从汉语句子中提取逻辑函子的一种方法*

靳光瑾 陆汝占

(上海交通大学计算机科学与工程系 上海 200030)
(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘要 文章介绍一种从汉语语句中提取逻辑函子的方法.该方法基于汉语配价理论,用组合逻辑方法将动词结构表示成逻辑函子,解决了多个NP竞争一个论元位置的问题.该方法体现了如何计算汉语语句语义的思想.

关键词 模型,函子,配价理论,移位.

中图法分类号 TP18

计算机理解自然语言首要的是理解逻辑语义,其次才涉及文化背景知识.逻辑语义基本上不因语种不同而有本质上的差异,因而使不同语种之间的翻译成为可能.计算逻辑语义的方法很多,其中最普遍、最成熟的是集合论-模型论.基于模型论的语义计算过程主要有两部分:(1) 如何将语句、篇章转换为高阶逻辑公式(转换规则与语言有关).表示形式及其转换是前提.下文将讨论一类汉语语句的转换形式.(2) 在给定模型下计算逻辑公式的语义,类同于一般逻辑系统模型解释过程.

Montague语义学给出了片断英语语句的模型论解释,是唯一的代数同构.[1,2]该理论成功的客观基础在于一条印欧语成立、而汉语不成立的规律:1句=1主×1谓.所以要想构造汉语语句到高阶逻辑式的转换,必须更多地注意非常型和含有省略、隐含的结构.[3]下文介绍一类非常型结构语句的函子的提取方法.

1 动词与函子

语句自动译成逻辑式应从汉语动词结构着手.本文提出的函子与扩展函子是介于句法和逻辑两者之间的中介形式.谓语动词连同支配成分构成了语句的核心,表现了句义的基本要素.动词结构对应了原子谓词式,再加上时态、模态算子就可以表示语态,从而表示了句子的基本逻辑含义.因此,动词结构逻辑形式是汉语语句语义计算框架的核心.

动词研究中直接可借鉴的是动词配价理论.句子中与动词搭配的有两部分,分别是必有支配成分和可有说明成分.一个动词的必有支配成分的数目是确定的,称为"价".通常有一价、二价,至多三价.[4]动词一般格式为:$V^n(x_1,\ldots,x_n)$,$1\leqslant n\leqslant 3$.论元项完全的动词格式称为函子,加入说明成分之后称为扩展函子.

例:昨天张三在会上批评李四.

在动词形式规范中,规定了函子的形式和语义类型.

动词"批评",函子 $V^3_{批评}(x,y)$,二价,论元 x,y,语法范畴 C 和语义类型说明 $YL,x:X,y:Y$,如下

x C:名词; X:(具体)人 / 机构 / 刊物

y C:名词/ 小句;Y:(对象)人 / 机构 /刊物|(抽象)事情 / 性能 / 意识

例句的函子 $V^3_{批评}$(张三,李四),扩展函子(昨天)↑(在会上)↑$V^3_{批评}$(张三,李四).寻常语句函子的抽象形式:一价 xV,二价 xVy,三价 $xVyz$.为了方便起见,写成 $V^1(x),V^2(x,y),V^3(x,y,z)$.

从语句提取函子的步骤为:

(1) 语句经句法分析,加注标记之后,从动词形式规范库中提取与谓语动词相应的动词格式及语义类型信息;

(2) 顺序提取动词前后名词性成分为候选论元;

(3) 候选论元与形式规范中规定的论元语义类型匹配合一;

---

* 本文研究得到国家自然科学基金资助.作者靳光瑾,女,1950年生,博士,主要研究领域为汉语计算语义.陆汝占,1940年生,教授,博士导师,主要研究领域为推理技术与定理证明,模型与软件集成,汉语语义计算.

