组合原则和自然语言虚化成分

邹 崇 理[1,2]

(1.四川师范大学 逻辑与信息研究所,成都 610066;2.中国社会科学院 哲学所,北京 100732)

摘要:计算机人工智能时代最重要的任务之一是自然语言的信息处理,逻辑语义学则是其基础理论,而组合原则又是逻辑语义学的基本原则,表现为部分决定整体的函项思想。自然语言的虚化成分是自然语言复合表达式中对整体意义不起作用的那些部分,自然语言违反组合原则的情况表现为句法和语义的不对应,意味着决定整体意义的"部分"这个概念应该受到限制,组合原则的经典表述在自然语言的某些场合受到挑战。就自然语言的某些语义领域而言,限制性的组合原则概念是关于组合原则具体精准的表述。

关键词:逻辑语义学;组合原则;自然语言;虚化成分

中图分类号:B815.3;O141 **文献标志码**:A **文章编号**:1000-5315(2017)01-0005-05

逻辑学是人文社会科学和自然科学共同的基础学科。1974 年,联合国教科文组织规定的七大基础学科依次为数学、逻辑学、天文学和天体物理学、地球科学和空间科学、物理学、化学、生命科学。由此可见,逻辑学在人类整个知识结构中的基础地位。20 世纪中叶以来,计算机科学技术的迅猛发展,导致席卷全球的信息革命,而自然语言是信息的重要载体之一,信息革命离不开自然语言的计算机处理。逻辑作为人类知识结构的基础,同样在自然语言的信息处理领域内发挥巨大作用。

自然语言的计算机信息处理过程是:首先,把需要研究的语言学问题用数学或逻辑的形式严密而规整地表示出来;其次,把这种严密而规整的表述表示成算法,建立各种自然语言处理系统;第三,对自然语言处理系统进行评测,不断改进质量和性能。逻辑语义学关于自然语言的研究主要对第一步骤发生作用,逻辑语义学是自然语言信息处理的基础理论学科,前者为后者提供了诸多重要的操作工具和指导思想。

从逻辑语义学角度对自然语言进行形式化研究,组合原则是其重要的方法论,那么怎样理解组合原则?组合原则是什么?

一 什么是组合原则

组合原则是逻辑语义学的基本原则。组合原则直观表述为:如果表达式 E 依据某个句法规则由部分 E_1 和 E_2 所构成,则 E 的语义 M(E)是依据某个语义规则把 E_1 的语义 $M(E_1)$和 E_2 的语义 $M(E_2)$合并起来而获得的。举例来说,表达式"伟大祖国"的语义是由形容词"伟大"的语义限制名词"祖国"的语义而获得。表达式"戴眼镜的女孩"的语义是以由定语从句"某某戴眼镜"的语义和名词"女孩"的语义合并而成。

严格讲,组合原则意味:一个复合表达式的语义是由部分的语义贴合这些部分的句法运算的意义所构成的函项。组合原则的数学定义深刻揭示了这样的特征[1]526。

令 $A=\langle A,F\rangle$ 和 $B=\langle B,G\rangle$ 都是代数,映射 $h:A\to B$ 是同态的,当且仅当,存在一个映射 $h':F\to G$ 使

收稿日期:2016-02-26

基金项目:国家社科基金重大招标项目"自然语言信息处理的逻辑语义学研究"(10&ZD073)。

作者简介:邹崇理(1953—),男,四川成都人,四川师范大学特聘教授、逻辑与信息研究所学术委员会主席,中国社会科学院博士生导师,中国逻辑学会会长,主要研究自然语言逻辑。

得对所有 $f \in F$ 和所有 $a_1,\cdots,a_n \in A$ 都有：

$h(f(a_1,\cdots,a_n)) = h'(f)(h(a_1),\cdots,h(a_n))$

在自然语言领域，A 是句法代数，B 是语义代数，h 就是从句法生成到语义组合的意义指派。A 是句法表达式的集合，B 是语义值的集合，F 是句法算子的集合，G 是语义算子的集合。f 是 F 中的某个算子，a_1，…，a_n 是 A 中的 n 个表达式，$h'(f)$ 是 G 中对应 f 的语义算子，$h(a_1),\cdots,h(a_n)$ 是 B 中对应 $a_1,\cdots,a_n$ 的 n 个语义值。复合表达式 $f(a_1,\cdots,a_n)$ 是 f 对 $a_1,\cdots,a_n$ 进行句法生成的结果，其语义 $h(f(a_1,\cdots,a_n))$ 就是语义算子 $h'(f)$ 对 n 个部分语义 $h(a_1),\cdots,h(a_n)$ 进行运算的结果，是语义算子贴合部分语义进行运算的函项。

