HNC（概念层次网络）理论

——计算机理解语言研究的新思路

黄曾阳　著

清 华 大 学 出 版 社

(京)新登字 158 号

内 容 简 介

自然语言理解是人工智能的一个重要分支,是几十年来未能攻克的世界性的重大科学难题。本书作者历经 8 年潜心研究创立的 HNC(概念层次网络)理论,赢得了自然语言语句理解处理的新突破。HNC 理论包括自然语言概念体系的理论模式、自然语言语义块和语句的理论模式、句群和篇章要点的表述模式、短期记忆和长期记忆的形成及相互转换模式、基于文字文本的自学习模式。本书反映了以上前两个模式的理论研究及其技术实现的成果。

HNC 理论在机器翻译、电话翻译、人机对话、智能检索以及自动文摘等语言处理领域有广阔的应用前景。

本书是 HNC 理论的第一部专著,全书分为正文和附录两大部分。正文汇编了作者在不同时期从不同角度对 HNC 理论的阐述,内容包括 HNC 理论概要、HNC 理解处理论文选录、HNC 理解处理的 52 个论题、HNC 理解处理问答和语义学日记选录。附录部分收录了作者的两封学术信函和他的同事或学生有关 HNC 理论的论文,是对正文的补充,内容涉及句类分析的技术实现、对 HNC 的阐发和评介以及 HNC 知识库的建设。

本书的读者对象为人工智能、语言信息处理、语言研究等领域的专家学者、研究生、本科生及其他高新技术的研究者和工作者。

图书在版编目(CIP)数据

HNC(概念层次网络)理论/黄曾阳著 .—北京:清华大学出版社,1998.11
ISBN 7-302-03200-9

Ⅰ.HNC… Ⅱ.黄… Ⅲ.自然语言-理论 Ⅳ.TP301.2

中国版本图书馆 CIP 数据核字(98)第 32558 号

出版者:清华大学出版社(北京清华大学校内,邮编 100084)
http://www.tup.tsinghua.edu.cn
印刷者:清华大学印刷厂
发行者:新华书店总店北京发行所
开 本:787 × 960 1/16 **印张**:33.5 **字数**:738 千字
版 次:1998 年 11 月第 1 版 1999 年 4 月第 2 次印刷
书 号:ISBN 7-302-03200-9/TP·1709
印 数:2001 3000
定 价:66.00 元

题　　词

语言信息处理包含多方面的内容,而我对自然语言理解这一块儿特别偏爱,因为它是语言信息处理其他领域的重要基础。机器翻译搞了几十年,但至今未能达到可供实际应用的水平,其主要原因正是自然语言理解没有获得根本性的突破。任何科学上的突破都需要非常规的思维即扩散性的求异思维。黄曾阳先生创立的面向整个自然语言理解的理论框架 HNC,在语义表达上有自己的特色,在语义处理上走了一条新路。鉴于汉语语法研究尚有诸多困惑,HNC 理论所走的以语义表达为基础的新路子对突破汉语理解问题尤其有实际意义。

1997 年 7 月 1 日

弁　言

HNC 是 Hierarchical Network of Concepts (概念层次网络)的简称,是关于自然语言理解处理的一个理论体系。这个理论体系的基本思路与传统计算语言学理论有本质的不同,如下表所示:

比较内容	传统方式	HNC
句子构成单元	短语	语义块
句子表述模式	句法树	句类物理表示式
每一模式的构成单元数量	不定	确定
模式总数量	不可穷尽	已穷尽
句子分析方式	句法分析	句类分析
分析所依附的基本知识	词性和句法成分	概念联想脉络
理解处理所运用的知识	句法知识为纲	句类知识为纲
知识表示方式	英语词语为主	HNC 符号,全数字化
词义表示	语义原语	HNC 符号
词义表示通用性	无,与语种有关	通用,与语种无关
知识库结构	单一	多层面
	复杂特征集	以概念层面知识为纲,以语言知识为目
语境	尚无对策	可自动生成
解模糊能力及语句合理性判断能力	弱	强,可接近常人水平
方法论	以综合与统计为主	以演绎和验证为主

从上面的对比可以看出,HNC 试图对立足于西方语法学理论体系的自然语言理解处理方案进行全面的改革,建立一种模拟大脑语言感知过程的自然语言表述模式和计算机理解处理模式。这当然是一项艰难的探索,甚至不是一代人的努力可以完成的。我从汉语以“字义基元化,词义组合化”方式构建新词(所以两千年来汉字只减不增)的独特语言现象受到启发,以充分基元化的约 1200 个汉字及其组合词语为素材,以建立概念联想脉络为目标,从自然语言概念体系表述模式的理论设计着手,开始这一漫长的探索。在第一步探索过程中,有幸发现了自然语言无限的语句可以用有限的句类物理表示式来表达。这是自然语言概念体系总体特征必然导致的结论。这个结论与传统认识大相径庭,因而也曾使我本人深感震惊,花费了三年的时间从各个侧面对此进行了验证。

句类物理表示式(简称句类)是对语句全局联想脉络的一种表达模式,分为基本句类、混合句类和复合句类三种类型。基本句类有 57 种一级子类,混合和复合句类各有 3192 种,汉语常用的混合和复合句类大约是 300 多种。这一组完备表示式的确定,是 HNC 联合攻关组(由中科院声学所、中国人民大学对外语言文化学院、北京语言文化大学语言信息处理研究所和中科院软件工程中心四个单位的有关专业人员组成)一年来取得的重大成果之一。

以上述语句物理表示式为基础的语句分析方式命名为句类分析。其基本过程是:第一步,进行语义块感知和句类假设;第二步,进行句类假设的合格性检验及合理性分析;第三步,对合格合理的句类进行语义块构成分析。

句类分析方式对汉语尤为适用,因为汉语拥有比较明确的语义块区分标志;汉语语义块的封闭性优于西语,像众所周知的"I saw a girl with a telescope near the bank"之类典型和普遍的语义块构成模糊,汉语是不存在的。传统汉语句法分析所遇到的基本困难,如述语动词的辨识,分词"瓶颈",词性兼类现象的困扰等,句类分析都已形成行之有效的解决方案。句类分析软件的实现是 HNC 联合攻关组一年来取得重大成果之二。

以上述语句完备物理表示式为基础的 HNC 知识库建设已从多年来的小作坊局面转变成符合现代高技术发展要求的规章齐备、分工精细、科学组装、质检完备的新阶段。这是 HNC 联合攻关组一年来取得的重大成果之三。

上列研究成果表明,HNC 理论预定的自然语言五项理论模式的探索,即

1. 自然语言概念体系的理论模式
2. 自然语言语义块和语句的理论模式
3. 句群和篇章要点的表述模式
4. 短期记忆和长期记忆的形成及其相互转换模式
5. 基于文字文本的计算机自学习模式

已完成了第一期目标:建立了前两项理论模式,实现了句类分析这一自然语言理解处理的全新方案。

这一阶段性成果应清华大学出版社之约,汇编成这部 70 万字专著:《HNC(概念层次网络)理论》。在这个基础上,HNC 联合攻关组拟定了一个远景发展目标和近期工作纲要。

远景发展目标是:让计算机能够像常人那样读懂自然语言的文字文本和听懂语音文本。这对于信息时代从当前的以数据处理为主的低级阶段向未来的以知识处理为主的高级阶段的转变和发展,显然具有决定性的意义。这一远景目标,过去一直处于"茫茫语海无舟渡"的困境,而现在可以说,出现了"蓦然回首可为期"的契机。为了实现这一远景目标,首先需要正式启动 HNC 预定的后三项理论模式的探索。这三项模式的共同问题同前两个模式一样仍然是概念联想脉络的激活、扩展、浓缩、转换与存储。这些联想脉络当然比语句层面的复杂得多,不可能用一组物理表示式来表达。但是,语句层面联想脉络表示式可构成事件联想脉络的基础。因为 57 组基本句类表示式并不是零散的各自独立的局域网络,而具有集群特征,在集群内部和集群之间都呈现出特定的交式和链式关联性。对这些关联性的揭示和表

达是下一步理论探索的中心任务。

近期工作纲要主要是两个方面。一是检验句类分析能否在双向机器翻译方面产生理论上预期的突破性进展。这一进展的标志是‘先懂后译’,从而根本改变译准率徘徊于70%左右的困境。二是检验句类分析在语音模糊消解及纠错方面的潜力。连续语音识别很难也没有必要做到90%以上的首选正确音节识别率,听觉实验表明,人类听觉预处理也只能听清楚连续语音流中70%的音节。计算机听懂语音文本的关键,当前已取决于后续理解处理系统的解模糊及纠错能力。

在机器翻译方面,首先要开展不同语种之间的句类偏好从而导致句类转换的研究,其次要开展不同语种之间语义块构成方式习惯差异的研究,第三是建立相应语种的HNC语言知识库和语法知识库。句类分析可保证‘看懂’前提的实现,因此,在上列两项研究和知识库建设取得一定成果的基础上,机器翻译可能取得人们盼望已久的突破性进展。

在语音识别的理解处理方面,解模糊处理属于句类分析的常规项目,难点在于纠错。纠错的基础首先是语境的建立,语境提供句群或篇章的要点信息。在这一要点信息和句类表示式的引导下,运用它们所提供的预期信息,就有可能发现语音识别的个别音节错误,并确定隐含在错误音节后面的正确概念类别,这是对大脑语言感知过程可实现的适当模拟。这里,句段信息要点的预期和认定是理解处理的中心内容,它只涉及概念层面的操作,这个处理步调是关键性的。至于从概念层面到具体词语的转换,只是一个不难解决的具体技术问题了。

语境的自动生成在根本上依赖于上述三项待探索的理论模式的建立。不过,在获得并运用完善的理论模式之前,也可以依据语音流中一些关键词语的HNC映射符号,通过特定的统计方式得到一些简明语境类型,并将它应用于纠错处理。HNC联合攻关组将从理论探索和统计方式两方面同时开展纠错处理的研究。

HNC已取得的进展不过是万里征途的第一步。虽然航向已经确定,但探索的艰辛依然。我们热切期盼得到有关部门领导的支持和指导,得到有关领域专家的匡正与合作。

黄曾阳

1998年10月

编 者 的 话

无限和不确定的表观与有限和确定的本质

黄曾阳先生的专著《HNC(概念层次网络)理论》出版了。作为‘HNC联合攻关组’的一个成员和《HNC(概念层次网络)理论》一书的编者,我非常高兴,有许多话要说。

计算机要智能化,语言研究要现代化,语言学和计算机科学的结合是历史发展的必然趋势。为了顺应这一历史发展潮流,我作为一个积极和计算机科学相结合的语言研究工作者,经中国工程院资深院士陈力为教授和全国计算语言学专委会首届专委会主任鲁川教授的引荐,在中文信息界的许多朋友支持下,于1986年开始先后担任全国计算语言学专委会的专委和中国中文信息学会理事及学术委员,相继参加了国内外有关信息处理的重大科研课题的研究,成为跨接语言学和计算机科学的语言研究工作者。近10多年来,我一直处在向中文信息界学习的过程中。通过学习,我认识到用流行在我国语言学界的‘语素—词—词组—句子成分—单句—复句’这一套汉语语法学去解决汉语的理解问题是走不通的。为什么?几千年来,汉语语言学的传统研究主要集中在‘字’的形、音、义上,相应建立了文字学、音韵学、训诂学。从1898年马建忠的《马氏文通》出版开始,汉语语法学出现以西方语言学理论研究汉语的状况,并成为汉语语法研究的主流派。应该说,100年来的汉语语法研究是有成绩的。但随着汉语语法研究的不断深入,愈来愈多的学者认识到,西方语言学理论总的来说是在形态语言的基础上建立起来的,汉语是非形态语言,用形态语言的理论去描写非形态的汉语,显然是不对路的。这种不对路的汉语语法研究成果当然就解决不了汉语信息处理的句法分析问题。要分词嘛,没有一个科学的词的定义,词的下面跟语素划不清界线,词的上面跟词组划不清界线。要标词性嘛,名、动、形的界限划不清楚,兼类问题解决不了,而且词类跟句子成分没有一一的对应关系,词性标注跟句法分析脱节。黄曾阳先生指出,信息处理用的词汇知识,必须下连网络、上挂句类,否则对计算机毫无用处。要分析句子成分嘛,主、谓、宾、定、状、补划不清楚。要分析句型嘛,首先就划不清楚单句和复句的界限。这不是我国语言学家和语言信息处理专家无能的表现,而是汉语语法研究的路子不对造成的。我国著名的老一辈语言学家张志公先生在20世纪90年代初提出,应该有勇气打破强加在汉语头上的印欧语的语法框架,创立一套适合汉语特点的语法体系。为此,志公先生提出了初步的设想。我曾试图努力落实志公先生的设想,但感到力不胜任。我于是考虑:适合汉语特点的语法体系创立出来之前能否抛开现有的语法学另辟汉语理解的蹊径呢?正在这个时候我有机缘接触到HNC理论。HNC理论引起我的注意,首先是因为它完全摆脱了我国现有的这

套语法学的束缚,而从语言的深层入手,以语义表达为基础,为汉语理解开辟了一条新路。经过一番学习,我进一步认识到 HNC 理论提出了可供工程实现的完整的自然语言理解的理论框架,它是一个面向整个自然语言理解的强大而完备的语义描述体系,包括语句处理、句群处理、篇章处理、短时记忆向长时记忆扩展处理、文本自动学习处理。目前,已赢得了语句理解的突破,并正在产品化。

自然语言理解的发展主要围绕三个方面:1.自然语言的表述和处理模式;2.自然语言知识的表示、获取和学习;3.研制开发自然语言的应用系统。其中,自然语言的表述和处理模式是根本,它决定着整个自然语言理解的方向和进程。黄曾阳先生经过八年的艰苦探索,在决定自然语言理解方向和进程的这一根本问题上提出了三大理论要点:(1)要把自然语言所表述的知识划分为概念、语言和常识三个独立的层面,对不同层面采取不同的知识表示策略和学习方式,形成各自的知识库系统。知识库建设的首要目标应定位于自然语言模糊消解,这是 HNC 理论对迄今为止的知识库建设进行总结后得出的论断。(2)建立网络式概念基元符号体系,即概念表述的数学表示式。这个符号体系或表示式应具有语义完备性,能够与自然语言的词语建立起语义映射关系,同时,它必须是高度数字化的,每一个符号基元(每个字母或数字)都具有确定的意义,可充当概念联想的激活因子。这个符号体系就是 HNC 理论设计的三大语义网络及五元组和概念组合结构等,它是计算机把握并理解语言概念的基本前提,称为局部联想脉络,是 HNC 理论的基本内容之一。(3)建立语句的语义表述模式,即语句表述的数学表示式。这一模式的完备性应表现为可表述自然语言任何语句的语义结构。为表述自然语言语句的语义结构,HNC 理论提出了语义块和句类的概念,在此基础上形成的句类格式就是语言的深层结构,它是语句分析的基点,称为全局联想脉络,是 HNC 理论的另一基本内容。以上三大理论要点,正是 HNC 理论在自然语言表述和处理模式上赢得突破性进展的表现。下面试进一步具体论述:HNC 是如何在上述三大理论要点的基础上赢得语句理解的突破的。

首先,解决了一个正确的定位问题。什么叫“理解”?不同的学科有自己特殊的认识。人工智能界多年来对“自然语言理解”的“理解”贪大求全,妄图一步登天,企求使计算机一下就能像人脑一样去理解语言。人脑异常精密复杂,其皱褶的全部表面约有一张报纸大,却拥有大脑(含 90% 脑组织)、小脑(与肌肉协调有关)、脑干(长约 75 毫米却含有控制“自律”功能的神经中枢)。人脑由 150 亿至 180 亿脑细胞组成,恰似人体司令部。在现阶段就要求计算机像人脑一样去理解语言当然就不可能实现。黄曾阳先生总结了这方面的经验教训,提出“消解模糊”作为“自然语言理解”初级阶段的标准,并认为口语有五重模糊:发音模糊、音词转换模糊、词的多义模糊、语义块构成的分合模糊、指代冗缺模糊,书面语只有后三重模糊。这五重或三重模糊的消解可进一步概括为“多义选一”的能力。“多义选一”是世界计算语言学的一个重大难题,也是人脑和计算机理解自然语言的首要任务。我认为 HNC 理论的这个定位至关重要。全世界研究自然语言理解近半个世纪,直到最近的八年才由黄曾阳先生找到正确的定位,那就是在自然语言理解的万里征途中以“消解模糊”作为坚实的第一步。

其次,创立了“消解模糊”的理论。创立一种理论首先要确定基本思路。什么是 HNC 理论的基本思路呢?HNC 理论的目标是建立一个模拟人类语言感知过程的理论模式。人对语言的理解本质上是一种认知行为,如果能描述大脑认知结构的具体模式,计算机就可以运用该模式对自然语言进行理解处理。HNC 理论把人脑认知结构分为局部和全局两类联想脉络,认为对联想脉络的表述是语言深层(即语言的语义层面)的根本问题。局部联想是指词汇层面的联想,全局联想是指语句层面的联想。HNC 理论的出发点就是运用两类联想脉络来“帮助”计算机理解自然语言。所以,用一句通俗的话来说,HNC 理论就是“帮助”计算机懂得人类语言的一种理论。这就是 HNC 理论的基本思路。从这一基本思路出发,能否设计好两类联想脉络就成为 HNC 理论成败的关键。

HNC 理论是怎样设计局部联想脉络的呢?自然语言的词汇是用来表达概念的,因此,HNC 建立的词汇层面这一局部联想脉络体现为一个概念表述体系。该表述体系是:概念分为抽象概念与具体概念,侧重于抽象概念的表达,对具体概念采取挂靠近似表达方法。外部特征和内涵是概念的两个基本特征,没有这两个基本特征便不成其为概念。HNC 理论对抽象概念的外部特征采用五元组来表达,对抽象概念的内涵采用网络层次符号来表达。其网络层次符号包含三大语义网络:基元概念语义网络、基本概念语义网络和逻辑概念语义网络。HNC 的五元组符号和三大语义网络的层次符号以及概念组合结构符号组合起来就可完成对抽象概念的完整表达,从而为计算机理解自然语言的词义提供了有力的手段。

HNC 理论又是怎样设计语句这一全局联想脉络的呢?语句联想的主要内容是语义块和句类两根支柱。语义块是句子的语义构成单位。主语义块 4 种,辅语义块 7 种。句类是句子的语义类别。有 7 个基本句类,它可构成 36 个混合句类。语义块和句类理论的基本论点是:语义块为句类的函数。语义块和句类的这种函数关系具体体现为句类格式。句类格式是指一个句子的主语义块的排列顺序。以句类格式为基点的语句分析叫做句类分析。基于 HNC 理论的句类分析,既不是基于规则的推理,也不是基于语料库的统计,而是用语句的物理表示式激活语句的全局联想脉络,黄曾阳先生认为这正是人脑感知语言过程的模式。

