进行转化。

3) 其他信息: 除上述两方面的信息之外, 还要描述函词变元关系中各成分之间的表层标量顺序。它通过特征 *them* 来表示。此外特征 *conf* 用来刻画变元之间的语义关系。如施事和受事关系, 整体和个体关系等等。各成分的语义信息由特征 *sem* 来描述。

正是由于加入了句子语义方面的信息, C-FAS 结构与以 GPSG 描述的句法结构之间不是简单地一一对应。虽然如此, 在一些局部结构中, 存在着 C-FAS 子结构与 GPSG 句法结构相对应的情况。更多的情况下 C-FAS 结构与 GPSG 结构之间并不对应。由于 C-FAS 结构描述了许多 GPSG 不包含的语义信息, 因而有一些 C-FAS 子结构, 没有相应的 GPSG 结构与之对应。如描述句型信息的 C-FAS 子结构 *illoc* 就无相应的 GPSG 结构。C-FAS 结构和 GPSG 结构之间的差异增加了两者之间转换的困难。

对于词汇的语义描述, 通常有两种方式, 其一为属性描述法, 其二为语义分类法。笔者采用一个树形语义分类体系来进行语义描述。本体系共分 16 大类, 即人物, 品德, 心理活动, 仪容, 处境, 交往, 争斗, 语言, 信息, 物质, 运动, 时间, 空间, 数量, 程度, 性质等。各大类又进一步划分为中类和小类。

## 4 实现策略

从 GPSG 句法结构转换到 C-FAS 结构, 重点是要从句法结构中获取所需的各种函词和变元关系。并且收集必要的语义信息, 完成对句子 C-FAS 结构中各级子结构的描述。显然, 通常 GPSG 句法结构并不能提供所需的全部语义信息, 如句型信息、主题变元结构信息以及词汇的语义信息等。为解决这个问题, 考虑从两方面入手。其一, 适当修改、扩充 GPSG 的功能, 使之能提供部分所需信息。如扩充控制特征集, 引进 *S\_agr*, *Do\_agr*, *Io\_agr*, 用来分别表示主语、直接宾语、间接宾语; 引进特征 *pos* (*position*), 用以表示各词在句中实际出现的顺序位置等。其二, 增强词典信息, 使词条不仅包括词形、拼音、词性和 GPSG 范畴, 还要包括语义、配制等方面的信息等等。

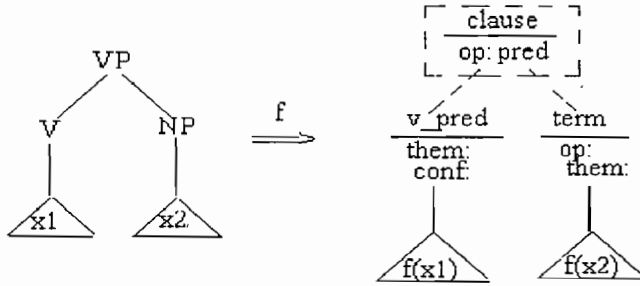
采用以下实现策略:

- 1) 从 GPSG 句法结构自顶向下转换生成 C-FAS 结构树。
- 2) 同时收集所需的语义信息完成部分语义描述(如主题结构信息 *them*)
- 3) 从词典中获得步骤 1) 2) 所不能提供的语义信息, 如词的语义信息等。