例子解读：令句法代数 A 的论域 A＝{张三，李四，散步，学习，张三散步}，语义代数 B 的论域 B＝{a，b，{b}，{a，b}，1，0}。再令 f(张三，散步)＝张三散步，h(张三)＝a，h(散步)＝{a，b}，$h'(f)$＝g。对此进行语义指派得：h(f(张三，散步))＝$h'(f)$(h(张三)，h(散步))＝g(a，{a，b})＝1，当且仅当 a∈{a，b}。

可以看出组合原则具有两个重要特征：(1)复合表达式语义组合“$h'(f)(h(a_1),\cdots,h(a_n))$”的根源依据是复合表达式的句法生成“$f(a_1,\cdots,a_n)$”，这就是句法和语义对应的思想；(2)复合表达式的语义不仅依靠其部分的语义“$h(a_1),\cdots,h(a_n)$”，还取决于合并这些部分的句法生成的意义“$h'(f)$”。

组合原则是现代逻辑的基石，在构造逻辑系统中起到方法论的作用。组合原则要求逻辑系统中每个句法(语形)形成规则必须对应一个语义解释规则。命题逻辑严格遵循了意义的组合原则，令 // // 为意义指派函项 h，则有：

Syn1.原子公式 p_1，p_2，∈Form；

Syn2.若 φ∈Form，则(¬φ)∈Form；

Syn3.若 φ，ψ∈Form，则(φ→ψ)∈Form。

Sem1.‖pi‖ ∈{0，1}；

Sem2.‖(¬φ)‖＝1 当且仅当‖φ‖＝0；

Sem3.‖(φ→ψ)‖＝1 当且仅当‖φ‖＝0 或‖ψ‖＝1。

句法规则 Syn1 对应语义规则 Sem1，Syn2 对应 Sem2，Syn3 对应 Sem3。显然，这里复合表达式的语义依据其部分表达式的语义，复合表达式的所指是部分表达式所指的函项。如‖(φ→ψ)‖＝‖→(φ，ψ)‖＝$h'(\rightarrow)$(‖φ‖，‖ψ‖)＝1，当且仅当‖φ‖＝0 或‖ψ‖＝1。

自 20 世纪 70 年代初开始，现代逻辑的方法扩展延伸到自然语言的研究领域，形成了以蒙太格语法(Montague Grammar)[2]247-270 为首的逻辑语义学群体，组合原则自然也成为逻辑语义学的灵魂。

蒙太格语法是强调组合原则的逻辑语义学理论。在其构造的三个英语部分语句系统那里，句法和语义处处严格对应。以 PTQ 系统为例，17 条句法规则对应 17 条语义翻译规则[2]247-270。每条翻译规则体现出：复合表达式的翻译是其部分表达式翻译的函项。句子、动词短语和名词短语三类合取复合表达式的句法规则及其翻译规则如下：

Syn1.若 φ，ψ∈P_t，则 F_8(φ，ψ)＝φandψ∈P_t；

Syn2.若 δ，γ∈P_{IV}，F_8(δ，γ)＝δ andγ∈P_{IV}；

Syn3.若 α，β∈P_T，F_9(α，β)＝α or β∈P_T。

Tra1.若 φ，ψ 分别翻译成 φ′，ψ′，则 φandψ 翻译成[φ′∧ψ′]；

Tra2.若 δ，φγ 别翻译成 δ′，γ′，则 δ and γ 翻译成 λx[δ′(x) ∧γ′(x)]；

Tra3.若 α，β 分别翻译成 α′，β′，则 α or β 翻译成 λP[α′(P)∨β′(P)]。

翻译起意义指派函项的作用。令 T 是翻译函项，拿 Tra2.来说，T(δ and γ)＝T(and(δ，γ))＝h'(and)(T(δ)，T(γ))＝λx[δ′(x)∧γ′(x)]。复合表达式的翻译依赖部分表达式的翻译。

组合原则的作用还体现在更多的领域内。

在计算机科学那里，连接许多通信处理器的大网络技术发展很快，人们特别关注超大系统的行为。在有关研究中，组合原则起到很大的作用：牵涉整个系统行为的证明应该是各个处理器的证明的函项。这方面的介绍参见文献。

