

句类中 E 块的作用及其判定策略^{*}

The Effect of E-Chunk on SC and Its Determinant

王厚峰

(中国科学院声学所国家重点实验室 北京100080)

(华中师范大学计算机系 武汉430079)

Abstract Sentences Analysis in HNC theory is called Semantics Analysis of Sentences Category (SASC), by which, the Sentence Category (SC) will be determined at first, and the type of Semantic Chunks will be obtained and their inner constituents are analyzed based on SC. E(igen) Semantic Chunk is very important for SCAS. It is similar to traditional predicate formally, but there exist differences between them in meaning. In this paper, we'll state main idea of SASC by analyzing features and constituents of E-chunk.

Keywords V-concept, E-semantic chunk, Predicate, Sentence category

1 引言

在一个句子中,谓语起着极其重要的作用。多年来,无论是以句法分析为主的自然语言处理,还是以语义分析为主的自然语言理解,都把对谓语的确定作为句子分析的关键因素。首先,从句子的构成形式上看,可以通过谓语的类型,判断句子是无宾语,单宾语,还是双宾语,甚至是小句作宾语;其次,从意义上看,又可以通过它得到句子的主语和宾语所具备的基本语义信息,是人还是物,是具体概念还是抽象概念,或者,从格的角度讲,是施事格还是受事格等。HNC理论同样肯定谓语所起的作用,同时认为,一方面“谓语”的内涵过泛,另一方面对谓语的复合构成缺乏深层的揭示。为此,它引入了表示特征的 E(igen)语义块基元的概念,由它构成特征语义块 EK。同时还引入了三类广义对象语义块基元: A(Actor), B(object), C(Content), 分别表示作用者,对象和内容,三者与特征语义块基元一起共同构成广义对象语义块的表示式。特征语义块与广义对象语义块一起,共同构成句子的主语义块。

特征语义块大体上相当于语言学中所指的“动词谓语”,至于汉语中的非动词充任谓语的情况, HNC统一将之定义为状态句的 S04句类,相应的谓语由称为 SC的语义块充任。由于具有 EK块(即动词作谓语)的句子具有广泛的代表性,而且在构造上和内涵上都非常复杂,因此在语言信息处理上需要着重讨论。本文将

针对具有 EK 的这类句子,着重阐述 HNC 特征语义块的特点及其发现和验证策略;并从语义上和形式上比较与谓语的差别,进而揭示句类分析的主要思想。

2 句类及有关的基本概念

为了建立自然语言的理解体系, HNC 理论提出了一系列新思想。从句子来看, HNC 理论认为,它所表达的概念在脉络上应该前后一致,要求句子的各个词之间既要符合格式上的规范,又要受意义上的约束。相应地,对句子的分析既应该从句子格式的角度又要从词与词之间的搭配关系的角度来判断其是否满足要求,总之,就是要检验其概念联想脉络的合理性。

什么是句子格式,句子究竟有多少格式?这是几十年来,语言学家和计算语言学家一直苦苦追寻的问题^[2,5]。HNC 理论从概念联想脉络的角度提出了7个基本句类和57个基本子句类,并通过大量语料考察发现,任何句子均可以通过这些句类(包括子类)或者这些句类的混合来表示^[6],从而为计算机自动处理语言提供了严格的形式规范。

HNC 理论将句子的下一级单位定义为语义块,语义块又分为主语义块、辅语义块和主辅两可的语义块。各类语义块的基元是固定的。主语义块基元有4种,它们构成句子的主体,形成句子的核心意义,而辅语义块既不依赖于句类,也不对句子的整体意义构成实质性影响,它们由通常意义下的一部分状语组成。HNC 理

^{*} 本文受“95”重大项目和“973”项目资助。王厚峰 副教授,博士后,研究方向为计算语言学和计算机网络。

论发现,基本句类(包括子类)最多可以有4个主语义块,标准(或规范)句子至少包含2个主语义块,具体类型和数目均由句类确定。因此,在HNC体系中,确定句类就成了分析句子的关键。如何确定句类,它们又该如何表征?这就不能不提到E块。