本文通讯联系人:陆汝占,上海 200030,上海交通大学计算机科学与工程系

本文1997-03-26收到原稿,1997-06-16收到修改稿

(4) 匹配成功者代入论元位置，这个匹配代入过程可用 λ-演算形式化(略)[5]；

(5) 语句中缺省成分(句法成分和逻辑成分)需自动添补，使函子论元项完备(求解缺省将另文介绍)。

例 1：他父亲死了。　句法：$NP1V$(了)，　函子：$V_{死}$(他父亲)

例 2：他吃了一个苹果。　句法：$NP1V$(了)$NP2$，　函子：$V_{吃}$(他，一个苹果)

例 3：我抽阿诗玛香烟。　句法：$NP1V2NP2$。　函子：$V_{抽}$(我，阿诗玛香烟)

函子的论元项只有名词性词组(NP)和动词词组(VP)这两种形式，后者又递归到函子形式。因此，论元项主要取决于名词词组。寻常语句如上所述步骤，可得函子初始形式。

2　非常语句的多 NP 竞争论元问题

语句中 NP 能作为函子的论元，主要是靠语义类型匹配合一的策略来实现的。在一般情况下，在两个(或者两个以上)NP 中选取一个作论元，排斥其余的，如例 2，“他”与“苹果”被恰当地分列在两处。这个策略是广泛有效的。但是复杂情况会导致该策略失败。语句中出现多个 NP、争当一个论元是其中很重要的一种。最典型的是两个同现的名词性词语有相同的语义类型，在非除作状语、补语的可能性之后，要确定这两者的语义角色，选取一个为动词论元，安置另一个以合适的语义角色。这里的 NP 之间语义上不是完全排他性的，而是有某种相关联系的。因此解决竞争的结果是要将两个 NP 复合成一个成分，让这个复合成分充当一个论元。以下列举 5 种比较典型的句型，例 5、6 中竞争施事论元，例 4、7、8 竞争受事论元。

例 4：苹果他吃了一个 / 例 5：他死了父亲 / 例 6：工厂塌了围墙 /

例 7：香烟我抽阿诗玛 / 例 8：皮鞋我买了一双黑色的

例 4：“苹果他吃了一个”，

句法上可认为是由语句“他吃了一个苹果”前移“苹果”的结果。“苹果”与“一个”语义类型相同，但所指范围不同，分别是个体类和单个个体。例 5 他死了父亲，是由例 1 后移复合 NP“他父亲”中的“父亲”的结果。移位部分和留有部分中有一个且只有一个为论元。究竟哪个是论元，标准在于“语义所指范围”。移位前的语句非常单一规范，呈主-谓结构，或者主-谓-宾结构，但是移位之后错综复杂。移位后复合 $NP=NP1+NP2$ 中 $NP1$ 与 $NP2$ 语义所指范围不同，反映在逻辑语义上是什么？对于计算机来说，就是给出一个汉语语句字符串，借助于词法分析和句法结构分析之后，标出的符号串为：(1) 名 1＋名 2＋V^2＋名 3，或(2) 名 1＋V^1＋名 2。这两种形式能否划一化？本质上是一个论元竞争问题，第 1 种形式中，3 个名词性成分竞争两个论元位置，其实名 2 从语义类型、句型诸方面来看，可基本确定为一个论元。实际上只有名 1 与名 3 竞争一个论元地位。第 2 种形式中，也是名 1 与名 2 竞争一个论元地位。两种形式归结起来就是前面说的 $NP1$ 与 $NP2$ 谁是论元问题。结果是将两者粘合之后“合占”一个论元。因此，设想将形式(1)中的名 1 与名 3，形式(2)中的名 1 与名 2 两个句法上不同的结构，逻辑式 $Form$ 化归为一种，即

$$Form(\text{名 1}+(\text{名 3} / \text{名 2}))\Rightarrow(Form(\text{名 3})/ Form(\text{名 2}))\odot Form(\text{名 1})$$

竞争的两者“粘合”(倒置)成一个表达式。作为函子的论元，这正是语句自动转换为函子的关键所在。定义这个粘合运算，按 $NP1$ 与 $NP2$ 的情况分别定义。

3　粘合运算

偏正结构的复合名词组 NP 中的 $NP1$，$NP2$ 及其它们之间的关系主要有以下几种：

(1) 人称代词＋名词：“他父亲”表示领属关系。

(2) 名词＋名词：

(2.1) “工厂围墙”表示领属关系；

(2.2) “阿诗玛香烟”表示种类关系。

(3) 数词＋量词＋名词：“三只苹果”，表示数量与物的限制关系。

(1)～(2.1)可解释为“$NP1$ 的 $NP2$”，(2.2)为“$NP2$ 中的 $NP1$”，(3)本身说明一个集合(3 只苹果)，它是整个“苹果”集合中的一个子集。也可以解释为“$NP2$ 中的 $NP1$”。