以上情况表明,HNC 理论科学地、成功地完成了两类联想脉络的设计。局部联想脉络和全局联想脉络不是彼此孤立的、割裂的,而是紧密相连的。连贯两类联想脉络的链条是作用效应链,这是 HNC 理论的理论基础和最伟大的创造。什么是作用效应链?作用,是指对事物产生影响;效应,是指作用产生的效果。概念层次网络理论认为,作用存在于一切事物内部和相互作用之中。作用必然产生某种效应。作用是源头,效应是结果。作用是事物发展变化的起因,效应是作用导致的结果。在达到最终的效应之前,必然伴随某种过程和转移;在达到最终的效应之后,必然出现新的关系和状态。过程和转移、关系和状态也是效应的一种表现形式。一个作用效应流程完成以后,新的效应又会引发新的作用,新的作用又会产生新的效应。如此循环往复,乃至无穷,这就是宇宙间一切事物存在、发展和消亡的基本法则,也是语言表达和概念推理的基本法则。句子的语义由“v”概念即语句核心的概念来表示,这与美国计算语言学家山克(Schank)的概念从属理论(conceptual dependency theory)是一

致的。可惜山克只主要考虑了“转移”类概念,他没有找到描述自然语言中“v”概念的完备集合,而 HNC 的作用效应链形成了这样的完备集合,完整地提出了“作用—效应—过程—转移—关系—状态”等 6 个环节,而且这 6 个环节形成一条链,这就叫作用效应链。它反映了一切事物的最大共性。自然语言的主要内容就是对作用效应链的 6 个环节进行局部和总体的具体表述,作用效应链揭示了语言表达的深层要素,形成了对自然语言进行总体表述的完整体系。它可以对任何语言的任何语句进行语义分类,并加以描述。

为使消解语句模糊的 HNC 理论得以工程的实现,黄曾阳先生设计了句类分析系统,开创了一条全新的语句理解的技术路线,那就是:从语义块感知和句类辨识入手,靠句类分析“消解模糊”。什么是语义块感知、句类辨识和句类分析呢?拿到一个语句,首先寻找表示“v”概念的词,并把它假定为特征语义块即语句的核心,据此判定整个语句的类别,这就是语义块感知和句类辨识。然后在句类知识的指导下进行语句合理性检验,这就是句类分析。如若检验成功,则句子理解正确,语句模糊即可消解;如若检验失败,则再做另外的假定和检验。在句类分析过程中,句类知识起着控制全局的指导作用,是“消解模糊”的最有力武器。

总而言之,HNC 理论之所以能赢得语句理解的突破,是因为它冲破了语句理解道路上的重重障碍。计算机理解语句,首先要抓到语句的核心。汉语的语句核心没有形态标志,拿到一个汉语的句子,计算机如何能抓到句子的核心呢?计算机如何处理带有两个以上语句核心(连动式、兼语式)的语句呢?这是汉语信息处理的一个老大难问题,这里的后一个问题也是菲尔墨的格语法无法解决的问题,HNC 理论终于突破了这道难关。抓到了语句核心之后,又面临着一个对语句核心用什么标准来分类的难题,HNC 理论用黄曾阳先生独创的作用效应链来给语句核心分类,因而也终于把语句核心分类这一难题解决了。对语句核心进行分类以后,又面临一个如何使语句核心和整个语句串通起来的难题,HNC 理论用语句核心的性质来给语句定类,什么样的语句核心就决定有什么样的句类,于是又把语句核心和整个语句的串通问题解决了。句子的语句核心和整个语句串通起来以后,HNC 便采取智能调动的举措,在句类的控制下进行语义块构成的分析。不同的句类有不同数量的语义块(语句的数学表示式)和不同性质的语义块(语句的物理表示式),由于句类又是有限的和确定的并具有覆盖自然语言语句全貌的功能,这样就解决了菲尔墨的格语法不知道有多少个格和不知道有多少类格框架等一系列的难题。在分析语义块的过程中,HNC 理论又把分词问题解决了。按传统的句法分析,分词是“瓶颈”;按 HNC 的句类分析,分词变成了句类分析获得成功时的水到渠成的“瓶底”。HNC 的句类分析之所以能冲破上述这些语句理解道路上的重重障碍,是因为 HNC 理论创立了局部联想脉络和全局联想脉络。这两个联想脉络透过自然语言无限和不确定的表观现象,抓到了沉淀在语句深层的有限和确定的本质,这就是 HNC 在词汇和语句层面的两个“完备”,即概念描述体系的“完备”和句类体系的“完备”。由于有了这两个“完备”,就赢得了语句理解的第一步。

自然语言理解,这是几十年来未能攻克的世界性重大科学难题。迄今为止,许多语言信息处理系统和产品多是基于统计的,例如,输入计算机时反复出现“完成”与“任务”相连,计

算机便能反应出“完成任务”为正确搭配。然而,这并非建立在对语言理解的基础上。15 年前,日本花费巨资搞了一个第五代计算机(又称智能计算机)计划,其中一个重要目标就是使计算机能理解人类语言,结果未获成功。美国微软公司 1998 年计划投入 26 亿美元,用于开发三项软件技术(自然语言理解、图像识别、三维图形设计),其中自然语言理解是所要开发的首要技术。由此可见,HNC 理论在语句理解上赢得的突破,对我国在高新技术领域的国际竞争具有重大的意义。

HNC 理论具有巨大的应用潜力和广阔的应用前景。多年来,在人工智能的许多应用领域没有重大的进展,其中一个主要原因就是自然语言理解未能获得根本性的突破。HNC 理论在语句理解上赢得的突破,将使机器翻译、电话翻译、人机对话、智能检索、自动文摘等语言处理的各个领域获得实质性的重大进展,并为我国创新语言信息产业带来曙光。令人可喜的是,为了 HNC 理论的产品化,在中国中文信息学会理事长、中国工程院资深院士陈力为教授和全国人大常委会副委员长、著名语言学家许嘉璐教授的积极推动下,近一年来组成了“HNC 联合攻关组”。这一“联合攻关组”包括中国科学院声学研究所、中国人民大学对外语言文化学院、北京语言文化大学信息处理研究所、中国科学院软件工程中心等单位。他们正在为 HNC 理论的产品化而紧张地工作。“联合攻关组”一年多来的研究实践充分证明,HNC 理论的发展和应用存在着巨大的潜力和广阔的前景。HNC 理论建立的语言表述和处理模型应该成为中华民族的财富,应该以它为基础开创我国的信息产业。

黄曾阳先生 50 年代毕业于北京大学物理系理论物理专业之后,在中国科学院声学研究所从事水声学和信号处理研究,搞智能探测;后转向研究语言声学。在多年的声学研究实践中,他体会到对语声只进行声学分析是远远不够的,还必须加上理解处理。要理解就涉及到语义问题,势必要进行信息处理用的语义研究,这就使他走上了研究自然语言理解的道路。黄曾阳先生在创立 HNC 理论的过程中,在外国的语言学理论和计算语言学理论中得到了有益的启示,诸如乔姆斯基的语言深层结构理论、奎廉的语义网络理论、山克的概念从属理论、菲尔墨的格语法。由于他是我国著名音韵训诂学家黄侃的侄孙,是我国著名音韵训诂学家黄焯的公子,他又从家学的传统语言文字研究成果中吸取了丰富的营养,如汉语的“字义基元化,词义组合化”便给了他很大的启示。全国人大常委会副委员长、著名语言学家许嘉璐教授对黄曾阳先生这一中外合璧,特别是弘扬祖国语言文字研究优良传统的研究路子极为赞赏,几年来多次和黄曾阳先生及 HNC 课题组成员进行学术研讨,最近一年许先生则更热情地大力支持 HNC 理论产品化的研究。“联合攻关组”计划在研制产品的同时,出版一部 HNC 理论的专著。由于黄曾阳先生忙于指挥语句理解产品化的工程和其他处理层面的设计研究,短期无暇顾及专著的撰写,我便提出将现有的论文汇编成书的建议。我的这一建议得到黄曾阳先生的赞同,同时得到中文信息学会理事长、中国工程院资深院士陈力为先生和中文信息学会秘书长曹右琦女士的热情鼓励,特别得到清华大学出版社的大力支持。我将全书分为正文和附录两部分。黄曾阳先生写的论著为正文,附录是黄曾阳先生的两封学术函件和他的同事或学生(硕士、博士)研究和评介 HNC 理论的论文。正文和附录互为补充,

相辅相成,相得益彰,以构成一个整体。

《HNC(概念层次网络)理论》的读者对象主要是人工智能、语言信息处理、语言研究等领域的专家学者、研究生、本科生以及其他高新技术的研究者和工作者。谨希望本书的出版对整个自然语言理解和中文信息处理的研究能起到促进作用。

从黄曾阳先生为本书写的‘后记’中可以看到,书中的论文是 HNC 理论创立过程中的不同阶段撰写的。为了保存历史的原貌,在这次结集时,有些论文不作过多的修改。由于本书各个部分是分头独立撰写的,因此,为了确保每篇论文的完整性,有些内容可能同时出现在不同的文章里。除此之外,还可能存在著者和编者未能发现的各种不足之处,热诚希望专家学者和广大读者多多包涵和不吝雅正。

林杏光

1998 年 8 月

于中国人民大学

导　　读

如何使计算机模拟大脑的语言感知过程,理解人类的自然语言,是信息时代从数据处理为主的低级阶段向知识处理为主的高级阶段发展所面临的巨大挑战。本书阐述的 HNC 理论,从一个全新的角度对这一挑战作出了回应。

HNC（Hierarchical Network of Concepts）是概念层次网络理论的简称,由黄曾阳先生创立,是面向自然语言理解的理论体系,因以概念化、层次化、网络化的语义表达为基础而得名。其中心目标是建立自然语言的表述和处理模式,使计算机能够模拟人脑的语言感知功能。该理论使自然语言理解获得了突破性的新进展,它所蕴涵的精深而丰富的思想对人工智能、语言学、计算机科学和认知科学等都具有重要的理论和应用价值,对中文信息处理和汉语研究尤其具有实际意义。

本书是 HNC 理论的第一部专著,绝大部分论文是第一次公开发表。全书分为正文和附录两大部分。

正文汇编了黄曾阳先生在不同时期从不同角度对 HNC 理论的阐述,分为五部分：

一、HNC 理论概要　是本书的总纲,提出了对大脑语言感知过程进行初步模拟的三大要点：1. 建立自然语言概念体系的完备表述模式；2. 建立自然语言语句的完备表示式体系；3. 以概念层面为纲,分别建立概念、语言和常识三个层面的知识库。该文对前两个要点的总体思路和完备性问题作了提纲挈领式的阐述。正文其他部分都是围绕着这三大要点而展开的。

二、HNC 理解处理论文选录（简称〈论文〉系列,亦称 Paper 系列,写于 1994 年冬至 1996 年春）　对上述三大要点进行了详细阐述,论文 1 和论文 6 涉及第一要点,论文 2 和论文 21 涉及第二要点,论文 6 和论文 7 涉及第三要点。论文 2 提出的句类分析思想体现了 HNC 理论的精髓,要同时参看〈论题〉系列的 13 和 14、6 和 7、〈问答〉系列的 30 到 32,并精读林杏光先生的“编者的话”,这有助于对这一思想的来龙去脉获得一个全面的印象。

〈论文〉系列共 21 篇。受篇幅限制,本书只选入 11 篇,其中 1 篇见附录（杜燕玲文）。其他各篇篇名及情况见“后记”。

三、HNC 理解处理的 52 个论题（简称〈论题〉系列,写于 1998 年 5 月至 7 月）　重点阐述了 HNC 理论技术实现的策略,兼及 HNC 思路的形成过程。对汉语处理的传统难题,如分词“瓶颈”、词性兼类的困扰、述语动词辨识、单音词辨识及其模糊消解等,都提出了 HNC 的解决方案。

〈HNC 理解处理的 52 个论题〉论文系列中,本书未收入的篇名清单如下,情况说明见

“后记”。

论题 3-1　动词团块处理
论题 15　论 S04 句类及其 BC 构成
论题 16　论复杂特征集和句类代码
论题 17　论连动句和兼语句
论题 18　一论句类转换
论题 19　再论句类转换
论题 20　三论句类转换
论题 21　二论中西语言的基本差异
论题 27　论块扩处理
论题 28　论句蜕处理
论题 29　论新词辨识
论题 30　论伪词辨识
论题 32　句类知识运用
论题 35　本体层与挂靠层
论题 36　概念节点之高、中、底层
论题 37　块内同行优先
论题 38　块间同行优先
论题 39　人类活动表述之层次表示
论题 40　论格式美
论题 41　论短时记忆
论题 42　论作用效应
论题 43　论过程转移
论题 44　论关系
论题 45　论状态
论题 46　论比较判断
论题 47　论基本判断
论题 48　论 lv 概念
论题 49　论 rw 概念
论题 50　论交式关联
论题 51　论链式关联
论题 52　论 HNC 技术的第一期目标——代小结

四、HNC 理解处理问答（简称〈问答〉系列，写于 1993 年冬至 1994 年夏）　系统阐述了 HNC 理论对自然语言理解处理的与众不同的思路和方案。对了解 HNC 艰辛探索的曲折历程提供了生动的写照。

五、语义学日记选录(简称〈日记〉,写于 1994 年冬至 1995 年春末) 通过对概念网络节点及其反映射汉字的阐述,剖析了各局部联想脉络的内部结构和外部连接。通过这些叙述片段,可以窥见 HNC 庞大符号体系的精妙所在。

附录部分收录了黄曾阳先生的两封学术函件和他的同事或学生有关 HNC 理论的论文,是对正文部分的有力补充,涉及以下三方面的内容:句类分析的技术实现;对 HNC 的阐发和评介;HNC 知识库的建设。

关于正文部分的更为详细的概括和说明,可参阅黄曾阳先生为本书写的"后记"。对〈论文〉序列和〈论题〉序列中未收入本书的部分感兴趣的读者,请访问以下网址:http://farad.ioa.ac.cn/hzy.html。

正文各个部分是黄曾阳先生在不同时期撰写的,其中较早的论述并没有根据 HNC 理论后来的发展加以修改,而是保持了原貌,因此前后有一些不同之处。这些不同,特别是 HNC 符号体系的调整,会给读者带来一定的困难,但它们体现了 HNC 理论的发展历程,这对未来的探索具有重要的参考价值和借鉴意义。

关于 HNC 符号体系的调整,请读者在阅读中注意有关的注释和说明,这里补充说明两点:1."语法"类概念 f 和综合类概念 s 早期没有独立出来,而是分别放在基元概念 7 行的 3 13 号节点和 12 行的 9 12 号节点;2. 概念高、中、底层符号之间的分界早期有明确标记,后来改用内涵约定的方式表达,这使得概念符号不易读懂,建议读者对此采取"不求甚解"的方法,只关注概念的高层即可。另外,文中的方头括号【】表示引用的是〈论文〉系列中的文章,如【2】代表"论文 2"。

本书中的许多内容可能需要反复阅读才能透彻理解,这有三方面的原因:其一,HNC 在很大程度上是思路全新的理论;其二,为表达全新思想,该理论定义了一系列必要的新术语;其三,黄曾阳先生习惯于理论思考,其论述大多惜墨如金,十分精练。反复阅读需要付出较多的时间,然而一定会从中得到丰厚的回报。

凡研究或关心与自然语言有关的科学问题的读者都将从本书中获益匪浅。

苗传江

1998 年 9 月 21 日
于北京语言文化大学
语言信息处理研究所

致　　谢

感谢清华大学出版社领导的支持和胆识，使这个本来不敢见“公婆面”的“丑媳妇”走上了面向社会之路。感谢本书的责任编辑薛慧女士为本书的出版做了大量精细的工作。感谢中国中文信息学会领导，特别是学会秘书长曹右琦女士对本书出版的热情鼓励和大力支持。感谢林杏光教授的当机立断和强有力的组织安排，使这部 70 多万字规模的专著竟然在一个月的时间里如期定稿。感谢陈力为院士的题词，他又一次在关键时刻对 HNC 给予了关键性的支持。感谢林杏光教授为本书精心撰写的“编者的话”。感谢 HNC 联合攻关组一年半以来令人难以置信的工作热忱和巨大成果，没有这个后盾，我是不敢出这本书的。感谢徐为方教授在她的本职工作忙重之余，为本书承担了校审工作。感谢儿子元敬昼夜加班承担了文稿的录入、打印和形成最终文件的工作。最后，还要对我在海内外的一些学生和战友说，感谢你们在 HNC 艰苦探索时期所奉献的才华和精力，我永远想念你们。

黄曾阳

1998 年 9 月 8 日

于中国科学院声学研究所

目　录

第 一 部 分

HNC 理 论 概 要

HNC 理论概要*

HNC 是‘Hierarchical Network of Concepts（概念层次网络）’的简称，它是面向整个自然语言理解的理论框架。这个理论框架是以语义表达为基础的，它对语义的表达是概念化、层次化、网络化的，所以称它为概念层次网络理论。

1 HNC 理论的形成

自然语言处理作为人工智能的一个分支，已有 40 年的发展历程，形成了计算语言学这一跨接语言、信息、认知科学和计算机技术的边缘学科。它的发展主要围绕以下三个方面：

1. 自然语言的表述和处理模式；
2. 自然语言知识的表示、获取和学习；
3. 研制开发自然语言的应用系统。

在自然语言的表述和处理模式方面，源于印欧语系的语法学和句法分析一直居于主导地位。八大词类、六种句子成分、短语结构和句法树成为语言分析的基本概念和依托。对于这一传统分析模式，仅在 20 世纪 70 年代，曾一度受到菲尔墨（Fillmore）和山克（Schank）的质疑和挑战。80 年代以来，语料库语言学的兴起使人们对统计模式产生了过高的期望，以致忽视了菲-山挑战的实质意义。

自然语言传统分析模式（含统计模式）的根本弱点何在？一言以蔽之，它不是描述语言感知过程的适当模式。

面对语音流的五重模糊（发音模糊、音词转换模糊、词的多义模糊、语义块构成的分合模糊、指代冗缺模糊），面对文字流的后三重模糊，大脑的语言感知应付裕如，表现了强大的解模糊能力，自然语言处理技术当前无从望其项背。

近 20 年来，自然语言处理囿于传统模式，不图突破。但是，它所面临的所有重大课题，从音词转换到机器翻译，从全文检索、信息抽取到智能阅读助手，都在呼唤语言表述及处理新模式的诞生；呼唤上下文联想处理向‘知其所以然’的语义理解前进；呼唤向语言感知的方向靠拢。随着网络时代的来临，这一呼唤的迫切性和严峻性在与日俱增。

响应这一呼唤才意味着真正的突破，但突破的契机何在？悲观论者认为：语言感知过程

---

* 本文发表于《中文信息学报》，Vol.11，No.4，1997。发表时该刊加有主编按语“HNC 理论概要”的作者黄曾阳先生创立的面向整个自然语言理解的理论框架，在语义表达上有自己的特色，在语义处理上走了一条新路。鉴于汉语语法研究尚有诸多困惑，HNC 理论所走的以语义表达为基础的新路子对突破汉语理解问题尤其有实际意义。

密切依附于大脑中万亿神经元的神经网络,依附于浩瀚无垠的世界知识海洋,在对这个“网络”和“海洋”的奥秘未作充分揭示之前,模拟语言感知过程是不现实的。

事情果真是如此悲观么?HNC 理论对此进行了近 8 年的探索,结论是,突破的契机是存在的,其要点是:

1. 要把自然语言所表述的知识划分为概念、语言和常识三个独立的层面,对不同层面采取不同的知识表示策略和学习方式,形成各自的知识库系统。

2. 建立网络式概念基元符号体系,即概念表述的数学表示式。这个符号体系或表示式应具有语义完备性,能够与自然语言的词语建立起语义映射关系,同时,它必须是高度数字化的,每一个符号基元(每个字母或数字)都具有确定的意义,可充当概念联想的激活因子。这个符号体系就是下文将要详细介绍的三大语义网络及五元组等,它是计算机把握并理解语言概念的基本前提。

3. 建立语句的语义表述模式,即语句表述的数学表示式。这一模式的完备性应表现为可表述自然语言任何语句的语义结构,即乔姆斯基所提出的语言深层结构。这个深层结构就是下文将要简要介绍的句类格式。以句类格式为基点的语句分析叫做句类分析,是对大脑语言感知过程的初步模拟,在上述五重模糊或三重模糊的消解方面,理论上,句类分析应能接近甚至超过常人的水准。

上述三点是形成 HNC 理论的基本背景。

但是,解模糊处理仅仅是自然语言理解的万里征途的第一步,仅涉及 HNC 理解处理系统(本文第三部分有简略介绍)的部分模块。作为自然语言的一种表述和处理模式,HNC 是开放的,并处于不断完善和深化的过程,在这一过程中,更需要不同学科的合作,特别是信息处理与语言学的合作,在 8 年的艰苦探索过程中作者深深感到这一合作的迫切性。现在这一合作的势态已初步形成,正是在合作者的鼓励和具体推动下(林杏光 1997),HNC 理论首次公开发表论文,主要目的在于扩大这一合作的势态。

2 HNC 理论的基本内容

人对语言的理解本质上是一种认知行为,如果能描述大脑认知结构的具体模式,计算机就可以运用这些模式对自然语言进行理解处理。我们把认知结构分为局部和全局两类联想脉络,认为对联想脉络的表述是语言深层(即语言的语义层面)的根本问题。什么是局部联想和全局联想呢?简单地说,局部联想是指词汇层面的联想,全局联想是指语句及篇章层面的联想。更简单地说,理解句子有两种思路:一是从组成句子的词语入手,一是从句子的整体结构和上下文语境入手,前者就是局部联想,后者就是全局联想。当然,人在理解句子的时候,这两种联想不是截然分开的,而是并存的、相互作用的,计算机理解语言也应该综合运用这两类联想脉络。HNC 的出发点就是通过建立两类联想脉络来“帮助”计算机理解自然语言。下面就分别介绍 HNC 建立的两类联想脉络。

2.1 局部联想脉络——五元组和语义网络

局部联想是词汇层面的联想,自然语言的词汇是用来表达概念的,因此,HNC 建立的局部联想脉络体现为一个概念表述体系,这个概念表述体系可以简单概括如下:把概念分为抽象概念和具体概念,对抽象概念用五元组和语义网络来表达,对具体概念采取挂靠展开近似表达方法。

概念有抽象与具体之分。在一般人看来,抽象概念总是比具体概念难于把握,中文信息处理界已有的汉语语义分类系统,其内容主要是对比较容易把握的具体概念的分类,这样的语义分类系统没有摆脱对客观事物进行科学分类的束缚,对抽象概念则几乎束手无策。实际上,从深层来讲,抽象概念比具体概念更具有基元性、系统性,更容易表达;具体概念是客观存在物在人的思维中的一种直接反映,它里面包含了许多世界知识,而对世界知识是很难进行详尽表达的。所幸的是,人对具体概念理解和认识的深度可以比抽象概念浅,所以可以采取实用原则:"不求甚解"。HNC 理论侧重于抽象概念的表达。

HNC 理论通过五元组和语义网络层次符号来完整地表达抽象概念,前者表达抽象概念的外在表现,后者表达抽象概念的内涵。

任何一个概念都需要从不同侧面予以表达,这种现象叫做概念的多元性表现。具体概念的多元性表现十分复杂,难以给出规范化的表达,抽象概念则有所不同,它的多元性表现在自然语言中有明显的迹象,这就是词性现象。印欧语系的词根或具有词根特色的词,可以加上不同的后缀分别构成动词、名词、形容词和副词,这种词性的转换就是抽象概念多元性的生动表现,也就是说,词根相同词性不同的词是对同一概念不同侧面的表达。汉语对抽象概念的多元性表现则没有相应的形式标志,而往往是同一个词兼有名词、动词、形容词、副词中的几个属性。汉语的词性模糊现象(即无形态变化)和西语以形态变化表现不同词性的现象都是抽象概念多元性的生动表现,形态变化的有无只是一种形式,本质在于抽象概念本身具有这种多元性表现的固有特征。

那么,抽象概念多元性表现的"多"是一个模糊的"多",还是一个确定的"多"?或者说,能否给以规范化的表达?或者再换一个说法,这个多元性表现的"多"是否存在某些基元(primitive)呢?答案是肯定的。抽象概念需要从动态、静态、属性、值和效应五个侧面加以表达,这就是抽象概念的五元组特性,简记为:(v, g, u, z, r)特性,它们是抽象概念多元性表现的基元。任何抽象概念都具有五元组特性,即都需要从五个侧面加以表达,不过,对某个抽象概念各个侧面的表达,自然语言中未必都有相应的词语,而且不同语种间存在着差别。反过来,自然语言中的一个表达抽象概念的词语必定是从五元组中的某个或某几个侧面来表达某个抽象概念。例如:"思考、思维、想法"就是分别从五元组的 vg, g, r 侧面对同一概念内涵的表达。五元组是词性的本质内容,是词性的基元。所以,不必为汉语词汇的大量兼类现象感到困惑。

为表达抽象概念的内涵,HNC 设计了三大语义网络:基元概念语义网络、基本概念语义

网络和逻辑概念语义网络。语义网络是树状的分层结构,每一层的若干节点分别用数字来表示,网络中的任一个节点都可以通过从最高层开始、到该节点结束的一串数字唯一地确定,这个数字串叫做层次符号。三大语义网络是抽象概念的三大聚类。

基元概念语义网络的一级节点分为两大类:一类是主体基元概念,另一类是复合基元概念。

主体基元概念有 6 个一级节点,分别是作用、过程、转移、效应、关系、状态,它们构成作用效应链。什么是作用效应链?作用效应链反映一切事物的最大共性。作用存在于一切事物的内部和相互之间,作用必然产生某种效应,在达到最终效应之前,必然伴随着某种过程或转移,在达到最终效应之后,必然出现新的关系或状态。过程、转移、关系和状态也是效应的一种表现形式。新的效应又会引发新的作用,如此循环往复,以至无穷,这就是宇宙间一切事物存在和发展的基本法则,也是语言表达和概念推理的基本法则。

这 6 个环节的源头是作用,结果是效应。自然语言的主要内容就是对这六个环节进行局部和总体的具体表述,我们对句类(见下文)的划分就是以此为标准的(这里顺便说明一下,山克的"概念从属理论"主要考虑了"转移"这一个环节,我们对"转移"二级节点的设计就部分吸收了"概念从属理论"的主要结果)。作用效应链既是用于表达概念的语义网络的核心,又是划分句类的标准,换句话说,它既是局部联想脉络的基础,又是全局联想脉络的基础,两个联想脉络通过它联系起来,所以,在一定意义上可以说作用效应链是 HNC 的理论基础。

复合基元概念主要涉及人类活动,这是因为,自然语言是人类的交际工具,其主要表述对象是人类活动而不是自然现象。复合基元概念总共设置了 8 个一级概念节点,根据人类活动的语境特征划分为三个层次,即生理本能活动、一般理智活动和社会性活动。

基本概念语义网络共有 9 个一级节点:序及广义空间、时间、空间、数、量与范围、质与类、度、客观的基本属性、含主观评价的基本属性。

逻辑概念语义网络分为两类:一类是语言逻辑概念,大体上相应于汉语的虚词,有 11 个一级概念节点,分为语义块区分标志符、语义块组合标志符、语义块及句间关系说明符三类。这 11 个一级节点的划分主要基于它们对语义块感知及句类辨识的作用,而不是它们的语法特性。另一类是基本逻辑概念,有两个一级概念节点:比较和基本判断。

HNC 语义网络的设计思想有两个来源:一是奎廉(Quillian)的语义网络、菲尔墨的格语法和山克的概念从属理论;二是汉语的"字义基元化,词义组合化"现象。第一个来源提出了"语义基元"的杰出思想并暗含着"总体表述"的宏伟目标,第二个来源则提供了语义基元的宝贵原料。汉语字少词多,仅用几千个汉字加以组合就构成许多的词。几千年来,汉语随着社会的发展而发展,新词不断增加,但组成词语的汉字却几千年很少变化。汉字字义的基元化和汉语词义的组合化是一个伟大的宝藏,HNC 语义网络的形成深深受益于这一宝藏的启发。

三大语义网络为表达抽象概念的内涵而设计,最终将用它来描写自然语言词汇的语义,

但网络本身却不是直接面向语言词汇的,而是面向构成词汇语义的概念基元的,适用于任何语种。网络上的任何节点本身都是概念,但这些概念只是庞大的概念海洋里的“元素”,即它们是概念基元,它们通过不同方式的组合而构成各种各样的、无数的概念,HNC 定义了 8 种组合结构,用以表达复合概念。

三大语义网络的设计,可以解决现代语义学中的两个难题。一是义素分析法的难题。义素分析法试图用分解的方法、用义素(语义原子)来描述词汇语义,它对一些词的意义进行了成功的描写,但是,语言的义素到底有多少,义素分析法没找到答案,因而不能落实到对全部语言词汇的描写中。三大语义网络的各个节点,即概念基元,大体上相当于义素,可以用来描写任何语言的所有词汇的语义。语义网络采用了分层的灵活结构,可以从高层到底层根据需要不断往下设置节点,而由于有上层的控制又不会零乱,从而解决了义素分析法的难题。二是语义场的难题。语义场理论看到了词汇语义的关联性和系统性,但是,语言中到底有多少义场,义场该怎样划分,义场之间、义场内部都是怎样的关系,对这些问题语义场理论都没能解答。三大语义网络建立了语言深层概念的网络,它是一个整体的设计,是一个完整的系统,它各个节点下的网络都形成相关联的概念的聚类,这些聚类就相当于语义场。更重要的是,通过语义网络,义场内部、义场之间都建立了联系,而且这各种各样的联系都可以通过层次符号显式地表达出来,从而使计算机能够掌握和操作。

五元组符号和语义网络的层次符号的适当组合可以实现对抽象概念的完整表达。这种表达方式能够显式地表达出自然语言概念之间的关联性,从而有助于计算机把握和理解。例如,“精神 - 振奋、无私 - 奉献、慷慨 - 就义、锦绣 - 山河、远大 - 前程、承担 - 责任、召开 - 会议”这些词语间的优先搭配在自然语言中是“理所当然”的,把这些搭配中的词用五元组和层次符号表示,各个搭配中的前后词语就会具有相同或相近的层次符号,而只是五元组符号不同,从而使它们之间搭配的“理所当然”得到显式的体现。可见,用五元组和语义网络层次符号表达语言概念的方法可以解决语义搭配(或称语义约束)的难题。传统的词性搭配不能解决语义问题,动词后可与名词搭配,但“动 + 名”结构根本无法保证语义的正确,这种语法正确、语义荒谬的困难必须借助语义约束来解决,但语义约束一直找不到表达和把握的手段。三大语义网络完成了概念之间关联性的设计,找到了解决语义约束问题的根本途径。

对概念关联性的表达是语义网络的首要目标。概念基元的首要价值与其说是给出复合概念的精确表示,不如说是给出概念关联性知识和联想脉络的线索。自然语言理解的中心任务是解模糊,如同音模糊消解、一词多义模糊消解等,这些模糊的消解统称为多义选一处理。对自然语言词汇的多义选一处理是人类理解自然语言过程中最频繁、最基本的操作。对这一操作过程的形式模拟不在于并行处理或快速计算,而在于以什么巧妙的方式完成大量语义距离(语义关联性)的计算。层次符号的构造方式把最频繁、最基本的语义距离计算变成了对层次符号的简单逐层比较。这是 HNC 用语义网络层次符号表达概念的基本出发点。层次符号是一种灵活的分层结构,它到任一层都代表一个概念,至于这个(些)概念与相应的语言概念之间,究竟谁是谁的近似,已无关紧要。重要的是,层次网络符号对概念的局

部联想脉络给出了明确的表示,便于计算机把握概念之间的关联性。

语义网络层次符号的设计为计算机理解自然语言的语义提供了有力的手段。当然,在工程实现上首先要用语义网络层次符号完成对自然语言词汇语义的描写,这是一项浩大而艰巨的工程,但这个瓶颈问题跟过去相比已有了本质的不同,过去缺乏语义描写的完备手段,现在手段已备,剩下的只是工作量的问题。

下面简单说明对具体概念的表达。

一般来说,具体概念的精确表达要比抽象概念困难得多,因为它涉及到许多世界知识,这些世界知识是人类认识积累的结果。但另一方面,人在理解自然语言过程中对具体概念的认识深度可以比抽象概念浅得多,天生的盲人仍能同常人一样掌握自然语言,道理就在这里。所以,对具体概念的表达,应采取大胆近似的方案,这是对具体概念进行层次符号设计的基本出发点。HNC 用‘类别符号 + 挂靠’的方式近似地表达具体概念。

具体概念的类别,从语言表达的角度来看,先分为物、人、物性三类(分别用符号 w, p, x 表示)比较合理。物有自然物与人工物之分,人工物又有现代与传统、物质与精神产品之分,当然还可以有各种各样的分类标准。人和物性也同样存在子类划分问题。在处理具体概念的分类问题时,不宜照搬自然科学的分类方法,HNC 的着眼点主要是引起概念的联想,而不是分类的科学性。

对具体概念的内涵,HNC 采用向抽象概念的基元概念和基本概念挂靠的方法表达。例如,人、一般人工物、现代产品这几类具体概念分别用符号 p、pw、w9 表示,基元概念里的概念节点 22b 表示自身转移,那么,向它挂靠的 pw22b 就表示交通工具;219 表示针对性接收,w9219 就表示现代探测设备;411 表示结合,p411 就表示夫妻;382 表示废弃,pw382 就表示垃圾;基本概念里的概念节点 711 和 712 分别表示正和负,p711 和 p712 就分别表示男人和女人。显然,这种挂靠的表示方式都是很粗糙的近似,但其重要意义在于:通过这一近似表示,计算机就能对有关概念之间的关联性有所‘领会’。挂靠式表示方式的目的,就是在具体概念与抽象概念之间建立一种关联,并把这种关联用符号显式地表示出来,以利于计算机计算语义距离。

挂靠的表示方式只适用于一部分具体概念,一些基本的物质概念仍然需要进行独立的层次符号设计。为此,我们设计了一个基本物的语义网络,这个网络有 7 个一级节点:热、光、声、电磁、微观基本物、宏观基本物和生命体。这些节点的设置仍是服务于联想脉络的建立,并不完全遵循自然科学的标准。

按照上述设计,对概念基元就可以写出下面的语义表示式:

$$F = \Sigma(\text{字母串})(\text{数字串})$$

F 代表概念基元的 HNC 符号。字母串由概念类别符号(ϕ, j, l, jl, p, w, x)构成,数字串由 16 进制数字的 0 d 构成。其中 ϕ 表示基元概念,j 表示基本概念,l 表示语言逻辑概念,jl 表示基本逻辑概念。

复合概念的语义表示式为:

$$F = \Sigma F_k$$

F_k 之间的连接通过 8 种概念组合结构符号来表示。

2.2 全局联想脉络——语义块和句类

全局联想脉络是语句及篇章层面的联想。语义块和句类理论是在语句层面设计的全局联想脉络,篇章层面的联想脉络本文暂不介绍。

简单地说,语义块是句子的语义构成单位,形式上可以是一个词、一个短语或者一个句子。语义块类似于传统语言学中的短语,但是,两者具有本质的区别,表现在:第一,从内涵上来看,语义块是语义,即语言深层的定义,短语则是语法,即语言表层的定义;第二,从形式上来看,语义块可包含或嵌套另外的一个甚至多个语句,而短语不能。另外,传统的短语更多的是被看作词的组合结构,而不是句子的直接构成单位。

语义块这一概念的提出是为了便于从语言深层(即语义层面)描述一个句子。用词或短语描述句子,无法清楚地界定一个句子是否完备,如果问一个句子应该或者可能有多少个词或短语,便难以回答。但有了语义块的概念,就可以明确回答一个句子有多少语义块以及每个语义块的类型等问题。

在通常情况下,一个语义块包含核心部分和说明部分。语义块按其语义功能分类,语义块的语义功能主要取决于其核心部分。

语义块分为主语义块和辅语义块两大类。主和辅是从句义表达的角度划分的,主语义块是句义的‘必不可少’的成分,辅语义块是句义的‘可有可无’的成分。主语义块有 4 种:特征 E、作用者 A、对象 B 和内容 C。辅语义块有 7 种:条件、手段、工具、途径、参照、因、果。

E,A,B,C 四大主语义块划分的理论依据是:一个语句表达的内容无非是两个方面,一是表达对象,二是对象的表现,前者是‘什么’,后者是‘怎么样’。作用者 A、对象 B 语义块是表达对象,内容 C、特征 E 语义块是表现。在表达对象中,B 是一般表达对象,A 是表达对象中的特殊对象;在表现中,E 是一般表现,C 是特殊表现。一个句子至少由一个对象语义块和一个表现语义块构成,但更为常见的结构是:两个对象语义块加一个表现语义块,一个对象语义块加两个表现语义块,两个对象语义块加两个表现语义块,还可以是多个对象语义块加多个表现语义块。所以,所谓‘一个句子只有一个中心动词’的语法规范与语言表达的需要并不协调。

为什么 E 语义块叫特征语义块呢?因为一个句子的基本语义信息就蕴涵在 E 语义块中。那么,什么是基本语义信息呢?它来源于作用效应链思想。一个句子总是对作用效应链的某一或某些环节的表达,所谓一个句子的基本语义信息就是指它所表达的关于作用效应链的某一或某些环节的信息。这样,作用效应链的 6 个环节自然就是基本语义信息的分类标准,因而也是 E 语义块的分类标准。不同类别的 E 语义块构成不同类别的句子,从而引入了句类的概念。HNC 的句类是句子的语义类别,与传统的句类是完全不同的概念,后者指陈述句、祈使句、疑问句和感叹句,基本上是句子的语用分类。

只表达作用效应链的一个环节的句类称为基本句类,表达两个或多个环节的句类称为混合句类。E 语义块的命名与作用效应链 6 个环节的名称相一致,即作用、过程、转移、效应、关系、状态。由这些 E 语义块构成的句子,分别命名为作用句、过程句、转移句、效应句、关系句和状态句。

E 语义块的核心部分一定是动词,而且,不同类别 E 语义块的动词来源于不同的基元概念。E 语义块的分类标准,也就是句类的分类标准。这个标准是与三大语义网络密切关联的,它实际上也就是 HNC 概念层次网络符号体系设计的基本准则之一。这样,E 语义块的辨识信息,或者说句类的辨识,就明确无误地蕴涵在概念的层次网络符号体系之中。

由于判断是人类思维活动的基本内容,也是语言表达的基本内容之一,我们据此又定义了一个句类:判断句。根据作用效应链定义的 6 个句类加上判断句,构成 HNC 的 7 个基本句类。每一个基本句类又分为若干个子类,子类的定义与相应基元概念网络的二级节点相对应。子类之下还可以再分子类。