根据HNC的考察,各句类的标准格式从形式上可以写成

$$JK_1 + EK + \sum JK_i \quad (1)$$

其中, $\sum JK_i$ 的个数与 JK_1 一起不超过三个, JK 表示广义对象语义块。 EK 代表句类,如为 X 时,表示作用句;为 T 时,为转移句……。它既决定了该类中 JK 的个数,又决定了 JK 的具体意义(如,某个 JK 是 XA , TA , XB , TB , 还是 XC , TC 等)。

那么,又该如何确定 E 块呢?这需要引入 v 概念。如果说HNC通过句类对句子进行了成功的语义分类,通过概念层次网络提供了描述词汇的符号体系,那么,这两个层面的相互关联则是通过由 v 概念向 E 语义块映射来实现的。

HNC理论所提出的概念层次网络符号是对概念的一种有效划分,为词义的描述提供了手段,但为了更完整地描述词汇,还需要考虑形态特征。这是HNC理论提出抽象概念五元组(v, g, u, z, r)的意图所在, v 就表示动态特征^[1],每一内涵在不同的情况下可能表现为不同的外部形态;例如,HNC将80定义为一般思维活动,但当这一概念表现为“动”的特征时,就用 $v80$ 表示,如“思考”;如果本身表示的是抽象概念这一“静”的特征,则用 $g80$ 表示,如“思维”;如果表示的是思维活动的结果,即,通过思维活动产生了某种效应,则表示为 $r80$,如“想法”。虽然HNC从五个侧面对概念作了划分,但并不一定每个概念都具有与五个形态特征对应的词,不同的概念有不同的偏好。

当概念层次网络符号与形态符号结合起来描述具体词时,既可以表示词的概念意义,又可以表示词的“词性”。上面所给的“思维”,“思考”和“想法”依次相应于名词,动词和名词,它们具有相同的概念意义。但是,这三个词具有的形态符号是不一样的,因此HNC所表示的“词性”(也称为概念类别)与通常所说的词性(如名词)又有所区别。

HNC理论无论是从句类的角度,还是从词层面的角度始终把握着一种“动”的思维哲理。从内涵上讲,提出了“作用-效应”链^[1],实际上这是从事物的动态过程来对每一环节进行描述的。从形态上讲,也反映了一种动态规律:一切事物都处在动态发展和变化的过程(v)之中,但必须有相对静止的观察点,变化的源头用(g)表示,变化过程中必然伴随着出现某些属性和特征

(u),在产生质变之前,必然会有量值(z)的变化和积聚,其结果必定会产生某种效应 r ,效应又可以成为新的源头……。这一过程循环往复,以至无穷。由此可以看出,无论在内涵上还是在形态上,概念都存在着一种天然的交叉点。

如何实现 v 概念向 E 语义块的映射呢?这又是“作用-效应”链。句类按照这一规律划分,表示基元概念的超级语义网络也按这一规律划分。当某一词条具有 v 特征,且又是一般作用性概念时,所对应的句类就是一般作用句。这是HNC从词汇层面面向句类层面映射的基本点。

但是,从 v 到 E 的映射并不完全一一对应,概念意义上的交叉性和词本身的多义性是造成一对多的主要原因。例如“撤回”,既是作用型概念又是效应型概念,说明如下:

例1 a)我们撤回了部队。b)部队撤回了。

其中,a)所强调的是“撤”这一作用;而b)强调的则是“回”这一结果。对于概念意义与句类间的一对多映射甚至不一致映射,HNC在词库上作了特殊描述。关于多映射的选一处理,HNC只需要利用基本句类知识和词的个性知识即可,在例1a)中“撤回”作为作用句,首先是因为它满足作用性的标准句类格式:

$$A + X + B \quad (2)$$

同时,我们在词知识库中可以规定,当“撤回”映射为作用句时,作用的对象是不可省略的。这就排除了例1b)作为作用句省略 B 的可能。反过来,也可以肯定例1b)是效应句而非作用句:

$$YB + Y \quad (3)$$

象这样,通过基本句类知识对语义块的数目和类型的要求这一共性以及词知识库所给的特殊个性知识,就可以完成由 v 向 E 语义块的映射。上面的 X 和 Y 均是 E 块所表示的具体含义。