逻辑表示：领属性关系有固有的(单值的：“父亲”“头”，双值的：“手”)和非固有的关系(多值的“朋友”“同学”)。固有关系用函数表示。$father(x)$，“x 父亲”，他父亲：$father$(他)；$hands$(他)：“他的手”，值是一个序偶(左手，右手)，“他的左手”表示成 $Lhands$(他)。非固有关系表示成“关系集”。“他朋友”$Friend$(他)；“小王是他朋友”表示成“小王∈$Friend$(他)，”或者说[$Friend$(他)(小王)]取值 1。采用叠置符号就有：$father\cdot$他，$Friend\cdot$他。于是它们的逻辑项就是 $F\cdot a$(叠置)。

$[_{代}代 + 名]\Rightarrow F \cdot a$,其中 a 为代词的指称表示.

(2.1)与上面非固有关系类似."围墙"都表示成关系集 *Encwall*,"工厂围墙":*Encwall* ○ *Factory*,其中○表示 of 关系.照理说,普通名词:围墙,本身应该表示为一个集合 *Set* 或谓词 Q,"y 为围墙"表示为 $y\in Set$ 或者 $Q(y)$,标志了 y 是围墙这样的特征.

$$工厂围墙 \Rightarrow 工厂+围墙 \Rightarrow Encwall \bigcirc Factory$$

语句成分自然序为工厂+围墙,○运算要颠倒,下面给出一种组合算子.

$$(由成分的对应表示)\Rightarrow (\lambda_n \lambda_A \cdot A \bigcirc U)\cdot Factory \cdot Encwall$$

$$(\lambda 代换规则)\Rightarrow Encwall \bigcirc Factory$$

两个集合的领属关系 $A \bigcirc B$ 定义了一集合,

$$A \bigcirc B \rightarrow \{y \mid y\in A \text{ and } \exists_x (x\in B \text{ and } Own(y,x))\}$$

所以"工厂塌了围墙",表示成函子

$$V_{塌}(Encwall \bigcirc Factory)$$

展开成逻辑式

$$\exists_{x,y}\{y\in Encwall \text{ and } x\in Factory \text{ and } Own(y,x) \text{ and } V_{塌}(y)\}$$

$$[_{属算子}名1 + 名2]\Rightarrow A_{属算子}\cdot Set1\cdot Set2 \Rightarrow Set2 \bigcirc Set1$$

(2.2)表示种与类的关系,逻辑上种类=类+种差.种类是对象类中的一个子集,种用来限制说明这个子集如何区别于类的.一般种、类都用普通名词("阿诗玛"尽管是品牌,但并不指称常个体,因此也同普通名词一样的功能),这两个普通名词分别对应两个集合,取两个集合的交,就是这个种类所指的对象子集.

这时"*NP*2 中的 *NP*1"的逻辑语义是 $A \otimes B$(两集合的交集).

$$"阿诗玛香烟"\Rightarrow 阿诗玛+_{中}香烟 \Rightarrow AShima \otimes Cigarette$$

对应于"*NP*2 中的 *NP*1"关系,定义一个组合算子,能交换次序,且表示"中"的关系.

$$(成分对应)\Rightarrow [\lambda_A \cdot \lambda_B \cdot A \otimes B]\cdot AShima \otimes Cigarette$$

$$[_{种}名1 + 名2]\Rightarrow A_{种算子}\cdot Set1\cdot Set2 \Rightarrow Set1 \otimes Set2$$

最困难的是复合 *NP* 结构数·量·名结构的逻辑表示,例"一本书""一双鞋".

名,普通名词,指人或物,对应一个对象集,所以"书"对应了集合 *Books*.