基本句类可以构成混合句类。所谓混合句类,是指两个以上的基本句类在一个句子中共现,诸如作用效应句、过程转移句、状态判断句等。自然语言的句子是丰富的、复杂的,但它们表达的信息总是由 7 个基本句类组成的,这正是基本句类之所以称为"基本"的原因。在自然语言中,基本句类的混合往往(或者说主要)是两两混合,因此,混合句类理论上应有 $6 \times 5 + 6 = 36$ 个。"6×5"是与作用效应链相对应的 6 个基本句类的两两混合;"$+6$"是它们与判断句的混合。

上面说明了语义块和句类的概念,它们之间是什么关系呢?一句话:语义块是句类的函数。这是 HNC 语义块和句类理论的基本论点。

E,A,B,C 四种主语义块是抽象概括的结果,它们在一个句子中的有无、个数和具体内涵随句类的不同而不同。这就是"语义块是句类的函数"所概括的内容。例如,拿作用者语义块 A 来说,作用句中的 A 语义块是"产生影响者",类似于一般所说的施事,而转移句中的 A 语义块是转移的发出者,过程句、关系句和状态句中则不涉及 A 语义块。再如对象语义块 B,作用句和效应句中的 B 语义块是"被影响者"或"接受者",类似于一般所说的受事,过程句、关系句和状态句中的 B 语义块是过程、关系、状态的体现者或承受者,而关系的体现者显然有两个,即关系的双方,它们都是 B 语义块,彼此之间不存在施事和受事的关系。在转移句中,B 语义块是转移的接收者,而转移"物"则是转移的内容,即 C 语义块。

我们把"语义块是句类的函数"具体体现为句类格式。句类格式是指一个句子的主语义块的排列顺序,例如作用句必须有三个主语义块:作用者 A、作用 X(即 E 语义块)和作用的对象 B,三者的排列顺序不外乎 6 种:A + X + B,B + X + A,B + A + X,A + B + X,X + A + B,X + B + A。选择这 6 种格式中的哪一种作为标准格式,不同语种间存在着差别,比如汉语和多数印欧语都采用第一种格式。标准格式中蕴涵着主语义块类别的辨识信息。

7 种基本句类和 36 种混合句类的提出为语句深层结构的表达提供了简明而完备的手段,所谓深层结构就有了计算机可操作的数学表示式。例如:

作用句：XJ = A + X + B　　张三打断了李四的腿。

过程句：PJ = PB + P　　李四的腿伤大有好转。

转移句：TJ = TA + T + TB + TC　　李四的朋友电告李四父母这个好消息。

效应句：YJ = YB + Y + YC；YBC + Y　　李四养好了腿伤。|李四的腿伤养好了。

关系句：RJ = RB1 + R + RB2；RB + R　　张三失去了他多年的女友。|张三跟他多年的女友吹了。

状态句：SJ = SB + S + SC；SB + S；SB + SC　　张三穿着皮大衣。|张三升官了。|张小姐很漂亮。

判断句：DJ = DA + D + DBC　　张三认为李四不该那样做。

反应句(作用句的子类)：

X2J = X2B + X2 + XBC + (X2C)　　张先生怕李小姐发脾气。

基本状态句(状态句的子类)：

S00J = SB + S00 + SC；SC + S00 + SB　　主席团坐在台上。|台上坐着主席团。

作用关系句(混合句类)：

XRJ = A + XR + RB　　张三挑拨李四和我的关系。

关系作用句(混合句类)：

RXJ = RB1 + RX + B　　张三多次帮助过李四。

这些表示式就是计算机赖以进行语句联想操作的基础。表示式中的每一项代表一个主语义块,这些主语义块的语义角色由该项的命名符号所唯一确定,它们是引发全局联想脉络的激活因子。

EABC 语义块在形式上似乎与传统语言学的主谓宾补相对应,其实它们是完全不同的概念,有着本质的区别：EABC 是语义层面的概念,是语言深层的描述量,它们是句类的函数,但与句子的格式无关；主谓宾补是语法层面的概念,是语言表层的描述量,它们与句类无关,但与句子的格式息息相关。EABC 语义块和主谓宾补是从不同层面或角度对句子的结构提出分析的模式,不能相互代替。

最后,简单叙述一下 EABC 概念的形成过程,这对于加深对这一概念的理解或许有所裨益。与主谓宾补相联系,语法学还有动词的及物和不及物以及双宾语等概念。但及物性的具体表现,仅在语法层面进行研究十分困难,它涉及宾语的分类问题,有的及物动词要求双宾语,有的不仅要求宾语,还要求补语。这些问题都必须进入语义层面,才能给出明确的答案。从理解来说,仅有及物的概念是远远不够的,重要的是：它“及”什么样的“物”? 开始的时候,曾以为这只是词汇层面的特性,后来才发现不是这样,它也是概念层面的重要特性,这一发现导致“语义块是句类函数”概念的形成。但应该说,是格语法理论的创立者菲尔墨最先想到了这一点,他是对宾语和主语进行语义分类的第一位先行者。可惜他的理论匆忙出台,在理论的总体性和层次性方面都比较欠缺。现在看来,主语和宾语的语义分类必须用 ABC 函数的概念,即将语义块作为句类的函数来处理才能给出完善的表述。至于双宾语,它

一定是转移型概念,而同时要求宾语和补语的一定是作用效应型概念。

3 HNC 理论的实现

上文介绍的两个联想脉络是 HNC 理论的基础部分,它的另一部分内容是自然语言理解的框架和具体实践。

HNC 理论走向应用的第一步是语义块感知和句类辨识。语义块感知就是找出一个句子中的各个语义块,句类辨识就是通过感知得到一个句子的 E 语义块,进而确定这个句子所属的句类。计算机能否感知到语义块关系到 HNC 能否指导实践、是否有应用价值的问题,张全的博士论文(张全,1996)对此做了肯定的回答。感知到语义块、辨识出句类以后,就可以运用句类知识对句子进行理解处理,这称为句类分析。在句类分析过程中,句类知识起着全局性的指导作用,主要有四方面的知识:一是句类格式知识,二是语义块构成知识,三是语义块之间的概念关联知识,四是语义块和句类的转换知识。语义块感知和句类辨识主要运用局部联想脉络,句类分析主要运用全局联想脉络,当然,处理过程中对这两个联想脉络的运用不是截然分开的。

以句类分析为基础,HNC 设计了自然语言处理系统的基本框架,这个框架由 9 个模块组成:1. 单音词感知模块;2. 语义块感知模块;3. 句类分析模块;4. 合理性分析模块;5. 短时记忆知识模块;6. 语境生成模块;7. 隐藏知识揭示模块;8. 要点主题分析模块;9. 短时记忆向长时间记忆扩展的模块。目前,部分模块已在计算机上得到实现。

自然语言处理离不开知识库,对知识库的设计和建立也是 HNC 理论的重要组成部分。人工智能早期一系列的挫折,使人们认识到知识的重要性。要使计算机表现出智能,唯一的办法就是使它拥有并运用知识。正是这一认识促成了 20 世纪 70 年代到 80 年代的"专家系统热",并取得了引人注目的成就。但这些专家系统的知识,都是局限于特定的领域,而一般自然语言理解(不包括特定领域的简单语言应用系统)所需要的知识则完全不同于通常的专家系统。它需要各种各样的知识,但可分为三大类:概念知识、语言知识、常识及专业知识。前两类知识的本质区别在于:语言知识与具体语种有关,而概念知识与语种无关。把概念知识从语言知识中独立出来是势在必然的发展。把常识及专业知识独立出来对知识库的建立是非常方便和有利的,这一点不言而喻。我们已经建立了比较完备的概念知识库,目前正在紧张地进行汉语语言知识库的建立。我们曾建立过地理知识库,使用效果很好,所以具有建立常识及专业知识库的成功经验,但常识及专业知识库的建立目前还不是自然语言理解处理的迫切任务。

自然语言理解处理的进展必须由信息处理工作者和语言研究者共同推动。令人高兴的是,在我国计算语言学前辈的推动下,这两方面力量开展联合研究的局面已开始形成,并初步组成了联合攻关的队伍。

主要参考文献

黄曾阳. 1996. HNC 理解处理论文选录. 中国科学院声学研究所声场声信息国家重点实验室自然语言理解课题组

林杏光. 1997. 正确引导汉语理解与汉语研究——事关人工智能开发的一个重要前提. 科技导报, 1997 (4)

苗传江. 1997. HNC 理论的基本内容. 中科院声学所‘HNC 知识库培训班’教材

张全. 1996. 基于 HNC 理论的语义块感知处理. 中国科学院声学所博士学位论文

Chomsky N. 1957. Syntactic Structures. Hague: Mouton

Chomsky N. 1965. Aspects of the Theory of Syntax. Cambridge, MA: MIT Press

Fillmore C J. 1968. The case for case. In: Bach E, Harms R eds. Universals in Linguistic Theory. New York: Holt, Rinehart and Winston

Qullian M R. 1968. Semantic memory. In: Minsky M Ed. Semantic Information Processing. Cambridge, MA: MIT Press

Schank R. 1973. Identification of conceptualizations underlying natural language. In: Schank R, Colby K Eds. Computer Models of Thought and Language. San Francisco, CA: W H Freeman and Company

Schank R. 1975a. Conceptual Information Processing. Amsterdam: North Holland

Schank R. 1975b. The structure of episodes in memory. In: Bobrow D, Collins A eds. Representation and Understanding. New York: Academic Press

Schank R. 1982. Dynamic Memory. New York: Cambridge University Press

Schank R, Abelson R. 1977. Scripts, Plans, Goals and Understanding. Hillsdale, NJ: Erlbaum

第 二 部 分

HNC 理解处理论文选录

自然语言语义网络的基本构成及其特性

引　　言

语义网络作为一种知识表示方式,早已为人们所熟知。但是,对自然语言语义网络的基本构成及其特性,似乎需要一个总体性的描述,而这样的描述还很欠缺。这一描述的目的在于阐发大脑认知结构(cognitive structure)的具体模式,以期有助于计算机运用这些模式,在语义层面进行自然语言的理解处理。

本篇及其姊妹篇【2】(注:方头括号【 】表示引用的是〈论文〉系列中的文章,【2】表示"论文 2",下同。〈论文〉系列亦称 Paper 系列。)是对自然语言语义网络进行总体性描述的理论尝试。

这个总体性描述理论将命名为概念层次网络理论,英文是 Hierarchical Network of Concepts,简称 HNC 理论。

这个理论将认知结构先分为局部和全局两类联想脉络。

局部联想脉络是本文的主题,全局联想脉络是【2】的主题。

建立两类联想脉络的出发点是试图形成一种预期及判断能力,以便计算机能够实行一种"自下而上"(bottom - up)与"由上而下"(top - down)相结合的理解处理模式。

我们将把这种处理方式称为句类分析。

句类分析是语义层面理解处理的核心模块。这将在【3】中作系统说明。

句类分析的知识基础包括概念层面、词汇层面、语句层面、语境层面的知识。这四个层面的知识应以语句层面为中心,并命名为句类知识。这类知识将在【14】到【21】中分专题讨论。概念层面和词汇层面的知识则在【6】【7】【8】中讨论。语境层面的知识与常识性知识最为密切,因而难以给出系统的描述,仅在【3】中予以初步说明。

上述四个层面的知识,从整个知识的海洋来看,仍不过是"沧海一粟"。基于这些知识的理解处理在自然语言理解的万里征途中仅仅是向前迈出了一小步。从应用的角度来说,这一小步的具体体现是能够模拟大脑语言感知过程对语句的理解,从而也许能在解模糊及纠错处理方面接近甚至达到人类的水平。因此,这一小步,对于"知识产业"(knowledge industry)或语言信息处理的发展也许能起到较大的推动作用。而中文信息处理则是我们首先关注的目标。

为了下面行文的方便,我们不得不引入一些新的术语,现在先对这些术语作一个简明的介绍,有些术语在后文还有详细说明。

1　五元组(v，g，u，z，r)

指抽象概念的类型特性,分别代表概念的动态、静态、属性、值和效应表达。每个抽象概念都具有这五个侧面的类型特征,也可称为抽象语言概念的形态或外在特征。

2　概念矩阵

这是五元组思想的自然推论。抽象概念的内涵和它的五元组分别构成概念矩阵的行和列。这个术语没有实质性的意义,主要是为了说明方便,将来对某一类语义网络可用‘某’行或‘某到某行’称之。

3　基本概念、基元概念、逻辑概念

这是我们对语言抽象概念的基本分类,并分别用类别符号 j，ϕ，l 予以标记。这三大类语言概念实际上就是三个超级语义网络,后文有详细说明。

4　类别符号集

除上述五元组符号 (v，g，u，z，r),三大类抽象概念的类别符号 (j，ϕ，l) 之外,还引入了下列概念类别符号:

具有三大类抽象概念综合特征的抽象概念	s
“语法”概念符号	f，q，h
具体概念物和人的符号	w，p
兼有抽象具体双重特征的物性概念	x

这 15 个类别符号专门用于表述概念的类别特征,不能用于层次符号的变量表示。它们是概念类别的基元表示,其中的基元概念符号 ϕ 在具体表达时可以省去。由这些基元符号还可以构成各种复合型概念类别。

5　层次符号集

由数字 0 到 13 构成,10 到 13 用小写字母 a，b，c，d 表示(16 进制)。

6　概念组合结构符号集

由下列符号组成:

作用、效应	#	□
对象、内容	&	|
包含	—	— 0…
挂靠结束	*	
展开	+	

偏正　　/

主谓　　‖

逻辑组合

　逻辑并　　,

　逻辑选　　;

　非　　!

　反　　∧

　括号　　( )

　一般逻辑组合　　(,1m,)

7　概念的一般表达式

Σ 类别符号串} 层次符号串} 组合结构符号} 类别符号串} 层次符号串}

其中的“类别符号串”也叫“字母串”,用类别符号集的字母表示。五元组符号一定在其他类别符号的后面。“层次符号串”也叫“数字串”。字母串代表概念的类别特征,数字串代表概念的层次性内涵,组合结构符号代表复合概念的组合结构。

8　层次符号串的两种基本形式

层次符号串 = 高层 (Σ 中层底层)

层次符号串 = (本体层) (挂靠层)

除上列术语之外,文中还会用到“语义块”、“句类”、“句类的子类”等术语,它们在【2】中有详细说明,读者可先顾名思义,仅作粗浅理解。

本文分下列 7 个题目:

1. 抽象概念的多元性表现及五元组
2. 中层层次符号的设计及概念局部联想脉络的基本特征
3. 基本概念
4. 基元概念
5. 语言逻辑概念
6. 概念的组合结构
7. 具体概念的近似表达

这个顺序安排不甚合理。按自然顺序,应先介绍基元概念,随后介绍语言逻辑概念,但基元概念与基本句类的划分密切相关,语言逻辑概念与语义块感知密切相关,对他们的表述不能不涉及到“句类”“语义块”以及赋予了特定意义的“对象”“内容”等术语,而这些术语的内涵比较复杂,在【2】中才有所阐述。所以把这两个题目安排在中间,希望有助于减轻新术语带来的困扰。但我们仍建议读者在阅读第 3 和第 4 节的同时,翻阅一下【2】的前两节。

下面转入正文的讨论。

1.1 抽象概念的多元性表现及五元组

任何一个概念都需要从不同侧面予以表达,我们将这一现象叫做概念的多元性表现。具体概念的多元性十分复杂,比较难以给出规范化的表述。抽象概念则有所不同,其多元性表现在自然语言中有明显的迹象,这就是词性现象。英语里的词根或带有词根特色的词,可配以不同的后缀分别构成动词、名词、形容词和副词。而汉语的抽象词,往往在同一含义下兼有动词、名词甚至形容词的属性,汉语的这种词性模糊现象和西语的以形态变化表现不同词性的现象都是抽象概念多元性的生动表现。西语的形态变化和汉语的不'理睬'这种形态变化(即所谓兼词性)只是一种形式,本质在于这些概念本身具有这一变化的固有特征。实际上这是所有抽象概念的共同特征,或者说是抽象概念语言表现的最大共性,它体现了实际语言概念之间的一类关联性,将称之为同行关联或同行优先,因为映射到层次网络符号,这些概念属于概念矩阵的同一行。

接下来的问题是:这个多元性表现的'多'是一个模糊的'多',还是一个确定的'多'?或者说,能否给以规范化的表述?或者再换一个说法,这个多元性表现的'多'是否存在某些基元(primitive)呢?答案是肯定的,抽象概念需要从动态、静态、属性、值和效应五个侧面加以表述,这就是抽象概念的五元组特性,简记为(v,g,u,z,r)特性,它们是抽象概念外在表现的基元。

五元组里的(v,g,u)来源于语法学,z 来源于明斯基(Minsky)的框架理论,r 来源于本文提出的'作用效应链'思想。简单的说,概念的动态性是一种广义的作用,而作用必有相应的效应。g,r 都是静态表达的名词,但分别代表因果两极。把两者区别开来,有利于概念的局部联想。现在,就来联系具体的语言概念对此作进一步的阐述。

"概念"这个词本身就是一个 r 型概念,因为它是'思考'的效应。从内涵来说,思考和概念是紧密关联的,其符号表达的数字串应该是相同或近似的,差别只应在字母串。按前述概念一般表达式,设'思考'的内涵用数字串 ik 表示,则'思考,概念'及其有关概念应分别表示为:

思考: vgik　　思维: gik
概念: rikm　　想法: rik

这些表示式给出了所列语言词汇的下述信息:"思考,思维,概念,想法"分别是同一内涵的 vg、g 和 r 型概念,汉语的'想法'和英语的 idea 一词与'思考,思维'的内涵层次性相同,但汉语的'概念'则要低一个层次。汉语的'想法'包含"概念",也就是说'想法'由一系列'概念'组成,但不能反过来说。这些表达式体现了层次网络符号体系设计的基本思想,在下一节里要专门讨论。这里再举一些以阐发 r 型概念为目的的实例。例中将不加说明地标明层次符号串的具体数字,上列'思考'等概念的 ik 及 ikm 的具体取值为'80'及'800'。

作用	gv00		
打击	vg00 # v362		
力量	gz00	力	g008
能量	zg00	能	z008
力度	z00	功能	r00
强	u00c22	弱	u00c21
效应	g30	效果	r30
效力	zg30	效率	z30
环境(自然)	w508	景象	rw508
		环境(广义)	r50
主宰	vg441	权力	rc441
运动过程	g109	轨迹	r109
演变过程	g10a	历史	r10a
智能活动	gv900	文明	r900
行为	gd0	威望	rd001
专业活动	gva0	财富	rwa0
经济活动	gva2	财产	rwa2

上列对比表明了类别符号 r 在概念表达中具有 g 不可替代的作用。从词性来说,环境与景象,运动过程与轨迹,演变过程与历史,行为与威望等都是名词,但仅用‘名词’这个概念,显然不能表明它们的两两差别。引入类别符号 r,就能清楚指明:景象是自然环境的效应物,轨迹和历史分别是运动过程和演变过程的效应,威望是行为的一种效应(其层次符号多了两层),财富是人类一般专业活动的效应物,而财产是经济活动的效应物。上列两两相关的概念,其内涵相同或近似,相应的层次符号也相同或近似。这种表达方式显然有利于计算机对概念关联性的把握,从而有利于对自然语言的理解。当然,这就需要将自然语言词汇映射成层次网络符号。我们已着手建立的自然语言知识库,就是基于这一构思而设计的。