3 E 语义块的内涵和构成

在以“句法”分析为主的句子分析中,为了有效地确定句子结构,首先需要寻找到谓语并确定与之相关的宾语,一般的处理方法是对动词进行细分类。目前,大多数系统将动词划分为14个子类:非谓语动词: vg ;不及物动词: vgo ;接体宾: vgn ;接形宾: vga ;接述宾: vgv ;接双宾: vgd ;接小句: vgs ;接兼语: vgj ;连动动词: vv ;补动词: vc ;情态动词: vd ;系动词: vi ;“是”动词: vy ;“有”动词: vh 。不难发现,这种分类主要考虑的是句法因素。由于语言十分灵活,在不同的情况下,同一个动词的表现形式也存在很大区别。

例2 a)张先生教我们。b)张先生教数学课。c)张先生教我们小孩。d)张先生教我们小孩数学课。e)张先

生教我们小孩画图画。

从上述例子所使用的动词“教”，不难看出，有如下二个特点。

◇所接宾语的形式一样，但宾语的内涵确存在很大的差异。在例2a)、b)中，虽然“教”所接的都是单宾语，而且均为体词，但其概念是不一样的，一者为对象，一者为内容。按上述分类所给的14个细类，并不能作这一区别，若想区别这一情况，势必会增加动词细类，使词性集增大。

◇动词自身的内涵一样，但所接的宾语的形式不一样。上述例子中的“教”出现了接体宾(vgn)、接双宾(vgd)和接兼语(vgj)的不同形式。这无疑增加了词性标注的复杂度和难度。例2c)的“教”所接的究竟是双宾语还是单宾语呢？不适当考虑语义而作判断具有一定困难。如果将名词“小孩”改为名词“数学”，情况就发生了变化。这是否需要名词作进一步的细分？

即使准确标定词性，在分析时也会出现困难。在例2d)中的两个宾语的分界线的划分就存在歧义，e)也有类似的问题。

从意义上讲，上述5个句子的“教”都表示传授知识，即传授者将信息(知识)传输给接收者，这就必然涉及到3个对象以及一个“传输”行为。三个对象是，传授者，接收者和传输的内容，并且，还可以进一步地判断，传授者和接收者必须是人(包括高等动物)，而传输的内容多为抽象概念，在HNC上表示为：

$$RB1 + R011T3 * 31 + TB + TC \quad (4)$$

这是关系信息转移句的句类形式，R011T3是E块的具体含义，表示关系信息转移。这一表示式也表明，对应的句子属于4块句。然而，4个主块并不一定必须同时出现在实际句子中，很多情况下可以省略后两个广义对象语义块之一。由于各类语义块有明确的语义要求，即使某语义块缺省，也不难判断。

根据上述分析，我们容易发现，例2a)所省略的是第二宾语——传授的内容；而在b)中省略的则是第一宾语——“接收者”；c)并不表示出现双宾语，而是省略了传授的内容(“我们小孩”整体作为接收者)，因为句中设有可作为内容的部分出现；由此也可以推断，d)中的分界就在“小孩”与“数学课”之间；同样e)中TC=“画+图画”。HNC理论引入的广义对象语义块C具有非常灵活的表现形式，作为TC表示的传输内容，既可以由抽象概念表示，也可以由一个句子表示(HNC称由一个句子表示的块为块扩)，还可以由句蜕来表示(如果一个句子通过结构上的变换，不再满足主谓结构，则称为句蜕)。e)实际上就是块扩。不过，这里的块扩由TB+TC共同构成。这是R011T3型混合句类的基本特征，也称为基本句类知识。这项知识是概念层面

的，属于共性知识，在概念知识库中单独描述，与具体词无关。

省略是句中常见的语言现象。由于这种现象会使得词的相对位置发生变化，从而影响句法分析的统计模型和规则模型，增加处理的复杂性，降低系统的性能。块扩和句蜕同样会导致这种情况。对动词而言，表现在兼类数增加。但句类分析并不害怕这种现象，就象我们在前面所讨论的，尽管一个词也会出现多个意义，且表示为v，同样涉及到多选一的问题，但多选一可以根据实际的现场信息与基本句类这一共性和词本身具有的个性联合解决，如上一节所给的例1。

通过上面的讨论，我们发现，仅仅从句法的角度对动词分类是不够的，那么，为什么不在分类中加入语义成分呢？如果这样做，必然会使动词的下一子类更加复杂，使词性集膨胀，从而增加词性排歧的难度和复杂性。