量,指计量单位,这里仅指个体量词."书"按"本"计,"鞋、袜"按"双"计.个体量词抽象为两类,个类:个、本、只,用 $[X]$ 表示.双类:双、对、付,用 $[X * X]$ 表示.这儿括号内的 X 是一种形式变元,可以换名,写成 $[Y]$,$[T]$ 都一样,但不能代换以任何具体的项.此处的 $[X]$ 仅仅是约束的形式项,不出现任何自由变元,可以用一个组合逻辑算子表示.它该是什么样子呢?就看它应具什么功能."量":$[X]$ 是指它后边一定要跟一个普通名词,它前边一定还有数量(子集的元素个数、基).它是一个前后衔接的联系.虽然作为一个项时,后边的普通名词会缺省,但句内一定会有,就是所说的拆开-移位.

因此"量":$[X]=\lambda_D \cdot \lambda_{A_{数量算子}}\{X \mid X\subseteq A \text{ and } |X|=D\}$.此式意为,$[X]$ 表示一个抽象的待定的子集,它是(后边会出现的)某个集合 A 的子集,且这个子集的基为 D.

"数",指数量多少.复合 *NP* 指称对象集中的一个子集,数量表示这个子集的基.例如:"二本(个、只)"句法上="二"+"本".

$$逻辑项:(\lambda_D \cdot \lambda_A\{X \mid X\subseteq A \text{ and } |X|=D))\cdot 2, \quad 化简:\Rightarrow \lambda_A\{X \mid X\subseteq A \text{ and } |X|=2\}$$

意指是一个抽象子集,它的基为 2,即元素个数为 2,待定是要看后边跟什么普通名词.注意这个逻辑项中"2"出现的次序在量 $[X]$ 后边,跟语序先"数"后"量"不一致.因此,前面还应该有一个"颠倒"功能的组合算子.

$$(\lambda_D \cdot \lambda_C \cdot C \cdot D)\cdot 2\cdot [X] \quad (化简)\Rightarrow [X]\cdot 2 \Rightarrow 上述逻辑项$$

$$"二本书" 句法上="二"+"本"+"书"="二本"+"书"$$

$$逻辑项:(\lambda_A\{X \mid X\subseteq A \text{ and } |X|=2\})\cdot Book \Rightarrow \{X \mid X\subseteq Book \text{ and } |X|=2\}$$

表示所有书集合中的一个子集,它有二本,符合"二本书"的原意.

$$"一双鞋"句法上="一双"+"鞋"$$

$$逻辑项:(\lambda_A\{X * X \mid X\subseteq A \text{ and } |X * X|=1\})\cdot Shoe$$

$$化简:\{X * X \mid X\subseteq Shoe, \text{and } |X * X|=1\}$$

表示一个序偶集,其中只有一个序偶,序偶中的元素都为鞋,表示了"一双鞋"的原意.

$$[_{数}数 + 量 + 名] \Rightarrow A_{数算子} \cdot n \cdot [X] \cdot Set \Rightarrow n \cdot [x]\copyright\ Set \Rightarrow SubSet$$

叠置　用组合逻辑方法定义的最大好处在于允许复合(叠置)运算.

例："三个兄弟中的二个"

$$\text{逻辑项：二个} \cdot (\text{三个} \cdot \text{兄弟})$$
$$\Rightarrow \lambda_A\{X \mid X \subseteq A \text{ and } |X|=2\} \cdot \lambda_A\{X \mid X \subseteq A \text{ and } |X|=3\} \cdot Brother$$
$$\Rightarrow \lambda_A\{X \mid X \subseteq A \text{ and } |X|=2\} \cdot \{X \mid X \subseteq Brother \text{ and } |X|=3\}$$

后面"三个兄弟"是确定的一个集合，记为 Bro.

$$\Rightarrow \{X \mid X \subseteq Bro \text{ and } |X|=2\}$$

的确表示了 3 个兄弟中的二个，这个集合也是"兄弟"中的某个特定的子集，元素个数为 2.

扩展复合 NP 结构：数＋量＋定＋名. 例如：一双黑色的鞋.