抽象概念表达的五个基本侧面是自然语言概念关联性的基本表现之一。我们将把它和概念的层次性、对比性、对偶性、包含性(这些将在下一节说明)统称为概念同行关联。表达这些概念的具体自然语言符号可以变化万千,但对内涵或其内核相同或相似的概念,我们将赋予相同或相似的层次符号,即相同或相似的数字串。从概念矩阵来看,它们都属于概念矩阵的同一行。这样,一部分语义距离的计算问题就简化为对数字串的逐层比较问题。自然语言的词,大多数是多义的,人在言语感知过程中必须进行大量的‘多义选一’处理。也许可以说,这是交谈或阅读过程中大脑里最基础、最频繁的操作。概念矩阵的表示方式就为计算机进行类似的‘多义选一’操作提供了一个简明有效的途径和工具。

把概念表达的五个基本侧面叫做五元组,从名称来说,是借用了乔姆斯基对形式语言四元组的说法,但实际内容并无共同之处。我们的五元组,首先是指抽象概念矩阵的外在表

现,它同抽象概念的内涵分别构成概念矩阵的列和行。其次,它是指‘词性’的基元,由这些基元可组成各种各样的词性,上面我们看到了 vg, gv, gz 型概念,当然还存在其他的各种两两组合甚至三个或更多的字母组合,例如 vr, rv, vu, uv, uu, vuz 等等。关于五元组组合的意义将在【6】中作系统阐述。

1.2 中层层次符号的设计及概念局部联想脉络的基本特征

层次符号是指一般概念表达式中的数字串,用于表达概念的内涵。这个数字串又分为两种基本类型:一种是高层、中层、底层三级串接,另一种是本体层和挂靠层相互串接。当然,中层和底层还可以循环串接,两种类型也可以混合。本节只讨论第一种类型的中层层次符号设计,第二种类型放在第 7 节里讨论。而高层层次符号的设计则分别在第 3、第 4 和第 5 节中阐述。

中层层次符号主要用于表达概念局部联想脉络的三类特征。第一是概念的对比性,第二是概念的对偶性,第三是概念的包含性。

概念的这三种特性在自然语言中俯拾即是。例如,强弱,快慢,大小,长短,高低,深浅,远近,亲疏,幼、少、中、青、老,优、良、中、差,咸、淡、甜、苦,红、橙、黄、绿、青、蓝、紫……等等都是对比性概念。正反,阴阳,真假,善恶,是非,对错,上下,前后,左右,内外,东西,南北,因果,入出,问答,进退,起止,生死,彼此,攻守,胜败,利害,得失,源、流、汇,回、来、去,开始与结束,产生与消除,扩展与压缩,建设与破坏,推动与抑制,积累与消耗,结合与分离,依存与排斥,支持与反对,主宰与从属……等等都是对偶性概念。整体、局部、个体,年、月、日,度、分、秒,体、面、线、点,国家、省、县……等等是包含性概念。

能否给上述三类概念一个精确的定义?一般来说,对语言现象的表述必须容许一定的模糊性或不确定性,追求或模仿数学或哲学式的精确定义往往是徒劳甚至是有害的。当然,也不能走向另一极端,仅仅依靠举例的方式代替定义。因此,这里将给出对比、对偶、包含性概念的必要说明,但不等同于定义。对比型概念是指共寓于同一高层概念下的一组概念,彼此间存在量的差异,而对偶型概念是指这一组概念彼此间存在质的区别,但量变会引起质变,因此,这两类概念不可能有截然分开的界限,“强弱、快慢”形式上很像对偶型概念,但实质上是对比型概念,因为,强与弱通常只是量的差异,然而在某种特定条件下也会转化为质的差异,例如强国和弱国不就有质的区别么?从物理和符号上说,对比是极性相同的,无正负之分,对偶是极性相反的,有正负之分。这就是对比与对偶的大致区别。

包含性是概念层次性的一种特殊形式,层次性就意味着包含,但某些概念的层次性可进行确定的离散式的分解,对这样的概念,值得从一般概念中分离出来予以特殊关注,并命名为包含性概念。而一般的层次性包含则称之为包容,以与这里特指的包含相区别。

对具有上述特性的语言概念,如果采用一种简明的样式予以表达显然有助于计算机的理解。我们将分别采用下列三种样式:

对比性概念： cnk 或 dnk k = (1,n)

对偶性概念： k 或 emk k = 0，1，2，3 或 k = 4，5，6，7

包含性概念： —，— 0，— 00…

式中的数字“c,d”是对比性的标志，“e”是对偶性标志(但对常用的对偶性概念基元省略这一标志)。对比性概念用了两个标志是为了区分排序的属性，数字 c 表示正序，d 表示反序。n 表示总级数，k 表示排序中的值。正序定义为自然顺序，或值的排列从小到大，反序定义为值的排列从大到小。例如，干部级别的值表示应取反序，值越小，级别越高，1 级最高。而技工级别则应取正序，1 级最低。若总级数不定，则约定取 n = 0，“第一，第二…… ”就属于这个情况。

根据这些符号的提示，从一个对比性概念就能联想另外（n - 1）个概念，从对偶性的一方就可以联想到另一方，从一个包含性概念可联想到它的上下方。这里应该说明，上面的符号约定只是有关特性的近似表达。这个近似主要表现在以下三点：

第一，允许对比性和对偶性之间存在模糊。但对这一点的具体表现暂时不管。

第二，对偶性概念在理论意义上至少有三方，即对偶的双方加上两者的“对立统一”方(以下简称统一方)。上面的符号约定只表达了这一共性，但对于对偶性概念的一系列个性则未予表达。其中最重要的一项个性涉及统一方的状况，这又分为三种情况：(1) 自然语言对于对偶性概念的对偶双方一般都配置相应的词汇，但对于统一方则不一定，有的配置，有的不配置，有的语种配置这一些，另一语种配置另外一些，这与该概念的使用频度和不同语种的习惯有关。这些细节可以给出详尽的描述，但我们认为，当前没有这个必要。(2) 具体的对偶性概念可以没有统一方，例如“左中右”是一个完整的对偶，但实际的“手”只有“左手，右手”，不存在“中手”，实际对偶性概念的这一特殊情况目前也不加以区别。(3) 从中层层次网络符号来说，有“1”必有“2”，有“5”必有“6”，反之亦然。但另一方面，有“0”不一定有“1，2”，有“4”不一定有“5，6”。例如，东西南北上下，前后左右的层次网络符号如下：

东	j21011	北	j21021
西	j21012	南	j21022
上	j21031		
下	j21032		
左	j21411	前	j21421
右	j21412	后	j21422

式中的类别符号“j”表示基本概念，层次符号的前两层“21”表示空间的序，基本概念的高层统一定义为两层(详见第 3 节)。所以，从第三层开始即进入中层，但第三层的“0”和“4”只表示该“序”具有对偶性，其具体特性在第四、第五层表示。第三层的“1，2，3；5，6，7”皆为空集。

对偶性通常表现为二重，但也有三重及多重对偶，例如“源、汇、流”就是一个三重对偶。

对偶性概念的二重表示经常有积极与消极之分,如果存在这种情况,则‘1,5’将是‘积极’的默认,而‘2,6’是‘消极’的默认。所有这些以及上一段里所说的细节知识,只在特殊情况下才加以精确表达,一般都采用上列近似表示式。

第三,包含性概念也有包含级数问题,上面的表示式里没有这项信息,所以它也是一个近似表示式。

层次符号是对语义的逐步逼近,它吸收了数学的近似展开思想,因此,层次符号一定是从高层向低层逐级展开。但层次符号又划分不同的聚类,同一聚类内部的层次符号是‘量’的展开,而不同聚类则是‘质’的展开,这与数学的纯粹按量展开又有所不同。本节讨论的中层层次符号是其聚类之一,其他的聚类在下面三节逐一讨论。

一般说来,对自然语言概念的语义表示不应该过分追求精确,而应以合乎需要的近似为目标。理解本身有深层与浅层之分,人的理解也往往只是近似,没有或不能达到精确程度的情况经常存在。当前的紧迫需要是使计算机对自然语言的理解水平由语法层面全面进入语义层面,否则就不可能承担起自然语言理解处理面临的一系列挑战。但这一历史性转折不可能一蹴而就,从中层层次符号的优化设计这一局部环节来说,上列表示的近似程度应以满足理解处理进入语义层面的基本需要为近期目标。

上述说法意味着承认自然语言概念是概念的精确表达,但对这一点必须加以注释。“精确”和‘模糊’这两个词本身就是一种模糊表示,“模糊”这个词由于模糊数学和各种所谓模糊技术的出现,其含义就更加模糊了。实际上,任何语言概念都有其精确和模糊的一面,即兼有精确和模糊的双重性。对‘模糊’的诠释,有必要区分科学性、层次性、脉络性、违例性模糊四类。汉语用‘百,千,万’表示不同程度的多,这是科学性模糊。自然语言对数和量的运用完全不同于数学,基本采用模糊方式,实际上是非常科学的。高层概念对低层概念的包容,是一种层次性模糊。本文给出的概念一般表示式,其层次符号到任何一级都是一个概念,它们向下包容,因而都是具有包容性模糊的概念。实际上自然语言的多数词汇属于这类概念,有些由于包容性过大而容易造成滥用,汉语的‘搞’字就是一例,其层次网络映射符号为v900,是一个典型的高层概念,在这个意义上它与英语的 get 相当。最近有人写文章抨击“搞”字的滥用,有道理,但也反映了作者对‘搞’的高层性认识不足。脉络性模糊也是自然语言词汇的常见情况。汉语对‘上、下’二字的运用就是典型的例子,它不仅表示空间的上下,也表示状态、层次、等级的上下,过程及转移趋向的上下。以上三种模糊可统称科学性模糊或不违例模糊,而科学性和包容性模糊也可称为无模糊表示,在某种意义上说,它们实际上是精确的。

词的违例性模糊是指该词超出联想脉络的多义性,包括滥用。这种情况西语远比汉语多见。语音识别输出的语音阵列是典型的违例性模糊(如果该阵列中根本不包含正确的音节,则不仅含虚警式违例模糊,而且是漏报性错误了)。自然语言理解过程需要进行的解模糊处理当然不包括第一、第二类模糊,对第三类模糊,应依具体情况决定是否作解模糊处理。必须予以消除的只是违例性模糊,可统称为多义选一处理。

对自然语言词汇的多义选一处理是人类理解自然语言过程中最频繁、最基本的操作。对这一操作过程的的形式模拟不在于并行处理或快速运算,而在于以什么巧妙的方式完成大量语义距离的计算。层次符号的构造方式把最频繁、最基本的语义距离计算变成对层次符号的简单逐层比较。这是我们设计上列概念一般表达式的基本出发点。这个表达式采用了灵活的分层结构,到任一层都代表一个概念,至于这个(些)概念与相应的具体语言概念之间,究竟谁是谁的近似,已无关紧要。重要的是:自然语言概念有违例模糊表示,而层次网络符号无违例模糊表示,对概念的局部联想脉络给出了明确的表示,便于计算机把握概念之间关联性。这里顺便指出一点,汉语非单字词的 72% 为单义词,因此,从语言深层处理的角度来看,汉语理解处理的难度有可能小于西语,这与仅在语法层面处理的情况有很大不同。我们认为,这是汉语的一项不同寻常的优势,我们应十分珍视这一优势提供的机遇。

1.3 基本概念

从本节起的相继三节,我们将讨论抽象概念的三大语义网络,即抽象概念的三大聚类,他们被分别命名为基本概念、基元概念和逻辑概念,并用类别符号 j, ϕ, l 加以标志。每一聚类形成各自的概念矩阵。这三个概念矩阵可视为三个超级语义网络,将用来对自然语言概念体系进行总体描述。也就是说,我们试图通过它们对语义场的说法给出具体的表述。

产生这一想法的来源有两个:一是奎廉的语义网络,菲尔墨的格语法和山克的概念从属理论。二是汉语的"字义基元化,词义组合化"现象。第一个来源提出了"语义基元"的杰出思想并暗含着"总体表述"的宏伟目标,第二个来源则提供了语义基元的宝贵原料。考虑到读者可能对第二来源不很熟悉,这里对它略加说明。

语言文字作为一个整体,都具有音、形、义三极,不过"形"这一极在西语里居于从属地位,所以传统语法理论只提音义两极。但汉语是典型的三极语言。两极意味着对义的表达只有音一种手段,这种语言基本不依赖于文字而独立发展。三极则意味着对义的表达有音形两种手段,文字与语言同步发展并对后者产生重大影响。对音的运用属于人类的本能,对形的运用则涉及更高级的智能,因此,汉语对音形两种表意手段的运用必然体现更多的智能性,这是它的长处。但同时又限制了它对语音本能的充分运用,这又是它的弱点。汉语的这种双重性在词汇构成方面表现得最为明显。语言的发展从词汇起步,词汇的基本功能是命名,在命名方式上,汉语与西语的巨大差异不仅是饶有趣味且极富启发性。古汉语的基本命名以单音节为限,几乎不越雷池一步,显得非常原始和笨拙。西语对一个命名的音节数量则不加约束,显得十分灵活和洒脱。但是,命名的需要随着社会的发展而层出不穷,当新的需要出现时,汉语采取以原有单音节汉字重新组合的方式予以表达,充分显示出其灵活和洒脱。西语则恰恰相反,原有词的音节数量一般已不适于再行组合,不得不采取另造新词的原始方式,从而显示出其灵活中的死板和洒脱中的笨拙。这样,汉字就成了一个 Chinese character 和 word 的混合怪物,两千余年来基本上只减不增。依靠约一千多个充分基元化的汉

字,汉语对新概念的表达应付裕如。这确实是一个有力的启示,表明菲尔墨和山克所追求的概念基元(primitive)的完备集合应该从汉语的这些充分基元化的汉字集合里去寻找。

换句话说,这一千多个充分基元化的汉字及其组合词汇是探求语义网络总体结构的宝贵原料,其中关于概念组合的结构知识尤为宝贵。当然,语义网络本身的信息仍然是隐含的,需要运用归纳和演绎的方法才能发掘出来。这就是本文形成过程所采用的基本方法。在这里作者不能不对创立汉字的祖辈表示敬意。可以设想,如果菲尔墨和山克先生粗通汉语,概念层次网络理论也许在 20 年前就出现了。

下面就来介绍基本概念语义网络的基本构成。这个基本构成就是基本概念的定义,这是一种‘可意会而不可言传’式的定义。所谓基本构成是指该网络的一、二级概念节点,如‘j 表’所示。

j 表:基本概念一、二级概念节点表

j0	序及广义空间
j00	序(一般)
j01	广义空间(广义位置及广义方向)
j02	广义距离
j1	时间
j10	时间
j11	时间的序
j12	时间的间隔
j2	空间
j20	空间
j21	空间的序
j22	空间的距离
j3	数
j30	基本数
j31	数的空间
j32	数的变换
j4	量与范围
j40	全体、局部、个体(包含性)
j41	量(单、双、半 ;少与多)
j42	范围(界、内、外 ;跨、起、止)
j5	质与类
j50	内容与形式
j51	质(对比、对偶)
j52	类(包含)
j6	度
j60	度(一般)
j61	量变的度

j62	质变的度
j7	基本属性(客观的)
j70	对比性
j71	对偶性,对仗性
j72	主要与次要
j73	特殊、一般;宝贵、普通
j74	本质、表象;
j75	相对与绝对
j76	常规与异常
j77	简单与复杂;纯与杂
j78	新与旧
j8	基本属性(含主观评价)
j80	正与邪
j81	真与假;实与虚
j82	善与恶
j83	美与丑
j84	对与错
j85	是与非
j86	积极与消极

j 表对基本概念语义网络的全部一、二级节点给出了定义,并约定基本概念的高层层次符号只有两层,从第三层起即转入中层或底层。这就是说,基于上列定义,其他的基本概念都可以通过上两节里分别定义的五元组及中层符号予以表达,这是基本概念的最大共性,即其个性可以通过对比性、对偶性或包含性予以表达。例如:

j40 -	全部	j40 - 0	局部	j40 - 00	个体
jv40 -	包括	jv40 -	属于		
j41 -	多	j41 - 0	单元		
		j41em1	单	j41em2	双
		j41em3	半		
jvzu41c21	少	jvzu41c22	多		
j420	界,边	j421	内	j422	外
		j4211	核心	j4212	外围
jv425	起	jv426	止	jv427	经
j4241	端	j4242	间		

在上列表示式里,如果五元组符号 g 单独出现,则予以省略,即 jik = jgik。

从形式上看,基本概念语义网络比较简单,双层结构特征比较突出,第一层是一个 9 × 5 概念矩阵。第二层的多数(除了两个基本属性的子网络之外)是 3 × 5 矩阵。下面就来对这

个双层矩阵作简要的说明。

第一层的九个一级节点可分为四个子类,时间、空间是第一类,量、质、度是第二类,两个基本属性是第三类,数是第四类。这四类当中,以“数”这一类最为特殊,也最为复杂。哲学史上的基本论争可以说正是由于“数”的参预而经常陷于僵局。上面关于基本概念最大共性的论点,对于“数”也不适用。可见其特殊性非同寻常。不过,语言对“数”的理解不必达到数学专业的水平,这就是设计“数”的二级节点的出发点。“数”按其特殊性安排在最后较为合理,但汉语的 HNC 知识库已按 j 表装建,不变动为妥。

两项基本属性是一切事物的最高层共性,是一般属性(由五元组的 u 或物性符号 x 表达)的属性。基本属性是哲学和社会科学的研究对象,而不是自然科学的研究对象,这个分野比较明显。把基本属性按其是否含有主观因素划分为两大类,其分野也比较清楚,j8 大体上相应于伦理学,是中国古代儒家学说的核心。目前有人认为,古老的哲学只剩下一个不可穷尽的探索领域,这就是伦理学,我们基本上同意这一观点,所以将 j8 独立出来设置一个一级概念节点。

量、质、度是基本属性的另一类,是自然科学和社会科学共同的研究对象。它们的二级节点,和时间、空间、序一样都是三个。这一巧合现象的出现,主要是由于引入了广义空间的概念,至于将“量与范围”质与类”并为一个一级概念节点,将“度”分为量变和质变两个子类,乃常理使然,无须解释。所以,这里仅对广义空间的概念作简要的说明。

序是宇宙万物的基本特征,故列为基本概念之首。一切过程是状态序列的时间表现,一切转移是状态序列的空间表现。有序才有规律,规律就是序的效应,其层次网络符号就是 jr008。序的基本特性是“位置、方向和距离”,这三个概念来于空间,但超出了空间。时间把“位置”换称“时刻”,把“距离”换称“间隔”,如此而已。人们说时间不能倒流,意味着时间的单向性,然而有向。“岗位和地位”是社会空间的位置,“趋向和志向”是社会空间的方向。至于距离,就是五元组 z 的差值,不过语言习惯上更多使用“差距,差异,差别”等等名称。汉语里的“位、向、距、上、下”诸汉字实际上就是在广义空间的意义上加以运用,是汉语字义基元化的生动表现之一。西语对上列概念的运用也有类似表现,不过不像汉语那么突出。由此可见,广义空间概念的引入,对于表达自然语言的概念关联性,对于概念的从高层向低层逐步逼近的表达方式,惟有利而无弊。