组合原则在形式翻译领域作用更大。为了考察逻辑系统之间的关系，比较表达力的大小以及获得相对的协调性，人们往往设立符合组合原则的翻译程序。最著名的例子是 Gödel 把直觉主义逻辑转换成模态逻辑的翻译。在直觉主义逻辑那里，联接词具有一种构造性解释，如 φ→ψ 被解释成：给定 φ 的证明，据此构造 ψ 的证明。令 Tr 为翻译函项，翻译程序定义为：

a. Tr(p) ＝ □p　对原子命题 p

b. Tr(φ∨ ψ) ＝ Tr(φ) ∨ Tr(ψ)

c. Tr(φ∧ ψ) ＝ Tr(φ) ∧ Tr(ψ)

d. Tr(φ→ψ) ＝ □p(Tr(φ) →Tr(ψ))

复合表达式的翻译，依据部分表达式的翻译来确定。Gödel 的翻译是一种组合翻译，逻辑系统之间大量的组合翻译可以参见 Epstein 的著述。

在不同自然语言之间也可以实行组合性的机械翻译。把一种自然语言翻译成另一种语言，要求保留源语言的意义，机器翻译的研究项目“Rosetta”企图基于翻译的组合原则来达到这一目标，两个表达式能够互译仅当它们的部分能够互译。

组合原则在理论和应用方面尽管有很大的价值和作用，但是也有其局限性。在 Pelletier 看来，下列句子对组合原则构成挑战：

Every linguist knows two languages.

The philosophers lifted the piano.

第一个句子因为两个不同的语言学家所熟知的两种语言是否彼此相同而产生歧义，第二句子不能确定是每个哲学家分别抬这台钢琴，还是哲学家们共同抬这台钢琴。两句的歧义既不来源于词汇，也不是由句法生成所引起。两句的情况都是：由相同部分的语义和相同的句法运算意义不能唯一确定整体表达式的语义。

组合原则是一个纯理论的数学概念，而具有相当灵活性和复杂性的自然语言中出现不遵循组合原则的例子在所难免。这些例子由自然语言的语用因素、歧义现象和句法语义的不对应所引起。本文所关注的自然语言虚化成分现象，涉及句法和语义的不对应，而并非遵守组合原则。即有的自然语言表达式，其句法构成是 f(A，B，C)＝ABC。按照组合原则粗略看，其整体语义依赖 A 的语义、B 的语义和 C 的语义。这里，由于 B 是虚化成分，其整体语义仅仅依赖 A 的语义和 C 的语义。那么，什么是虚化成分？

二　自然语言虚化成分

我们注意到，在自然语言中，有那样一些语言表达式，在句法上把其左边具有语义所指的表达式跟其右边具有语义所指的表达式隔开，而自身在整个表达式的语义解释中没有所指物。如汉语的结构助词，英语不定式中的小品词 to 等，我们称之为虚化成分。简言之，虚化成分是复合表达式中那些不起语义作用的句法成分。

我们以汉语结构助词“的”为代表来看学术界对此的处理情况。刘壮虎在《逻辑学增刊》上构造的复合谓词逻辑系统中，把诸如“大的蚂蚁”之类短语的逻辑语义分析看作是“大”对应的谓词对“蚂蚁”对应的谓词进行复合的结果，结构助词“的”在这里没有对应的逻辑词项。张秋成在专著《类型逻辑语法研究》[3]146 中的推演是：

美丽　　的　　姑娘

Lex　　Lex　　Lex

n/n：美丽　(n/n)\(n/n)：lM.M　n：姑娘

\E

n/n：lM.M(美丽)

/E

n：(lM.M(美丽))(姑娘)

λ－转换

n：美丽(姑娘)

最后结果表明，表达式“美丽的姑娘”的语义表达“美丽(姑娘)”中没有“的”所对应的语义表达。而王欣在她的博士论文《类型逻辑语法与现代汉语“是”和“的”》中给出的推演为：

a. 中式　家具　　b. 中式的　家具

n/n　n　　n/n　n

n　　n

她直接指出，在 a 和 b 里“中式”和“中式的”的范畴都是 n/n，加不加“的”意思都相同[4]157。也就是说，这个“的”对语言分析来说是多余的成分。

通常对汉语的动词短语表达式“飞快地跑步”是这样处理的：令“飞快”的逻辑式为 α，“跑步”的逻辑式是 γ，“飞快地跑步”的逻辑表达就是 α(γ)。从语义角度看，这里“地”是没有所指的。况且，在英语中，“飞快地跑步”对应“run rapidly”，“漂亮的女孩”对应“pretty girl”，英语上述短语的语义只需要两个成分起作用，这里汉语表达式中的“地”和“的”均可以看作是没有语义所指的虚化成分。