定，指修饰性定语，用来指称对象种类，用谓词表示，"X 为黑色"，记为 $black(x)$，世界上所有黑色的东西组成的集合 $Black$，与"鞋"集合 $Shoe$ 的交集，

$$[_{定}定 + 名] = A_{定算子} \cdot Set1 \cdot Set2 \Rightarrow Set1 \otimes Set2$$
"一双黑色的鞋"句法上＝"一双"＋"黑色的鞋"
$$\text{逻辑项：} \{X * X \mid X \subseteq (Black \otimes Shoe) \text{and } |X * X|=1\}$$

综上所述，逻辑式：

$$Form(\text{名 } 1 + \text{名 } 2 + V + \text{名 } 3) \Rightarrow V^2(\text{名 } 2, Form(\text{名 } 1 + \text{名 } 3)) \Rightarrow V^2(\text{名 } 2, Form(\text{名 } 3) \odot Form(\text{名 } 1))$$
$$Form(\text{名 } 1 + V + \text{名 } 2) \Rightarrow V^1(Form(\text{名 } 1 + \text{名})) \Rightarrow V(Form(\text{名 } 2) \odot Form(\text{名 } 1))$$

其中粘合运算如下，当 $v=Form(N1)$，$u=Form(N2)$ 时：

$u \odot v = v \cdot u$	叠置关系	领属	$N1$：人称代词
$u \bigcirc v$	Own 关系	领属	$N1$：名
$u \otimes v$	交集，"与"	种类	$N1$："类"
$u \copyright v$	子集	数量名	$N1$：数量

上述定义规则和化简规则，都是可由模块实现的. 依据词法、句法特征标记后，可将语句中多个 NP 粘合成一个复合 NP，从而确定函子的论元项. 这是函子自动抽取中的一个关键步骤.

$$Form \text{ 例 } 4 \Rightarrow V_{吃}(他, Form(苹果 + 一只)) \Rightarrow V_{吃}(他, (1 \cdot [x])\copyright Apple)$$
$$Form \text{ 例 } 5 \Rightarrow V_{死}(Form(他 + 父亲)) \Rightarrow V_{死}(Father \cdot 他)$$
$$Form \text{ 例 } 6 \Rightarrow V_{围}(Form(工厂 + 围墙)) \Rightarrow V_{围}(Encwall \bigcirc Factory)$$
$$Form \text{ 例 } 7 \Rightarrow V_{抽}(我, Form(香烟 + 阿诗玛)) \Rightarrow V_{抽}(我, Ashima \otimes Cigarette)$$
$$Form \text{ 例 } 8 \Rightarrow V_{买}(我, Form(鞋 + 一双黑色的)) \Rightarrow V_{买}(我, (1 \cdot [x * x])\copyright (Black \otimes Factory))$$

参考文献

1 Dowty D *et al*. Introduction to Montague Semantics. D Reidel, Pub., 1981

2 Montague R. The proper treatment of quantification in ordinary English. Formal Philosophy, Thomason R H, 1974

3 沈阳. 现代汉语空语类研究. 济南：山东教育出版社，1994
(Shen Yang. Empty Category in Modern Chinese. Jinan: Shandong Education Publishing House, 1994)

4 胡裕树，范晓. 动词研究. 郑州：河南教育出版社，1995
(Hu Yu-shu, Fan Xiao. On Verbs. Zhengzhou: He'nan Education Publishing House, 1995)

5 朱一清. λ-演算的语法和语义. 南京：南京大学出版社，1992
(Zhu Yi-qing. The Syntax and Semantics of λ-Calculus. Nanjing: The Publishing House of Nanjing University, 1992)

A Method for Extracting Logical Functors from Chinese Sentences

JIN Guang-jin　LU Ru-zhan

(*Department of Computer Science and Engineering　Shanghai Jiaotong University　Shanghai*　200030)
(*State Key Laboratory of Novel Software Technology　Nanjing University　Nanjing*　210093)

Abstract　In this paper, the authors present a method based on valent grammar and combinatory logic for extracting logical functors from Chinese sentences. The method represents a verb structure as a logical functor, and solves the problem of several NP competing for one position of argument. The proposed method enbodies the thoughts of computing the meaning of Chinese sentences.

Key words　Model, functor, valent grammar, movement.