其中步骤 1) 是重点。它主要由一组转换规则制导实施。转换规则的左部表示输入的

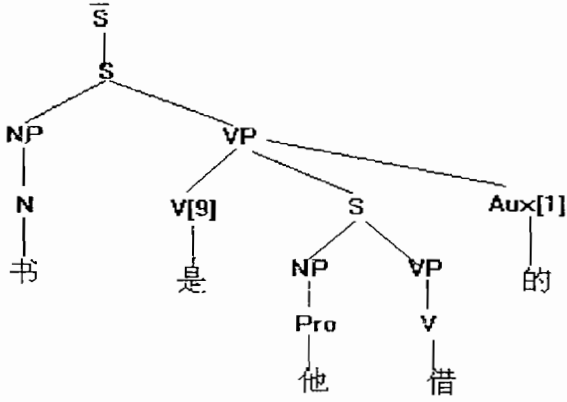
GPSG 句法子结构模式，右部表示产生输出的 C-FAS 结构子树。例如：

例1：

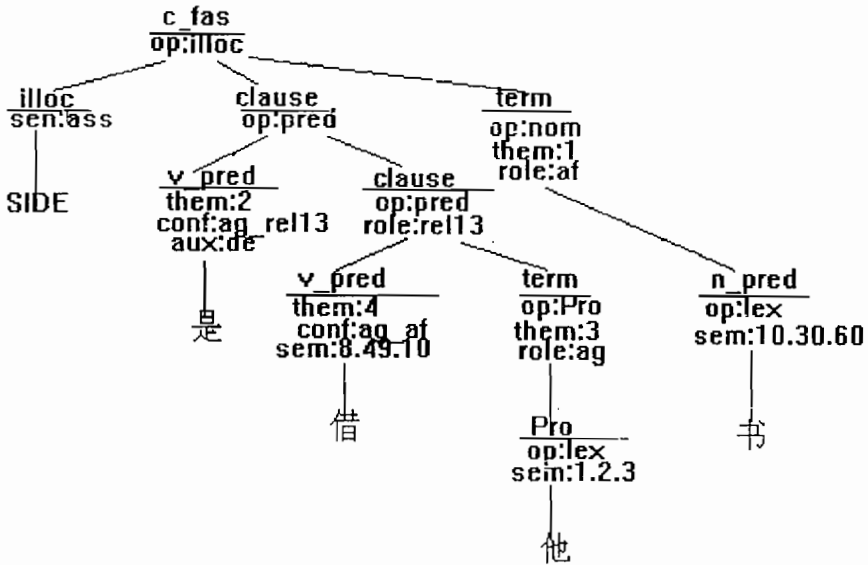


(注：例 1 中产生两棵子树，做为上一层 C-FAS 结点 (clause) 的子树)

例 2：例句：“书是他借的。”，此句的 GPSG 句法结构为：



经转换得到的C-FAS 结构为：



(为了方便起见,以上各例中的范畴特征都是以简略形式表示的。其中 op 表示操作,其值 illoc 表示非内嵌子句, pred 表示谓词操作, nom 表示名词短语;特征 them 的值表示各词汇的表层位置, conf 表示配制信息,等等。)

## 5 结 论

用 C-FAS 结构分析和描述汉语句子的语义,放松了 GPSG 中语义结构与句法结构严格对应的限制。这有助于克服在机译系统中从 GPSG 结构映射到内涵逻辑表达式所遇到的困难。同时,对于一个进行多层转换的机译系统来说, C-FAS 结构起了一个承上启下的作用。它是源语言到目标语言单纯句法层转换的深化,也为更进一步实现篇章内容的概念转换打下了基础。用 C-FAS 结构描述汉语句子语言是一个有意义的尝试。当然本模型有许多不足之处,有待于进一步深化和改进。

## 参 考 文 献

- [ 1 ] G. Gazdar, G. Pullum, 《 Generalized Phrase Structure Grammar 》  
Basil Blackwell, Oxford, U. K. 1985
- [ 2 ] 陈建民,《现代汉语句型论》,语文出版社 1986年11月
- [ 3 ] 吴竞存,梁伯枢,《现代汉语句法结构与分析》 语文出版社 1992年6月
- [ 4 ] 陈宗明,《汉语逻辑论》,人民出版社 1993年11月
- [ 5 ] 陆致极,关于广义短语结构语法,《国外语言学》,1989年4月
- [ 6 ] 金顺德,广义短语结构语法中的语义解释问题,《国外语言学》,1990年4月
- [ 7 ] 姚天坊,盛焕焯,何厚存, FAS ----自然语言中一种句子语义的表达结构,  
《1992年全国机器翻译学术会议论文集》
- [ 8 ] Stephan Busemann, Crista Hauenschild, From FAS Representations to GPSG  
Structure, KIT—REPORT 74, TUB, JULY 1989