我们从类型逻辑语法角度审视，首先确立包含下列词条的词库：

飞快⇒np\s)/(np\s)：α　地⇒B：Ø(空逻辑式)跑步⇒np\s：γ

然后可以看到从□飞快地跑步推出 np\s：α(γ)”的树模式图示：

飞快　地　跑步

(np\s)/(np\s)：α　B：Ø　np\s：γ

删去虚化成分

np\s)/(np\s)：α　np\s：γ

np\s：α(γ)

上述删去虚化成分的推演需要提供新的推演工具。这就是：从 A：α，B：Ø，C：γ 推出 A：α，C：γ。对此抽象，去掉与范畴配对的 λ－项，再把虚化成分 B 及其

左右范畴 A 和 C 构成一个所谓“省略槽”的复合范畴[A{B}C],即得:[A{B}C]→(A·C),这就是新的推演工具。三元复合范畴[A{B}C]是删去虚化成分的起点,据此揭示包含虚化成分的表达式的语义特征,如“迅速地跑步”,“美丽的女孩”和“玩得高兴”,其中的“地”、“的”和“得”所属范畴就是起间隔虚化作用的 B。三元复合范畴的语义解释如下:

$v([A\{B\}C])=\{x\mid \exists yz[Sxyg(B)z \,\&\, y\in \| A \| \,\&\, z\in \| C \|]\}$

按照上述定义:Sxyg(B)z 意味:x 是 y、g(B)和 z 毗连的结果,具有语义所指的符号串 y 和 z 分别属于 A 和 C,而 g(B)指起虚化作用的符号串,B 是 A 和 C 之间的虚化范畴(对应的语义所指为空逻辑式)。于是有:

限制 0:$\forall B\forall x[x\sim g(B)\Rightarrow x\in v(B)]$

这里 x~g(B)的直观理解是:x 是 g(B)(起虚化作用的符号串)。限制 0 表明起虚化作用的符号串是‖B‖中的元素。

于是,我们提出基于[A{B}C]的范畴逻辑系统[5]370-381。其公理是:

公理 0:A→A

公理 1:A·B↔B·A

公理 2:[A{B}C]→(A·C)

公理 3:D·[A{B}C]→[(D·A){B}C]

公理 4:[A{B}C]·D→[(A·D){B}C]

公理 5:D·[A{B}C]→[A{B}(D·C)]

公理 6:[A{B}C]·D→[(A{B}(C·D)]

公理 7:[A{B}C]·[D{B}E]→[(A·D){B}(C·E)]

系统的规则有(Lambek 演算的 5 条推演规则):

$$\frac{A\cdot B\to C}{A\to C/B}\qquad \frac{A\cdot B\to C}{B\to A\backslash C}$$

$$\frac{A\to C/B}{A\cdot B\to C}\qquad \frac{B\to A\backslash C}{A\cdot B\to C}$$

$$\frac{A\to B\quad B\to C}{A\to C}$$

此外,系统还有两条独特的推演规则:

规则 6:
$$\frac{A\to B}{[A\{D\}C]\to[B\{D\}C]}$$

规则 7:
$$\frac{A\to B}{[D\{C\}A]\to[D\{C\}B]}$$

对构成其他复合范畴的算子,传承 Lambek 演算 L 系统的语义解释如下:

$v(A\cdot B)=\{x\mid \exists y\exists z[Rxyz \,\&\, y\in v(A) \,\&\, z\in v(B)]\}$

$v(C/B)=\{y\mid \forall x\forall z[Rxyz \,\&\, z\in v(B))\Rightarrow x\in v(C)]\}$

$v(A\backslash C)=\{z\mid \forall x\forall v[Rxvz \,\&\, v\in v(A))\Rightarrow x\in v(C)]\}$

按照惯例给出系统的框架语义,这是一个由三元可及关系 R 和四元可及关系 S 组成的混合框架。系统的语义特色在于下述框架限制:

限制 0:$\forall B\forall x[x\sim g(B)\Rightarrow x\in v(B)]$

限制 1:$\forall xyz[Rxyz\Rightarrow Rxzy]$

限制 2:$\forall xyzu[Sxyzu\Rightarrow Rxyu]$

限制 3:$\forall xyzuvw[Rxyz \,\&\, Szuwv\Rightarrow t[Sxtwv \,\&\, Rtyu]]$

限制 4:$\forall xyzuvw[Rxyz \,\&\, Syuvw\Rightarrow t[Sxtvw \,\&\, Rtuz]]$

限制 5:$\forall xyzuvw[Rxyz \,\&\, Szuvw\Rightarrow t[Sxuvt \,\&\, Rtyw]]$

限制 6:$\forall xyzuvw[Rxyz \,\&\, Syuvw\Rightarrow t[Sxuvt \,\&\, Rtwz]]$

限制 7:$\forall xyzuvwst[Rxyz \,\&\, Syuvw \,\&\, Szsvt\Rightarrow \exists ab[Sxavb \,\&\, Raus \,\&\, Rbwt]]$