语义网络节点的设计应遵循三条原则:一是制定共性表达的优化分层准则,二是考虑网内关联性表达的便利,三是考虑网间关联性表达的便利。但第二点与第三点往往相互矛盾,难以兼顾。高层节点的设计以照顾第二点为主,底层节点反之,以相互补充。但基本概念语义网络的设计采取了置底层于不顾的做法。这样做是否会留下后遗症,目前还难以作出预测。

上述语义网络设计的三条原则适用于所有的语义网络。但基元概念和语言逻辑概念网络的设计还另有自身的基本原则。对基元概念,二级概念节点的设置还要考虑与基本句类[2]二级分类的匹配。对语言逻辑概念,其一级概念节点的设置主要是考虑语义块[2]感知

处理[11]的需要。

1.4 基元概念

在抽象概念的三大语义网络中,基元概念语义网络居于中心地位。其高层节点的设计最为复杂,涉及语言深层的基本结构。所以,有些问题要放到【14】到【21】里讨论。这里,从基元概念一级节点的设计谈起。

基元概念的一级节点分为两大类,一类是主体基元概念,另一类是复合基元概念。第一类基元概念由作用效应链构成,故亦称作用效应链,它是基元概念语义网络的核心,某些复合基元概念语义网络就直接定义为对它的挂靠。挂靠是一个新术语,下文有具体说明。复合基元概念主要涉及人类活动,这不难理解,因为自然语言的主要表述对象是人类活动而不是自然现象。对后者的表述,自然语言符号是远远不够的,还需要数学及各专业的特殊符号体系,但自然语言的配合仍不可或缺。当前专家系统的研制,正是由于这一配合的暂缺而必须借助软件工程师的介入,如果计算机能够如同人一样理解自然语言,就有可能逐步减少这种介入,最终实现由专家自行研制自身专家系统的诱人前景。

下面就来分别讨论第一类和第二类基元概念。

1.4.1 主体基元概念

第一类基元概念是主体基元概念,也就是作用效应链,由下列六个一级节点组成:

$\phi 0$ 作用
$\phi 1$ 过程
$\phi 2$ 转移
$\phi 3$ 效应
$\phi 4$ 关系
$\phi 5$ 状态

类别符号 ϕ 在实际表达时可略而不用,这六个节点是自然语言对万事万物进行总体表述的六个基本角度,也是一切事物发生、发展和消亡的六个基本环节。作者在四年前曾写道:“作用效应链反映一切事物的最大共性,作用存在于一切事物的内部和相互之间,作用必然产生某种效应,在达到最终效应之前,必然伴随着某种过程或转移,在达到最终效应之后,必然出现新的关系或状态。过程、转移、关系和状态也是效应的一种表现形式。新的效应又会诱发新的作用,如此循环往复,以至无穷,这就是宇宙间一切事物存在和发展的基本法则,也是语言表达和概念推理的基本法则”。

这六个环节的源头是作用,结果是效应。自然语言的主要内容就是对这六个环节进行局部和总体的具体表述。这里顺便说明一下,山克的“概念从属理论”主要考虑了“转移”这

一个环节,我们对“转移”二级节点的设计就部分吸收了“概念从属理论”的主要结果。

各环节的二级节点设计如 φ 表(表中已省去 φ):

φ 表:基元概念一、二级概念节点表

00	作用
008	物理作用
009	化学作用
01	作用的承受
02	作用的反应
03	作用的免除
04	约束
10	一般过程
100	过程的基本特征
1008	过程的持续
1009	过程的进展
100a	过程的重复
109	运动过程
10a	演变过程
10b	生命过程
11	过程的序
12	过程的因果及源、汇、流
13	过程的趋向及转化
14	过程的新陈代谢
20	一般转移
200	转移的入出(基本特征之一)
204	转移的起止(基本特征之二)
209	定向转移
20a	传输
20b	自身转移
21	接收
219	针对性接收
21a	物的接收
21b	信息的接收
22	物的转移
23	信息转移
24	交换、替代及变换
30	一般效应

300 效应的基本特征
309 变化
30a 实现及完成
31 产生与消除
32 利与害
33 显露与隐蔽
34 扩展与缩小
35 立与破
36 推动与抑制
37 连断及通阻
38 选、存、弃
39 合分及聚散
3a 获得与付出
3b 积累与消耗

40 一般关系
400 关系的基本构成(方面)
4001 我方
4002 你方
4003 他方
4004 双方
4005 此方
4006 彼方
408 关系的相互性
409 关系的紧密性(距离)
40a 关系的传递性
41 结合与分离
42 依存与排斥
43 支持与反对
44 主宰与从属
45 使用与舍弃
46 拥有与失去
47 适应与干扰

50 一般状态
500 状态的基本特征
508 自然状态
509 生命状态
50a 人的状态
51 形态

52	动态
53	势态
54	结构
55	层次
56	等级

下面对主体基元概念二级节点的设计原则分别进行说明。

1.4.1.0 作用概念基元

作用概念基元分为五个二级节点,分别表述了作用的五个基本侧面。但这五个侧面的相互关联性有本质的不同。前三个节点呈链式关联:作用 - 作用的承受 - 作用的反应是缺一不可的三级链条。作用的免除是反应的一种特殊形式,约束是作用的一种特殊形式。因此,03 与 02 交式关联,04 与 00 交式关联。为什么将 03 和 04 独立出来,不作为 02 和 00 的一个子节点来设置呢?这涉及多方面的考虑,但主要的一点是为了句类分析的便利。

上面我们使用了“交式关联”、“链式关联”、“句类分析”等术语。本论文集里将经常使用这些术语,这里稍加解释。

“交式”来源于数学的“交集”概念,不同概念的内涵存在“交集”是普遍现象。两概念节点的“交集”不是空集,就表明两概念节点之间呈现交式关联。作用效应链的六个基本观察角度不可能截然分开,相互之间存在“交集”,具体表现为 ϕ 表和 j 表各概念节点之间存在交式关联。这类关联知识在论文【14】到【20】里有详细说明,并构成概念知识库的重要内容之一[6]。

不同词性的词汇搭配,也可视为交式关联性的表现之一。五元组的表示方式将这一类关联性形式化为层次符号的相同或相近,如第 1 节的示例所示。这样显然有利于计算机对这类关联性的把握。如“精神振奋,斗志昂扬,意气风发”、“无私奉献,慷慨就义”、“锦绣山河,远大前程”、“承担责任、召开会议”之类的搭配,在层次网络符号里都是“同一”概念节点的不同五元组表现,他们的优先搭配乃“理所当然”。我们把这一“理所当然”现象叫作同行优先,因为它们属于概念矩阵的同一行。与词性搭配的概念相比较,可以这么说,词性搭配准则只给出了搭配的必要条件,而没有给出充分条件,而同行优先准则则同时给出了充分必要条件。因此,同行优先准则在解模糊处理时能产生更好的效果。

同行优先有狭义与广义之分,狭义同行的“同”是指同一概念节点,广义同行的“同”则指具有交式关联的不同概念节点。上面示例中的“远大”与“前程”是广义同行,其他都是狭义同行。

当然,同行优先准则里不仅有交式关联,也有链式关联,上面例子里的“承担责任、召开会议”就属于链式关联,其他是交式关联。

所谓链式关联是语句要素[2]之间的关联,在文【6】中,将这种关联细分为 11 类,并按类建立概念关联性知识库。

这里所说的交式或链式关联性是以概念节点为对象的,所谓联想脉络就是指概念节点

之间的这种交式及链式关联性。通过概念节点而不是通过具体词汇来表示这一脉络的构成是概念知识库的基本思路[6]。

1.4.1.1 过程概念基元

除作用外,其他主体基元概念的二级节点之间基本不存在链式关联,而以交式关联为主,0 号二级节点的代表性或包容性更为突出。这样,以后在谈到二级节点的个数时,除作用基元外,都不包括 0 号节点。

过程概念基元有四个二级节点,它表述了过程的四个基本侧面。这四个侧面两两强交式关联。即过程的始与因、终与果强关联,过程的趋向、转化与新陈代谢强关联。这四个侧面是过程的完备表述么?对过程的表述主要是以下四个方面,第一是过程的起源,第二是过程的结局,第三是过程本身的特征,第四是过程的类别。上列四个二级节点概括了前两方面的表述需要,后两方面则放到 0 号节点里予以表述。为什么不直接按这四个方面设置二级节点呢?这主要是考虑到语言表达的侧重倾向,也就是过程句[2][16]二级分类的需要。上一节已指出,这是主体基元概念二级概念节点设计的基本原则之一。转移基元概念二级概念节点的设置与过程恰恰相反,把转移的类别特征作为二级节点设置的依据,原因也在于此。

1.4.1.2 转移概念基元

转移实际上是过程的一个子类。上述关于过程的四方面内容,转移同样存在,过程的基本特征(由 100 节点表述)完全适用于转移。但转移语言表述的侧重点却与过程不同,它侧重于转移的类别,这就是转移网络二级概念节点设计的基本依据。物和信息的转移分别是转移的两大类别,这不难理解,“概念从属理论”就用它们作两个概念基元的命名。另外两个二级节点则比较特殊。“接收”只是转移全过程的后一半,把它与物和信息的转移并列为转移的类别之一,形式上似乎很不协调,但如果注意到转移的本质在于“接收”,你就不觉得奇怪了,因为它意味着转移的结局和目的。从感知到审查,从古老的天文观测到各种现代化被动式探测设备,都只涉及接收,然而它们显然是一类重要的信息转移。

将交换、替代和变换并为一个子类是为了体现这样一项共性——“涉及两项转移内容并且两相交互”。这项共性在过程里比较平淡,但在转移里却相当突出,贸易和谈判就是人们熟知的例子。交换的交互性是明显的,变换的多数具有可逆性,替代虽然一般不具有交互性,但必然涉及两项转移对象。

1.4.1.3 效应概念基元

作用与效应的关系可以非常贴切地与磁性的两极相比拟。两者不可分离,但却是作用效应链的两极,作用是源极,效应是末极。这是概念层次网络理论的基本观点。

效应二级节点的设置原则与转移相同,即以效应的类型为依据。但效应的类型更为复杂,ϕ 表中所列清单的完备性,很难仅从理论上加以论证。从汉语字词的 HNC 符号映射过程来看,迄今为止,还不曾发现有引入新节点的必要。这当然不能作为完备性的证据,但应该引起对完备性提法的思考。概念基元对于语言理解的作用或意义类似于化学里的元素或数学里的正交函数,但没有必要达到元素周期表或完备正交函数系的水平,因为概念基元的

首要价值与其说是给出复合概念的精确表示,不如说是给出概念关联性知识和联想脉络的线索。

1.4.1.4 关系概念基元

关系基元概念二级节点的设计,采取与转移相同的思路。理由当然也是出于关系句二级句类划分的便利,详细论述见【18】。

1.4.1.5 状态概念基元

状态基元概念二级节点的设计,采取与过程大体相同的思路。但在中层层次符号的设计方面采取了两项特殊措施。第一是将 500 与基本概念挂靠,第二是引入了"动态"和"势态"两个二级节点,并将它与基元概念(包括状态概念自身)挂靠。这里挂靠的意思等价于并,500ik = (500, jik),52ik = (52, ϕik),53ik = (53, ϕik)。这里的 ik = ik…,层次不限。五元组符号统一放在 5 的前面。按照本文前面引入的术语来说,这里的 500, 52, 53 是本体层,jik, ϕik 是挂靠层。

这里的 52"动态"实际上没有相应的语言词汇,它试图表达古希腊哲学家赫拉克利特的"你不能两度跨进同一河流,我们存在,又不存在"这一名言里的思想。这个思想对于揭示语言概念的关联性十分有效。从下列概念的对比表示中可以清楚地看到这一点。

结束	v112	暂停	v52112
显现	v331	闪现	v52331
力的作用	v008	碰撞	v52008
竞争	vb30	赶超	v52b30
过程进展	v1009	爆炸	v521009

这就是说,暂停是结束的动态,闪现是显现的动态,碰撞是作用力的动态,赶超是竞争的动态,爆炸是过程进展性的动态。从这些例子可以看到,在"52"里含有暂时、瞬间和突然等等的意思。这些信息细节都不难通过扩展 52 的本体层予以表达,但目前并没有这样做,因为主要目的在于用它表达概念之间的关联性。

引入"势态"概念的目的与"动态"相同,即主要用于概念关联性的表达。如酝酿是开始 v111 的势态 v53111,后效是结束效应 r112 的势态 r53112,风险是损害效应 r322 的势态 r53322,预测是关于势态的判断 (v810, l10, g53)。汉语势态一词的内涵非常丰富,《过秦论》和《封建论》里的"势"字就是 53 含义的精确映射,英语里无相应的词。它有"潜在的存在""带有某种程度偶然的必然"等多方面的含义。现代物理学的混沌学可以说就是对势态的研究。

状态表达的 6 个基本侧面如 ϕ 表所示。这 6 个侧面当然相应于状态句的 6 个二级子类。其中的"动态"和"势态"侧面显然比较特殊,而排列次序却插在"一般侧面"中间,表面上显得很不协调。这一不协调性与"数"在基本概念的排序相似。但这一不协调性有其内在的合理性,而且不会对软件的设计和运作造成不利影响,不必予以调整。

1.4.2 复合基元概念

复合基元概念总共设置了 8 个一级概念节点,如下表所示:

φ6	生理及本能活动
φ7	心理活动及精神状态
φ8	思维活动
φ9	理智活动
φa	专业活动
φb	追求活动
φc	社会性活动
φd	规约性活动(行为)

显然,复合基元概念主要针对人类活动的表述,这一点不言而喻。

表述一般从分类着手,但对人类活动的分类,显然不宜采用简单的线性准则。如同对一般事物的语言表达需要从作用效应链六个不同的基本角度加以透视一样,对人类活动也需要从不同角度进行透视。

人类与动物的区别首先是理智表现,φ9 是这一表现的标志。9 行以后的各项活动都具有理智的特征,所以 9 行与它们广义同行优先。对理智表现或智能活动当然需要从不同角度进行透视,但不同角度的划分显然是一个十分复杂的问题,我们在这里建议采用‘专业性、追求性和规约性’的标准,并把它们安置在概念矩阵的 abd 三行。这三行的层次符号完全另行设计,不向主体基元概念挂靠,高层的层数统一定义为三层(主体基元概念为两层)。

在语言表达中,理智活动的专业性、追求性和规约性带有明显的语境特征。对这三类活动的表述往往不是一两句话的规模,而需要大块文章。这就是说,这三行的语言概念绝不会在文本里孤零零地出现。因此,对这三类概念作简单的统计,即可取得宝贵的语境信息。

那么,为什么不将专业性、追求性和规约性连续编号,而在中间插入表示社会性的 c 行?这是由于考虑到规约的基础是人类活动的社会性,离开了这个基础,规约的概念将更加呈现出‘公说公有理,婆说婆有理’的混乱状态。这种安排,同思维之放在 8 行,属于同一思路。d 行是基元概念矩阵的最后一行,把规约性活动放在这一行,暗含着一种对未来的憧憬。如果因此而对软件的设计与运作带来不便,作为‘始作俑者’将深感歉意。

上述三类智能活动不直接向主体基元概念挂靠,意味着它们是主体基元概念的综合。但专业和追求活动必将形成某种作用,并产生某种效应,所以,由 ab 两行 v 概念构成 E 要素的句子通常是作用效应句或含作用效应的混合句。

对人类日常活动的表述,大部分应能直接向主体基元概念挂靠,否则基元概念就失去了存在的意义,或基元概念的设计存在根本缺陷,因为,语言的表达主要涉及人类的日常活动。

人类的日常活动显然有不同的层次,从最低的生理本能活动到最高的社会性活动。这个层次的划分不宜过于学术化,三层比较适当,即在最低和最高之间加一个中间层,并将中

间层与一般智能活动合而为一。语言的词汇就大体遵循这一分类准则。这样,本能活动的6行和上述的9行与c行又构成一个“集群”。这个“集群”与主体基元概念直接挂靠。

然而,本能活动是一切生命体的共同特征,并非人类所特有。自然语言对人与一般生物本能活动的区别给予了足够的注意。所以,6行在挂靠之前加了一层以表示这一区别。这就是说,本能活动的本体层为两层,而一般智能活动和社会性活动的本体层为一层。

复合基元概念的概念节点表和综合类概念节点表将分别在文【6】的6.2.3节和文【2】的2.3节中给出。

1.5 逻辑概念

本节讨论两类逻辑概念,一类命名为语言逻辑概念,大体上相应于汉语的虚词。用类别符号l予以标记。另一类命名为基本逻辑概念,大体上相应于西语的系词和情态动词,用类别符号jl予以标记。

1.5.1 语言逻辑概念

语言逻辑概念的层次符号按本体层、挂靠层的串接方式设计。本体层表示语言逻辑概念自身的意义,是本节讨论的主题。挂靠层取自基元或基本概念的高层(少数情况进入中层)层次符号。因此,语言逻辑概念的一般表达式有下列两种基本形式:likmn或ljikmn。ik为本体层,mn为挂靠层。前者表示向基元概念挂靠,后者表示向基本概念挂靠。本体层统一定义为两层。挂靠层通常为两层,也可以不挂,或一直挂到中层,即mn可以延伸。语言逻辑概念同样具有五元组的特性,因此,在l或lj后面可加五元组符号。

前面我们看到,基本概念的高层是两层,主体基元概念的高层也是两层,现在语言逻辑概念的本体层又是两层,这种两层性显然有利于理解程序的设计。语言逻辑概念的一级概念节点分为三类:

(1) 语义块区分标志符　　i = 0　3

(2) 语义块组合标志符　　i = 4, 5

(3) 语义块说明符　　i = 6　b

下面分别对这三类语言逻辑概念作具体说明。

1.5.1.1 语义块区分标志符

一个句子通常由若干个主、辅语义块组成。这些语义块的排列顺序,不同语种各有自己的偏好,但任何语种都容许语义块顺序的变动。在变动时,语义块作整体搬迁。一个语义块通常又分为说明与核心两部分,作为一个整体,它们通常是排列在一起的,不容许其他语义块插入。但主语义块之一的特征语义块却经常例外,其不同成分分置两处。其他语义块也有分置的情况。语言表达时对语义块运用的这些复杂情况就产生了一种需要,就是对不同语义块和语义块的不同成分给出相应的标志。

由于语义块有主辅之分,所以这个标志自然也分为两大类,就是主语义块标志和辅语义块标志,简称主、辅标志。在形式上这些标志又分两类,一类是单一性标志,放在所标志语义块的前面,另一类是两两搭配的标志,其中包括类似于括号功能的搭配标志,如汉语里的"在……下"在……方面"等等。考虑到对语义块区分标志的上列要求,其本体层第一层的层次符号"i"设计如下:

i = 0　　主语义块单一标志符
i = 1　　辅语义块单一标志符
i = 2　　主语义块搭配标志符
i = 3　　辅语义块搭配标志符

语义块区分标志符本体第二层的层次符号"k"用于表示语义块的具体类型。主语义块分为四种,分别命名为特征、作用者、对象、内容,简称 E、A、B、C 语义块。辅语义块分为七种,分别命名为方式、工具、途径、比照、条件、动因、目的,主语义块"k"依次取值 0　3,对辅语义块"k"依次取值 1　7。