除了意义和搭配上的区别外，在构成上，E块与动词谓语也有区别。E块除了其动词外，还包括与之相关的前后附加信息，其一般表示式为^[1]：

$$EK = QE + EQ + E + EH + HE \quad (5)$$

等号左边的EK表示特征语义块整体，右边的E表示特征语义块基元，也称为特征语义块要素(在不引起混淆的情况下，我们将不加区分地使用E)；其中，

$$EK_n = EQ + E + EH \quad (6)$$

为E块的核心部分^[7](Q, H取自汉语拼音的“前、后”的声母符号)。通常以如下四种形式出现：

$$EK_n = E \quad (7) \quad EK_n = EQ + E \quad (8)$$

$$EK_n = E + EH \quad (9) \quad EK_n = EQ + EH \quad (10)$$

例3 a)我们对他们的不良行为应该予以批评。

(予以+批评→EQ+E)

b)张医生马上为李小姐做手术。

(做+手术→E+EH)

c)我们要贯彻执行党的方针政策。

(贯彻+执行→EQ+EH)

由例3b)可以看到，核心部分不仅有动词，也可以有名词，也就是说，可以包括某些动宾结构或VP。

对于附加部分，HE与通常所说的“述补”结构中的补语有相似之处，但是表示时态的助词“着，了，过”以及表示趋向的词“起来，出去，去”等也属于HE中的一种(在HNC上称为hv)；“状心”结构中的状语则与QE有相似之处，但由短语结构充任的状语不属于QE(它们大都形成HNC中称为的辅块)，某些不足以作为发现核心E的副词也不属于QE。QE是发现E核心的激活信息，它们主要由情态(上例中的“应该”)，时态(上例中的“马上”)和逻辑态(如“正是”，“历来”)组成。

HNC通过在意义上对句子分类，通过在构形上寻

找可以激活核心 E 的附加部分,为核心 E 的假设和检验提供了极大的方便,下一节将根据 E 的特征给出其判断策略。

4 全局 E 的确定策略

从上两节的讨论,我们可以看到,E 块具有如下三个特征:

(1)意义特征.E 块由具有 v 特征的词充任,具体的 E 类型由相应的词所具有的概念意义确定;如例1中的 X 和 Y,例2中的 R011T3。

(2)格式特征.E 的具体意义决定着句类,也就决定了语义块的类型和数目,式(2)表示为3块句,式(3)表示为2块句,式(4)表示为4块句;它还唯一地确定了 E 语义块与各广义对象语义块之间以及广义对象语义块相互之间在语义上的搭配关系。这种关系并不随语义块的省缺而被破坏。省缺是常见的语言现象,在句类指导下,省缺部分可以判断,甚至可以根据上下文予以恢复。

(3)构成特征.E 具有通用的构成表示式,它是识别全局 E 的关键。

E 语义块的三大特征为句中全局 E 的确定和句类分析提供了条件。句类分析与句法分析在处理步骤上的明显不同是,句类分析采用了假设-检验方式,而句法分析则采用了短语构成(或分解)方式。句类分析的基本步骤如图1。基本流程中的前三步,实际上就是 E 块的确定过程:

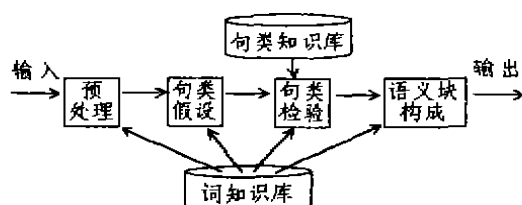


图1 句类分析的基本流程

预处理(也称感知) 是为全局 E 的假设做准备的,主要完成三项工作。

•分段处理,如果相邻的两个汉字(可能包括前后的汉字)不能成为一个词,那么,相邻的汉字之间就存在分段符,分段处理就是寻找汉字串的所有分段符过程;

•词的假定,相邻的两个分段符之间的部分称为一个段。对每个段,按段长为奇数和偶数原则假定词——偶段优先考虑双字词,奇段优先在奇位置考虑单字词;

•从假定的词中标记具有 v, l 概念特征的所有词。 l 类概念中有三类激活信息,一是语句规范格式的主语

义块标记符,二是辅语义块的标记符,三是特征语义块的 QE。

句类假设 是根据上述预处理假设全局 E,主要包括:

•排除不能充任全局 E 的词(如“的”和指示代词所指的 v 概念的词);

•对剩余的 v 概念的词,进行优先性排队,根据式(5),如果某个 v 概念的词含有的前后附加成分越丰富,则作为全局 E 的优先性越高(E 的构成特征);

•根据排队结果及 v 对应的概念意义(E 的意义特征),假设全局 E。

句类检验 根据假设的句类,检验各语义块之间是否满足搭配关系(E 的格式特征)。

语言是非常复杂的,一个句子可能会有多 v ,而且可能会有多个 E 被假定,甚至有几个 E 被检验通过,但最终会有不同的优先级,有的还会通过语义块构成来否定。为了说明分析过程,我们将简要地分析例2和例3a)。

例2a)~d) 只有一个 v ,根据前面分析,各句很容易假设为(4)式表示的句类并通过检验。对例2e)就受到二个 v “教”与“画”的干扰,但根据 v 的排队原则——“在尾尾优先,其余首优先”,尾部没有 v ,应优先假定“教”为全局 E。同样符合式(4)并容易通过检验。反过来,即使假定“画”为全局 E,由“画”的概念意义,对应的句类为作用性句类:

$$A + X + B \quad (11)$$

由词本身的个性知识可以判断,A 是不能由原型句蜕(即块扩形句蜕)来充当,也可以否定“画”为全局 E。

对例3a)有“应该”作为 QE,“予以”作为 EQ,真正起作用的是由“批评”为 E。根据“批评”的概念意义,对应着转移作用句,标准格式为:

$$TA + T3X + B \quad (12)$$

对应的规范格式为:

$$TA + B + T3X \quad (13)$$

其中,B 的前移是由主语义块指示符“对”引导的(标准格式向规范格式的转化均由指示符引导),根据共性和个性知识的要求,TA 必须是“人”,B 可以是“人”或抽象概念或者两者同时出现,不难看到,检验符合条件。

有关 E 块的排队和排除的详细讨论见文[1]。

总结 HNC 理论是一种新理论,也是自然语言处理的一种新思路。基于这一体系的句类分析,E 块的判断具有举足轻重的作用。本文针对这一中心点,简要地与句法分析上的“动词谓语”作了比较,分析了 E 块的特征,意义及其形态构成,并以此为出发点,给出了 E 块判断的基本策略。

(下转第25页)

满足 $p'_1(M) = \frac{1}{poly} - \Delta = \frac{1}{poly}$ 。由于 M 中的图都是好的, 故 $p'_1(M) = \frac{1}{poly}$ 就意味着 $p(M) = \frac{1}{poly}$ 。所以算法 A 能在分布 p (图上的均匀分布) 的图 G 的 $\frac{1}{poly}$ 因子成功找出大团。由此可得下面的结论:

结论3 假设对 $\delta > 0, k \leq (2-\delta)\log_2 n$, 存在一个确定性多项式时间算法 A , 在由分布 p'_1 导出的图中找 k -团, 且概率为 $\frac{1}{q(n)}$ ($q(n)$ 为多项式), 则必存在多项式 $q'(n)$, 使得算法 A 在由分布 p 导出的图中找 k -团, 且概率为 $\frac{1}{q'(n)}$ 。

在随机图中找大团是难的这一假设下, 上述结论表明当团 K 的大小足够大时, 例如把 $\frac{3}{2}\log_2 n$ 大小的团随机插入到随机图 G 产生 G' , 在 G' 中找任何大团仍是难的。由此我们就可构造单向函数 $f: G \rightarrow K \rightarrow G$, 这里 G 是 n 个顶点的图结合, 而 K 则是所有 k 个顶点的集合 (即 $\{1, 2, \dots, n\}$) 的所有子集全体。 $f(G, K)$ 是对图 G 的改动, 使得所插入的子图完全由 K 导出, 显然由 K 是容易得到新图 $G' = f(G, K)$, 而由 $f(G, K)$ 要高概率地求出 K , 则在随机图中找大团是难的假设下, 将是难的, 因此 f 符合单向函数的准则。当然由 G 和 G' 可容易得到 K , 利用这一特性我们就可构造一个具有分级权限的秘密密钥管理模式。其依据就是在随机图中随机插入任意常量个团 $K_1, K_2, \dots, K_n$ 生成 G' , 把 $\{K_i\}$ 作为秘密密钥, 而 G' 作为公开密钥。 P_1 生成随机图 G 并在 G 上随机插入团 K_1 产生图 G_1 送给 P_2 , P_2 在 G_1 上随机插入团 K_2 产生图 G_2 送给 P_3 , 继续这个过程直到 P_n, P_n 在 G_{n-1} 上随机插入团 K_n 产生图 G_n 作为公开密钥。由于 P_1 知道 G , 故 P_1 可以得到 $P_2, P_3, \dots, P_n$ 的