依据上述提供的框架语义解释及其限制,可以证明系统的可靠性和完全性。可判定性证明也可按照惯例给出[5]370-381。

上文已强调,系统的最大特色就是公理 2:[A{B}C]→(A·C)。意味从 A:α,B:∅,C:γ 推出 A:α,C:γ。从句法角度看,复合表达式“ABC”的部分表达式是“A”、“B”和“C”。公理 2 的潜在显示为:“ABC”即“[A{B}C]”的整体语义就是“A·C”的语义,即“α(γ)”。而这仅仅取决于部分表达式“A”的语义“α”和部分表达式“C”的语义“γ”,复合表达式的语义并非如组合原则所要求的是由所有部分表达式的语义来决定。

在自然语言复合表达式中间的部分表达式是虚化成分的条件下,这时的组合原则就是受限的,其表述就是:复合表达式的语义是由除去作为那个虚化成分的部分的语义以外的其他部分的语义贴合这些部分的句法运算的意义所构成的函项。受限组合原则的定义为:

令 A=〈A, F〉是句法代数和 B=〈B, G〉是语义代数,映射 h:A→B 是同态的,当且仅当,存在一个映射 h′:F→G,存在 f∈F 并且存在 $a_1,\cdots,a_n\in A$ 满足:

$h(f(a_1,\cdots,a_i,\cdots,a_n))=h'(f)(h(a_1),\ h(a_{i-1}),\ h$

$(a_{i+1})\cdots,h(a_n))(1<i<n)$

其中，$h(a_i)$是作为虚化成分的部分的语义。

例子解读：动词短语表达式“飞快地跑步”的句法生成：f(飞快，地，跑步)。令“飞快”的逻辑语义为 α，“跑步”的逻辑语义是 γ，“飞快地跑步”的逻辑语义就是 $h(f(飞快，地，跑步))=h'(f)(\alpha,\gamma)=\alpha(\gamma)$。从语义角度看，这里“地”是没有逻辑语义的，在复合表达式“飞快地跑步”的整体语义组合中不起作用，是语义虚化的部分表达式。“美丽的姑娘”和“中式的家具”中的结构助词“的”也都是自然语言中的虚化成分。

三　结论

由于自然语言的丰富多样性，句法和语义的对应及意义的组合原则往往表现出异彩纷呈的局面。这给人们留下研究的空间，探讨作为数学概念的组合原则怎样通过具体生动的自然语言而呈现出多种多样的表现形式，是逻辑语义学介入自然语言信息处理领域所期待的工作，是逻辑学作为基础工具学科作用于计算机人工智能科学的价值所在。

参考文献：

[1]Janssen T, Partee. *Compositionality* [C]// Johan van Benthem et al. (eds.). *Handbook of Logic and Language* [M]. Amsterdam: Elsevier, 2011.

[2]Montague R. *Formal Philosophy* [M]. New Haven: Yale University Press, 1974.

[3]张秋成.类型逻辑语法研究[M].北京：中国人民大学出版社，2007.

[4]王欣.类型逻辑语法与现代汉语“是”和“的”[M].北京：北京语言大学出版社，2009.

[5]ZOU Chongli et al. The Categorical Logic of Vacuous Components in Natural Language[C]//Van Ditmarsch et al. (eds.). *Logic, Rationality, and Interaction*, *LNAI* 6953. Berlin: Springer-Verlag, 2011.

Principle of Compositionality and Vacuous Components in Natural Language

ZOU Chong-li[1,2]

(1.Institute of Logic and Information, Sichuan Normal University, Chengdu, Sichuan 610066;
2. Institute of Philosophy, Chinese Academy of Social Sciences, Beijing 100732, China)

Abstract: Principle of compositionality is the most important principle for Logical Semantics, a theoretical foundation of NLP (Natural Language Processing), which is one of the most important tasks in the age of artificial intelligence. Principle of compositionality embodies the idea that the meaning of the whole expression is the function of the meanings of its components. However, as a typical phenomenon of anti-syntax-semantics-correspondence, the vacuous components in complex expressions of natural language make no contribution to the meaning of the whole expression. Therefore, the meanings of its parts should be manipulated under a certain restrictions of principle of compositionality. This paper proposes the restricted principle of compositionality as a more accurate expression of the Principle.

Key words: logical semantics; principle of compositionality; natural language; vacuous components

[责任编辑：帅　巍]