语义块区分标志符主要相应于传统语言学里的介词,但包括按传统概念难以注明词性的义项以及在不同语种里词性标记不同但语言逻辑意义相同的词。前者如汉语"给,予,使"的逻辑意义,后者如汉语的"比"与英语的"than"。它们都是搭配型语义块区分标志符。"给,予,使"的本体层映射符号相同,都是 lq22,但挂靠层不同。"给,予"挂靠"0200 + 00",表示直接标志的是作用对象,搭配的是作用。"使"挂靠"0200 + 30",表示直接标志的是作用对象,而搭配的是效应。从"给(予)敌人以沉重打击"、"使大家满意而归"这类常见的句例可以看到上述区别。而这一区别的得以表达就是依靠了挂靠层的引入。

语义块区分标志一般表示式 likmn 或 ljikmn 概括了语言表达的一类需要,它的每一级表示式,li, lik, likm 都有确定的意义。重要的是:这些表示式只有包含性模糊,而没有脉络性和违例性模糊,这是它与介词的根本区别。在本节的开头我们说到,语言逻辑概念大体上相当于汉语的虚词,这里又说到,语义块区分标志大体上相当于传统语言学的介词。那么,为什么不直接采用传统语言学的术语和表达方式?因为,介词的概念不区分语义块的"主"与"辅",而连词的概念不区分语义块的"分"与"合"。而从语义块感知和句类分析[2]来说,这一区分是至关紧要的。

主、辅语义块类型的划分只是语义块意义揭示的第一步,第二步是确定它与句类的函数关系。主语义块的这种函数关系更为明显,把握这个函数关系是跨入语言深层的关键一步。语义逻辑概念层次符号挂靠层的引入,对第一类语言逻辑概念来说,就是为了标明这种函数关系。

语义块区分标志符还有语种的共性和个性,搭配的真假,主辅转换等问题。这些就不在这里讨论了。

1.5.1.2　语义块组合标志符

分合是天然的对偶,既有语义块区分标志,就必然有语义块组合标志。组合标志的语种

个性远比区分标志突出。它相应于多种词性的虚词,包括关系代词和部分连词、介词以及按传统概念难以注明词性的词。介词如英语的‘ of ”,连词如汉语‘ 和、与、及 ”,英语的‘ and ”。

理论上,语义块的组合标志应该包括下一节所介绍的全部组合结构符号。l4 是组合标志,l5 是组合说明。l4k,l5k 的 k 定义,实际上就是建立数字 k 与相应组合符号的对应关系。但语义块组合中最常用的只是偏正、并、或三种。因此,目前仅定义了 l4 的部分 k 值:

k = 1　　偏正
k = 2　　反偏正
k = 3　　并
k = 4　　或

1.5.1.3　语义块说明符

语义块说明符定义如下:

l6　　时态说明
l7　　暂空
l8　　辅要素说明
l9　　指代逻辑
la　　E 要素逻辑说明
lb　　语句间逻辑说明符

1.5.2　基本逻辑概念

如果说语言逻辑概念在不同程度上依赖于基元概念和基本概念而存在,那么,基本逻辑概念是独立的。因此,前者采用挂靠结构,而后者不采用。

基本逻辑概念涉及基本判断,基本判断是思维活动的基本操作。而基本判断的起点或基础是比较,基本内容是‘ 是否有无 ”。这些,就是基本逻辑概念语义网络设计的基本依据。根据这一观点,基本逻辑语义网络的一级节点只有两个,这就是比较和基本判断:

jl0　　比较
jl1　　基本判断

比较的二级节点按比较的类型分为三类:

jl00　　两相比较
jl01　　集合内相互比较
jl02　　与一个标准比较

基本判断的二级节点设计比较复杂,有多种方案可供选择。这里仅介绍目前选定的方案,但不加解释。

jl11	判断内容
jl111	是,肯定
jl112	否,否定,不
jl115	有,存在
jl116	无
jl12	判断的势态(纯客观):可
jl12c31	可能
jl12c32	能够
jl12c33	必然
jl13	判断的势态(含主观)
jl13c21	应该,应
jl13c22	必须,必

对上列表示式应作两点说明:第一,省略了五元组符号。第二,基本逻辑概念的高层也是两层,从第三层以后,即进入统一的中层层次符号表示。

基本逻辑概念 jl0, jl11 将构成基本判断句特征语义块的要素,而 jl12, jl13 是任何句类特征语义块说明成分的活跃块素之一,即传统语言学的情态部分,它们也可视为语义块的区分标志之一。

语言逻辑概念设计的指导思想是为语义块感知处理服务,l 网络层次符号前两级的设计,即本体层层次符号的定义,主要以此为立足点。

1.6 概念的组合结构

以上三节介绍了三类概念基元,包括反映人类活动的复合概念基元,这些概念基元只是概念海洋里的"元素",概念的海洋乃由这些"元素"以各种方式组合而成,这是常识。传统语言学曾从主谓宾补的观点对这些组合结构作过一些描述,提出了"联合,偏正,后补,述宾,主谓"五种组合结构。这个描述,从语言表层来看,它不够完备。对于"枪毙,点焊,宵禁"之类的词汇,对于"水到渠成""心宽体胖"之类的短语,它不能给出适当的表述。从语言深层来看,组合结构必然是多侧面和多层次的,上述表述方式不能适应这一需要。

如果对上述五类结构的说法加上"逻辑,因果"两项补充,大体上就完备了。因果是逻辑概念之一,逻辑结构含因果结构。为什么把它独立出来?因为这里所说的"因果"是指作用与效应。从作用效应链的观点来看,必须把它独立出来,其理由之一是:由这类组合结构形成的词汇含有对句类分析特别宝贵的信息,在文【2】中对此有详细论述。

基于上述,本文将概念组合结构分为四个基本类别,每一类别又分为两个子类,如下表所示:

类别名称	子类名称	符号	传统命名
1 作用效应类	作用	#	后补
	效应	□	后补
2 对象内容类	对象	&	述宾
	内容	|	述宾
3 逻辑类	并与选	,;	联合
	非与反	!	
	逻辑	(,1m,)	无
4 语法类	偏正	/	偏正
	主谓	‖	主谓

在形式上,上表极易造成一个误解,似乎新定义是将传统定义的后补和述宾两种结构换了一个名称,并一分为二。所以,这里有必要对新旧定义、特别是它们的对照关系作进一步的说明。

让我们从“说服,说破,说谎”这三个词谈起。按传统定义,它们都属于后补结构。但按照新的分类准则,它们分属于作用、效应和内容。为什么要作出这种划分?因为在句类分析时它们将提供截然不同的关键信息,“说服”将构成作用句,“说破”将构成效应句,而“说谎”将构成转移句。在相关语义块的构成方面,对“说服”必须分别配置对象和内容,而且可以扩展为语句,而“说破”不具有这一特性。对“说谎”可配置对象及内容,但也可不配置而不影响语句的完整性。对三者这一重要差别的表述,“后补”二字显然是无能为力的。而将它们分别纳入“作用、效应、内容”型组合结构则可以显示这一差异。

新定义的一类组合结构并不相应于旧定义的某一类,表中所列对应关系只是示意性说明。例如作用效应类,它主要相应于后补,但也可含有旧定义的联合和偏正,像“诬告,诬赖,诬陷”之类,旧定义可列为联合或后补,但新定义一定是作用。像“水压,水灾,水电”之类,旧定义应为偏正,而新定义则为效应。

对象与内容也不只包含旧定义的述宾,但述宾一定是对象或内容。宾语这个概念过于宽泛,所以传统语言学也曾对宾语作过更细的分类,这主要是在“格语法”理论的推动下进行的。但宾语的分类涉及语言深层的句类分析,当年的分类工作不可避免地显得不得要领。现在我们知道,宾语不仅是句类的函数,而且首先要分为对象和内容两大类。这一点在【2】里有详细阐述。不过,宾语首先应该一分为二的现象在语言表层也不难观察到,例如“杀人”与“放火”、“练兵”与“练武”、“立法”与“违法”,都是述宾结构,但显然有对象与内容之分。人和兵是杀和练的作用对象,而火和武是放和练的内容。同一个法,立法对法产生影响,违法并不对法产生影响,所以前一个法是立的对象,而后一个法是违的内容。

逻辑类的第一个子类是旧联合类中的一部分,联合这个概念同宾语一样,过于宽泛,所以其中的另一部分纳入作用效应类。逻辑类的第二个子类显然是原分类法的疏忽,上面举的“枪毙”等例子显然必须归于这一类,汉语里的这类词汇比较丰富,显然不能勉强纳入偏正

类或其他。这个子类的进一步划分,采用辅语义块[2]的分类准则。

语法类沿用原来的命名,但实际内涵已大为缩小。缩小后的偏正还需要作二级分类,这项分类工作也相当复杂,这里就不来讨论了。

在上列四类结构中,除了逻辑第一子类外,都存在正反两种形式。"红旗漫卷西风"是句子里典型的主、谓、宾反结构,这种反结构在词汇级同样存在。上一节曾标明,汉语的"的"是正偏正结构符号,而英语的"of"是反偏正符号。像"心窄,心硬"这样的词,用反偏正来表述就比较恰当。如果不引入反结构的概念,它们只能纳入主谓结构,但与"心慌,心领,心碎"相比,你就不难发现,后者才是标准的主谓结构,如果把前者同它硬绑在一起就未免有些勉强了。但勉强并不是完全不合理,这就是说,一个具体词汇组合类型的选定不是绝对的。

从理解处理的角度来看,组合概念的具体结构并不重要,重要的是它的功能表现。在上列组合结构符号中,突出代表功能表现的是作用效应型组合 # 和□。功能表现将用"语义结构方程"来表示,在使用结构方程时,实际的偏正、并选、逻辑组合结构,根据需要可虚用作用型或效应型来表示,这个问题在〈问答 34〉中有详尽的论述。

1.7 具体概念的近似表达

一般来说,具体概念的精确表达要比抽象概念困难得多,但另一方面,人在理解过程中对具体概念的认识深度可以比抽象概念浅得多,天生的盲人仍能同常人一样掌握自然语言,道理就在这里。这就是说,对具体概念的表述,应采取大胆近似的方案,这是对具体概念进行层次符号设计的基本出发点。

具体概念的类别,从语言表达的角度来看,先分为物、人、物性三类比较合理,分别用类别符号 w、p、x 予以标记。物有自然物与人工物之分,人工物又有现代与传统、物质与精神产品之分,当然还可以有各种各样的分类标准。人和物性也同样存在子类划分问题。在处理具体概念的分类问题时,不宜照搬自然科学的分类方法,我们的着眼点主要是引起概念的联想,其次才是分类的科学性。因此,对三类具体概念的子类都不引入另外的类别符号,而用类别符号组合的方式予以表达。例如,人工物、现代产品、精神或信息产品将分别用 pw、w9、gw 予以标记,类别符号的组合意义当然带有模糊性,但只要不产生违例模糊,是可以作为相应物类的定义符看待的。

表达 p、w、x 具体特性的层次符号,将仿照语言逻辑概念的做法,向基元概念和基本概念挂靠。例如,基元概念里的概念节点 22b 表示自身转移,那么,向它挂靠的 pw22b 就表示交通工具;219 表示针对性接收,pw219 就表示现代探测设备;412 表示结合,p412 就表示夫妻,pe411 - 就表示家庭;392 表示废弃,pw392 就表示垃圾;j711 和 j712 表示正负,pj711,pj712 就分别表示男人和女人;j20 - ,j20 - 0,j20 - 00 分别表示"体、面、线",xjz20 - ,xjz20 - 0,xjz20 - 00 就分别表示体积、面积和长度。

显然,上述表示方式都是很粗糙的近似,但重要的是通过这一近似,计算机就能对有关

概念之间的关联性有所‘领会’。挂靠式表示方式的目的,就是在具体概念与抽象概念之间建立一种交链式关联的符号表示,并尽可能制造出‘同行’的数据格式,以利于计算机计算语义距离。

挂靠可以多重。例如:

货币	gwc248	证券	gwc248 + va24
建筑物	pw6554	住宅	pw6554 + v6550a9
		办公大楼	pw6554 + va0
		走廊	pw6554 - + v22a
		门窗	pw6554 - + v220

挂靠表示方式,当然只适用于一部分具体概念。一些常用的具体概念,主要是物,仍然需要进行独立的层次符号设计。我们把这一部分具体概念定义为基本物,并用类别符号 jw 予以表示。

jw 的层次符号定义如下表所示:

jw 表

jw0	热
jw1	光
jw2	声
jw3	电磁
jw4	微观基本物质
jw5	宏观基本物
jw51	气态物
jw518	大气
jw52	液态物
jw528	水
jw53	固态物
jw538	土
jw6	生命体
jw61	植物
jw62	动物
jw63	人体
jw61 -	植物部件及组织
jw62 -	动物部件及组织
jw63 -	人体部件及组织

将表中的 w 换成 x 就表示一些基本物性,例如 jx0 表示温度,jx1 表示色彩等。如同基本概念和基元概念是抽象概念的概念基元一样,jw 表定义的具体概念是物表示的概念

基元。

具体概念的表示,除了层次符号的挂靠,物概念基元的定义之外,还利用了类别符号的连用知识,上面示例中的 pw 就是一例。这种表示方式在【6】中有进一步的说明。

结 束 语

本文初稿写于 1991 年,当时的题目是“概念层次网络理论概述”。到今天又过去四年了,但作者仍感到,本文仍未到公开发表的水平,因为层次符号的底层设计仍在探索之中,而高层设计的合理性检验是离不开底层设计的。

但是,底层设计是一个复杂的系统工程,我们寄希望于与语言学家及同行们的合作。

1995 年秋

自然语言的深层结构及句类分析

引　　言

语法学有句类、句型、句式等概念,这些概念只涉及语言表层的分类和分析,并没有揭示语言的深层结构。因此,语法分析不仅不能辨认所谓‘语法正确、语义荒谬’的句子,如‘无色的绿色思想在狂怒地睡觉’(Colourless green ideas sleep furiously),“所有的石头都死了”,也不能辨认大部分所谓搭配不当的语法错误,如‘秋天的北京是美丽的季节’,“盐在血液循环中起着重要地位”。

对上述‘语义荒谬’或‘语义搭配不当’错误的判断,显然已超出了语法的范畴。它们已不是语法层面、而是语义层面的研究对象,或者说,不是语言表层、而是语言深层的研究课题。从应用的角度来看,对此类错误的判断已成为自然语言理解处理的基本需要。这就是说,理论和实际应用两方面都产生了一种迫切的需要,就是在语义层面建立语句分析的理论模式和方法。这是一项任重道远的探索,本文只是一次尝试。

语言深层的根本问题是联想脉络的表述,联想脉络有局部和全局之分。简单地说,局部联想是指词汇层面的联想,全局联想是指语句及篇章层面的联想。这两种联想当然不可能截然分开,界限模糊永远是语言的基本特征。但局部联想和全局联想的概念仍然是有益和有效的。从这个意义上说,本文是关于语句层面全局联想的阐述,其姊妹篇[1]是关于局部联想的阐述。

在形成本文的过程中,是否创立一些新术语的问题曾困扰作者良久。借用语言学原有的术语而赋予新的含义是一种可供考虑的方法,但建立并穷尽自然语言语句物理表示式的总目标使作者在情绪上不愿意接受这种方法。所以,本文将先简要介绍有关的术语,然后转入正文的叙述。

语义块:语句的下一级语义构成单位。它可以是一个词、一个短语,甚至可包含另一个句子,或由另一个句子蜕化而来。在通常情况,一个语义块包含核心部分和说明部分,其核心部分也称为语句要素。语义块以其要素命名。

语句要素:简称要素,即语义块的核心部分,有主辅之分。

主要素:四种,是句类的函数。分别命名为:特征要素,作用者,对象和内容,分别用符号 EABC 表示。

辅要素:七种,也是句类的函数,但依赖性较弱。分别命名为:手段,工具,途径,比照,条件,原因,结果和目的,分别用符号 Ms, In, Wy, Re, Cn, Pr, Rt 表示。

主语义块 :以主要素为核心构成的语义块。四种主语义块以相应主要素命名,即特征语义块、作用者语义块、对象语义块和内容语义块。简称 E 块、A 块、B 块和 C 块,后三者也合称广义对象语义块,在形式上可统一记为 JK。A、B、C 实质上是广义对象语义块的基元表示或块素,A、B、C 可以连用,从而构成复合语义块。

辅语义块 :以辅要素为核心的语义块。通常带有语义逻辑概念的辅块标记 l1 或 l8。七种辅语义块也以相应的辅要素命名,对应的辅块标记为 l11 l17。在形式上辅语义块可统一记为 fK。

基本句类 :七种,分别命名为 :作用句,效应句,过程句,转移句,关系句,状态句,判断句。相应的符号为 :X,Y,P,T,R,S,D。每一基本句类又分为若干子类,子类的定义与相应主体基元概念网络的二级节点相对应。

各句类特征要素的符号 :与句类符号相同。

广义对象语义块的符号表示 :由句类标志和自身标志联合构成,如转移对象和内容分别记为 TB、TC,效应对象和内容分别记为 YB、YC。当需要标明子类信息时,则在两类字母间加数字符号予以标志,如物和信息的转移对象和内容分别为 :T2B,T2C,T3B,T3C。数字与主体基元概念层次符号第二层相对应。

句类格式及其标准格式 :句类格式定义为主语义块的排序。标准格式定义为主语义块的约定顺序。显然,标准格式与语种有关。本文给出的标准格式以汉语为参考。例如,汉语作用句、反应句、作用效应句的标准格式分别是 :

作用句	A + X + B
反应句	X2B + X2 + XAC
作用效应句	A + X + B + YC,YC = E + EC

这里,作用句的 A、B 前面省去了 X,这只是一项书写的约定。反应句是作用句的一个子类。作用效应句是复合句类之一。上列表示式就是语句的物理表示式,因为该式的每一项(语义块)都有明确的物理意义。关于语句的物理和数学表示式在【21】中有详尽说明。

混合句类 :是指两基本句类的混合,理论上应有 $6 \times 5 + 6 = 36$ 种。这里的 6 表示与作用效应链相对应的六个基本句类,6×5 是它们的两两混合,+ 6 是它们与判断句的混合。这里混合的意思是指一个 E 块同时含有作用效应链的两个甚至多个环节的信息,不是指一个语句里存在两个 E 块。

复合句类 :指一个句子存在两个甚至多个 E 块,而且它们分别含有作用效应链不同环节的信息。

2.1 语义块、E 语义块、句类及其格式

这是四个相关联的概念。

语义块的定义是 :语句的下一级语义构成单位。传统语言学把这个单位叫做短语。两

者的区别可以这样来表达,前者是语义和语言深层的定义,后者是语法或语言表层的定义。语义块与短语的本质区别在于:语义块可包含或嵌套另一语句,而短语不能。语义块按其语义功能分类。

首先是特征语义块 E。本节只讨论 E 语义块。为什么叫特征语义块?因为一个句子的基本语义信息就蕴涵在 E 块中。那么,什么是基本语义信息?这里不能不再次引用在【1】中已引用过一次的一段话:"作用效应链反映一切事物的最大共性,作用存在于一切事物的内部和相互之间,作用必然产生某种效应,在达到最终效应之前,必然伴随着某种过程或转移,在达到最终效应之后,必然出现新的关系或状态。过程、转移、关系和状态也是效应的一种表现形式。新的效应又会诱发新的作用,如此循环往复,以至无穷,这就是宇宙间一切事物存在和发展的基本法则,也是语言表达和概念推理的基本法则"。所谓一个句子的基本信息就是指它所表达的关于作用效应链的某一或某些环节的信息。这样,作用效应链的六个环节自然就是基本语义信息的分类标准,因而也是 E 语义块的分类标准。不同类别的 E 语义块构成不同类别的句子,从而引入了句类的概念。