秘密密钥 (虽然 P_1 不知道这些密钥是哪个成员), 而其他成员则不可能得到 P_1 的秘密密钥。一般地, P_i 可以求出所有 $P_j (j > i)$ 的秘密密钥。由此即得分级密钥管理模式。

结语 用格的归约理论可以分析研究计算部分密钥位与整个密钥位的计算复杂性, 研究公钥密码系统的安全性, 而团问题应用于密码系统更是为计算密码研究拓展了方向。但目前还存在这样一些问题: 怎样建立一个实用的基于格归约密码系统, 是否能利用团问题直接建立一个公钥密码系统。所有这些都有待于我们不断研究和探讨。

参考文献

- 1 Ajtai M. Generating Hard Instances of Lattice Problems. STOC, 1996. 99~108
- 2 Ajtai M, Dwork C. A public-key cryptosystems with worst-case/average-case equivalence. STOC, 1997. 284~293
- 3 Babai L. On Lovasz's lattice reduction and the nearest lattice point problem. Combinatorica, 1986. 5: 1~3
- 4 Boneh D, Venkatesan R. Hardness of computing the most significant bits of secret keys in Diffie-Hellman and related schemes. Proc. of Crypto, 1996
- 5 Goldreich O, et al. Public key cryptosystems from lattice reduction problems. Electronic Colloquium on Computational Complexity Tr96~056
- 6 Jules A, Peinado M. Hiding Cliques for Cryptographic Security. STOC, 1998. 678~684
- 7 J. Y. C. An Improved Worst-Case to Average-Case Connection for Lattice Problems. Proc, 38th FOCS, 1997. 468~477
- 8 Boneh D, Venkatesan R. Rounding in lattices and its cryptographic applications. SODA, 1997. 675~681
- 9 L. Kučera. A generalized encryption scheme based on random graphs. In Graph-Theoretic Concepts in Computer Science, WG'91, Lecture Notes in Computer Science 570, Springer-Verlag, 1991. 180~186

(上接第67页)

自然语言是非常复杂的, 其处理也是十分困难的。虽然在理论上, HNC 已经显示出了强大的解模糊能力, 但是否有真正的实用价值, 需要做大量的验证工作。值得庆幸的是, 由国家语委和“95”重大攻关课题所资助的汉语音-字流转换系统, 经过测试, 与 HNC 理论上的预期基本吻合。这使得我们更加满怀信心去实现“973”资助的“英-汉”互译系统的研究。

本文所给的许多思想已经在“语音-字流转换”系统中实现。因此, 本文的形成也是 HNC 联合攻关组共同努力的结果。

参考文献

- 1 黄曾阳. HNC (概念层次网络) 理论——计算机理解语言

研究的新路. 清华大学出版社, 1998

- 2 鲁川. 自然语言处理的语言模型. 见: 1998年中文信息处理国际会议论文集. 清华大学出版社, 1998
- 3 李临定. 现代汉语句型. 商务印书馆, 1986
- 4 罗振声, 郑碧霞. 汉语句型自动分析和分布统计算法与策略的研究. 中文信息学报, 1994(2)
- 5 罗振声, 孙长健, 孙才. 汉语句型成分自动分析中谓语识别策略的研究. 计算语言学进展与应用, 清华大学出版社, 1995
- 6 刘志文. 自然语言语句的 HNC 表示. 语言文字应用, 1998(2)
- 7 陶明扬. 特征语义块的构成与复合. 中国人民大学96级硕士论文, 1999
- 8 张艳红. HNC 理论的主辅语义块及其变换. 北京语言文化大学96级硕士论文, 1999