E 语义块的命名当然应该与作用效应链六个环节的名称相一致,即:作用、过程、转移、效应、关系和状态。分别记为 X(作用)、P(rocess)、T(ransfer)、Y(效应)、R(elation)和 S(tate)。由这些 E 语义块构成的句子,分别命名为作用句、过程句、转移句、效应句、关系句和状态句。

这里对作用和效应采用了特殊的 X、Y 标记,这不是由于英语无相应的词汇,而是为了突出这两个概念之特殊重要性。因为,作用效应链的六个环节,也可高度抽象为广义作用和广义效应两极。这两极,又是言语表达的两个基本参照点,所谓主动式和被动式,就是分别立足于作用极和效应极的两种表达方式。但是,立足于效应极的表达并非一定要用被动式,是否采用被动句式是形式而非实质,汉语对效应极的表达就很少采用被动式。这一点,在文【14】【15】中有详细阐述。

一个语义块往往分为核心和说明两部分,这两部分多数情况以偏正或反偏正结构组合而成。语义块的语义主要取决于其核心部分。就 E 语义块来说,不难想象,其核心部分一定是动词,而且,不同类别 E 语义块的动词应来自于相应的基元概念。但是,对这一"想象"的后一点应立即加以补充:它只是充分而非必要条件,因为,如上所述,作用效应链的任一环节都可以充当广义作用或广义效应。这里,又是作用和效应扮演了特殊的角色。换句话说,上述'想象'对于过程、转移、关系和状态是正确的,这些 E 语义块的类别可由相应基元概念的层次符号唯一确定,但作用效应语义块例外。从理论的严谨性来说,例外的说法是不符合逻辑的,但这只是一个佯谬。奥妙在于概念的组合或组合概念。广义作用或广义效应是通过概念的组合体现出来的,在形成作用型或效应型组合概念时,其源可以是任何概念。关于这个问题,在文【1】中已有详细讨论。

上面关于 E 语义块的论述仅涉及三个超级语义网络之一的基元概念,细心的读者一定会想到尚未提到的基本概念和逻辑概念。在上述六个类别的 E 语义块中,状态类的核心部

分可包含基本概念的动词,因为状态概念 500 挂靠基本概念。选择状态基元概念与基本概念直接挂靠,其原因之一就是为了状态 E 语义块或状态句辨识的便利。状态句是所有句类中唯一可以没有 E 语义块的句类,这个问题将在文【19】中详细阐述。现在回到逻辑概念,显然,语言逻辑概念不会构成 E 语义块的核心,因为,它们只是语言表达的服务工具。可构成 E 语义块核心的只能是基本逻辑概念 jlv。事实正是如此,或者说,这正是我们专门设置 jl 类概念的原因。由 jlv 构成 E 语义块核心的句子定义为基本判断句。每个基本判断句就是形式逻辑的一个命题。主语和谓语的概念即源于此。所以,把语句作为命题来处理,实质上只适用于七个基本句类之一的判断句。判断是思维活动的基本内容,一般判断句的 E 语义块将由 v8 来标志,但 v8 只是一般判断句的必要而非充分条件。关于基本判断句和一般判断句将在文【20】中作专题论述。

上面,我们说明了 E 语义块的分类标准,它同时也是句类的分类标准。这个分类标准实际上也就是概念层次网络符号体系设计的基本准则,这样,E 语义块的辨识信息,或者说句类的辨识信息,就明确无误地蕴涵在层次网络符号体系之中。我们将把 E 语义块或句类的辨识叫做初级句类分析。显然,在句类分析时,对层次网络符号的运用,就体现了分析目的和分析方法的统一。但这只是问题的战术方面,更重要的是句类分析的战略性方面,这个问题将在文【3】中阐述。

上面的论述淡化了 E 语义块辨识的困难,现在需要对 E 语义块辨识的难点作一个概括性的说明(系统论述在【11】)。但这涉及到句类格式的概念,虽然这个概念的介绍最好等到全部主语义块说明以后,但提前介绍亦无不可。

所谓句类格式是指一个句子的主语义块排列顺序,例如,作用句必须有三个主语义块 A、X、B,三者的排列顺序不外乎六种:A + X + B,B + X + A,B + A + X,A + B + X,X + A + B,X + B + A。这六种格式都有充当标准的资格,但也有理由认为第一种格式比较符合天然的顺序,语言的实际情况也支持这一观点。在所谓 VO 型语言中,以第一种格式为标准的占 63%,汉语和多数印欧语,包括英语,都采用这个标准,而以第五、第六种格式为标准的仅分别占 34% 和 3%。当然,对自然顺序的理解显然带有民族性和地域性,中国人习惯的自然顺序是先整体、后局部,先共性、后个性,而西方人恰恰相反,所以东西方的地址表示顺序正好互相颠倒。因此,所谓标准格式必然具有语种性。本文给出的标准格式乃以汉语的习惯为准。

既然标准格式具有语种性,那引入这一概念的意义何在?答案是简单的 标准格式本身就是主语义块类别的辨识信息,因为,当句子取标准格式时,语义块的类别信息就隐含在它们的顺序中。只有当句子偏离标准格式时,才需要对某些主语义块加上类别标志符以利于辨识,这种需要是汉语虚词和西语介词的主要来源之一。当然,语言的运用是复杂的,打破上述约定的情况可能出现,例如,汉语作用句的标准格式是:

A + X + B

但‘红旗漫卷西风’这样的非标准格式并不给出语义块的类别标志。当艺术性与科学性发生

冲突时,语言倾向于把艺术性放在第一位,而牺牲科学性,但这种牺牲必须以不引起语义模糊为前提,这就要求 A、B、C 语义块具有鲜明的个性,不致相互混淆。“红旗漫卷西风”甚至“枕流漱石”的表达方式就满足这个条件。

E 语义块辨识的困难主要来于语义块的分离现象,而后者又与非标准格式相联系。让我们先看一下句例:

中科院声学所和自动化所正在联合研制汉语人机对话系统。

中科院声学所和自动化所正在进行汉语人机对话系统的联合研制。

汉语人机对话系统正由中科院声学所和自动化所联合进行研制。

中科院声学所和自动化所联合承担了汉语人机对话系统的研制任务。

这四个句子在形式上代表了作用句的四种格式:

第一句　A + X + B

第二句　A + B + X

第三句　B + A + X

第四句　X1A + X1 + X1C

前三句是作用句,第四句是作用承受句的标准格式之一[15]。第一句是作用句的标准格式[14],第二、第三句是作用句的非标准格式。前三句里的 A 语义块“中科院声学所和自动化所”,B 语义块“汉语人机对话系统”,都保持不变。但 E 语义块却分别出现了三种表达方式:

1　正在联合研制

2　正在进行……联合研制

3　正……联合进行研制

后两种表达方式里的 E 语义块是分离的。这一分离现象使得语义块顺序或格式的说法似乎有点模糊不清,其实不难加以澄清,就是语义块的位置以其核心部分为准。这里 E 语义块的核心部分是“研制”,而“正”或“正在”,“联合”、“进行”或“联合进行”都是说明部分。但是,问题需要展开:第一,这种分离现象只出现在 E 语义块么?第二,区分“核心”与“说明”部分的准则容易把握么?下面就来简单讨论一下这两个问题。

语义块应具有封闭性,即其各部分连在一起,不容许其他语义块插足其间,当句子改变格式时,语义块仅作整体搬迁。从上面的句例,我们看到了 A、B 两种语义块确实具有这种封闭性,那么,这一特例中的表现是否具有一般性?对此至少可以作出这样的判断,就是 A、B 语义块的封闭性符合思维的正常步骤或习惯,口语可能打破这个习惯,但书面语通常是遵循的。这就是说,广义对象语义块在标准格式里是不分离的,但在非标准格式里也可以分离。

E 语义块的分离可分为两大类:一是核心与说明部分的分离,二是核心本身的分离,主要是高层表示与低层表示的分离。第一类分离是人们比较熟悉的,它又有四种具体表现:核心与情态说明分离;核心与时态说明分离;核心与逻辑态说明分离;最后是核心与其他说明分离。这一类分离与语种密切相关,例如核心与时态说明的分离在西语是不存在的,但汉语

则经常出现,这是由于两者对时态采用了完全不同的语法手段。

E 语义块的分离现象,也可看作是一种词汇间的远搭配现象。在一般情况,词汇间的优先搭配主要是概念关联性的表现,语法功能仅居于从属地位。但 E 块分离现象却主要是语法功能的表现,第一类分离是如此,第二类分离更是如此。上列句例中"进行"与"研制"的分离就是第二类分离。这里,进行是高层概念,研制是低层概念。这种高低层概念远搭配的现象各语种都存在,不过相对说来,汉语也许比西语更普遍一些。因为对于复合概念,汉语更多采用组合而不是另造新词的方法。E 语义块的高低概念分离是概念组合的一种方式,从组合结构来说,这种分离的组合显然只适用于作用、效应和内容三种结构。由此可作出推论:E 的高低分离又是一个类别信息,表明该 E 优先于 X 或 Y,该句子优先于作用、效应句。语言无绝对规则,所以这里采用"优先于"这样的模糊表达方式,这是需要申明一下的,因为作者并未对上述推论作过语料求证。

概念层次的高低,在层次网络符号里一目了然,所以高低分离现象并不难发现。在自然语言里,用于分离 E 语义块的词汇也比较集中,如汉语的"搞干做令使"和"进行、从事、实行",英语的"get make take do let"。机器翻译的难点之一就是分离结构 E 语义块的转换。

最后,说明一下 E 语义块核心与说明部分的辨识问题。语言逻辑概念的层次网络符号的设计,从某种意义上说,是以语义块的感知处理为中心目标。这就是说,lik 所提供(定义)的信息都是直接针对语义块切分和组合的需要,其中 l6 和 la 专门用于指示 E 语义块的说明部分,它同 jlvu1k,k = 2,3 一起,包揽了 E 语义块的上述三种说明部分。情态说明一定是 jlvu1 类概念,时态说明一定是 luv6 类概念,逻辑态说明一定是 luva 类概念。以上关于 E 语义块的构成特性,在文【14】中有详尽讨论。

2.2 广义对象语义块

本节讨论主语义块的另外三类:作用者 A、对象 B 和内容 C。对于狭义作用句,作用者的意义等同于施事,对象 B 大体上等同于受事。但作用的语言表达不仅需要施事和受事的概念,还需要兼有施事与受事双重资格的概念,作用的承受者和作用的反应者就是典型的例子。它们先作为受事接受作用,随后又发挥作用而扮演施事的角色。对于作用效应链的过程和状态这两个环节,在通常的语言表达里,可以不涉及作用者,因而无所谓施事和受事。这里的表达对象通常只有一个,即过程和状态的体现者或承受者。与此相反,作用的表达对象必须有两个,施事和受事。当然,关系的表达对象也必然有两个,但关系的双方与作用的双方既有共性,又各有个性,通常不能用施事和受事的概念予以表述。转移的表达对象最为复杂,包括转移"物"、发出者和接收者、起点和终点、途径和工具,这里的"物"是广义的,包括信息。传统语言学里的所谓双宾语,就是指转移物和转移接收者,双宾语句一定是转移句。这里还应说明一点,途径和工具当然不是转移的"专利",但这两个通常情况下非必须的辅要素在转移句里往往是必须的。

上列语言现象表明:语义块的语义功能表述必须作为作用效应链的函数来处理。因此,对语义块的定义或命名,既不能沿用语法传统的大简化方式,也不能沿用格语法的大海捞针式寻求语义角色的方式。两种方式的共同弱点是缺乏总体指导原则,后者尤为严重。

HNC 语义块的定义方式是,在作用效应链的指导下,仿照概念层次网络符号的表达形式,对语义块的表达也采用字母串与数字串的混合结构。不过,这里的主语义块字母一律大写,辅语义块字母则采用一大一小的方式。字母与数字都有两种定义方式,如下表所示:

语义块字母定义	
类别字母	E,A,B,C;Ms,In,Wy,Re,Cn,Pr,Rt。
函数字母	X,P,T,Y,R,S,D。
语义块数字定义	
类别字母之后	相同类别语义块的各方。
函数字母之后	基本句类的子类, 通常与主体基元概念节点的二级甚至三级层次符号对应。

上面我们详细阐述了对语义块的表述方式,还没有直接涉及 A、B、C 的定义,在讨论这个问题之前,让我们先看几个句例,这样,既有助于上述表述方式的由虚返实,又有利于下面讨论的由实入虚。

张先生怕李小姐发脾气。	X2B + X2 + XAC
张先生怕李小姐的脾气。	
李小姐的脾气,任何人都害怕。	
张先生怕李小姐怕得要命。	
李小姐一发脾气,	XAC + X2B + X2 + X2C
张先生就不敢参加任何娱乐活动了。	
张先生一碰到李小姐发脾气,	
就吓得不敢参加任何娱乐活动了。	
张先生提拔李小姐当公关部主任。	A + XY + B + YC
李小姐破格提升公关部主任是张先生决定的。	
李小姐被张先生提拔为公关部主任。	
李小姐警告过张先生不得草率从事。	T3A + T3 + TB + T3C
李小姐对张先生说过不要草率从事。	
不得草率从事的警告,	
李小姐事先就对张先生讲过。	

这三组句例分别相应于反应句、作用效应句和信息转移句,在例句的右方给出了相应句类的标准格式,其中反应句有两种格式。反应句有四项要素:反应,反应者,反应者的后续表现,引起反应的作用者及其表现,它们分别用 X2, X2B, X2C, XAC 表示。这里的 X2B 兼有施事与受事的双重性,反应例句中的张先生是李小姐脾气的受事,又是娱乐活动的施事。反应表达的第一表达对象应该是反应者,所以,它构成第一标准格式的第一号语义块。第一标

准格式有两种特殊形式,一是 XAC 扩展为语句,二是省去 XAC 的表现而只留下对象 XACA,或者省去对象而只留下表现 XACC。对反应句,关键在于把握:第一,反应的作用者及其表现的可分可合特性;第二,反应者及引起反应的作用者都有表现,表现被称为内容,在符号上用 X2C 和 XACC 予以区别。

在上面的说明中我们采用了

$$XAC = XACA + XACC$$

的分解方式。这种分解方式是普适的,而且可以递推。如

$$EB = EBB + EBC \text{ 或 } EBC + EBB$$

$$EBB = EBBB + EBBC \text{ 或 } EBBC + YBBB$$

$$EBC = EBCB + EBCC \text{ 或 } EBCC + EBCB$$

每一步分解一般按对象和内容二分,因此,语义块构成的这种分解也叫做对象内容分解,当然也可以是不同对象或不同内容的二分。语义块的这种分解有良性与非良性之分,良性分解指对象与内容存在确定的顺序,如对象在前,内容在后,或反之。作用句的对象 B 就属于良性分解。但一般复合语义块的分解是非良性的,如例句中的 XAC。

第一组例句的两种标准格式,分别相应于语言表达的效应和作用两极。前四句是立足于效应极的表达,后两句是立足于作用极的表达。反应是效应的具体表现之一,第二种标准格式在形式上与作用效应句雷同是毫不奇怪的,实际上它就是作用效应句的特殊形式之一。但反应句的‘作用者’与一般作用者不同,它通常含有作用内容,且两者可分可合,所以采用了以 XAC 代替 A 的表达形式。作用者与作用内容的分合,对象与效应内容的分合是语义块切分、组合分析的难点之一,将在文【11】中专门讨论。

语义块 XAC 中的 XACC 或 XACA 都可以省略,甚至整个语义块都可以省略,这是该语义块表达的基本规则。第四个例句就省略了 XACC。这里顺便指出,此例句里的补语‘怕得要命’并不具备充当 X2C 的资格,因为它不是具体的表现,不过仅表明怕的程度而已。“要命”的映射符号是 ju60c44,比‘很’的程度还高一级,相当于汉语的‘极”,只能构成 E 的说明部分,是 E 块构成中的 HE 部分[14]。

例句显示了语义块切分组合和句类分析中一系列复杂的问题,这将在后文及本论文系列的其他文章中逐步展开讨论。现在让我们回到本节的主题,并从语义块 A、B、C 的定义谈起。我们已给它们分别取了中文名字:作用者、对象和内容,这三个名字所用到的词汇在《现代汉语词典》里都有权威解释,先将有关义项转录如下:

作用　　对事物产生影响

对象　　行动或思考时作为目标的人或事物

内容　　事物内部所含的实质或意义

事物的概念通常是不包含人的,所以对象词条里用了‘人或事物’的说法,但作用词条里的事物显然应该包含人。下面为避免行文的累赘,将在包含人的意义上使用事物一词。这样,对

作用者和对象 B,可仿效词典的表达方式,给出下面的定义:

作用者 A　　对事物产生影响的事物

对象 B　　语言表达时作为目标的事物或事物之一(除了作用者)

在对象的定义里用了"之一"一词,并加了"除作用者"的附注,这就是说,一个句子里的表达对象可以不只一个,作用者是作为特殊表达对象的特殊命名。但这个定义显然并没有把"表达对象"完全说清楚,或者更准确的说,它只对作用句是清楚的,这时的作用者和对象有明确的含义。在其他情况,表达对象的含义和数量都是不明确的。这里关于表达对象的定义只是基本定义。这就是上一节对语义块进行定义时,除了基本定义之外,还要另加函数定义和数字定义的原因。作用效应链六个环节所需要的表达对象有明显的差异。过程和状态只需要一个表达对象,就是过程和状态的的体现者,按定义,用符号 PB 和 SB 表示。关系的表达对象必须有两个,就是关系的双方,按定义,用符号 RB1 和 RB2 表示。转移的表达对象则比较复杂,在转移的八个表达对象中,仅定义接收者、起点和终点为转移对象,并分别用 TB、TB1 和 TB2 表示。转移的发出者定义为转移作用者 TA。转移"物"则定义为转移的内容 TC。这项定义表面看起来只是一个约定,但实际上涉及到语义块的语句扩展性这一根本概念,也涉及到对象和内容的定义。下面将从多个侧面对此进行阐述。至于转移的另外两个表达对象——途径和工具,则通常作为辅要素来处理。

一个语句的内容无非是两方面,第一是表达对象,第二是对象的表现。对表达对象,上面分为 A、B 两类语义块。B 是一般表达对象,而 A 是表达对象中的特殊对象。在一个句子里,对象可以不只一个,表现也可以不只一种,仿照对象的分类,表现也可分为一般和特殊两种,我们将一般表现定义为 E,将特殊表现定义为内容 C。这就是引入"内容语义块 C"这一概念的第一个来源。

上面例句中的"发脾气"不敢参加任何活动"当公关部主任"不得草率从事"显然都是表现。与他们搭配的"怕"提拔"和"警告"也是表现。对这两类表现如何区分一般与特殊?在一个句子里的表现如果不只一个,那么表现之间的关系显然有两类(这里不计简单的并合关系)。一类是一般的顺序关系,另一类是链式关系或因果关系,后者是一种特殊的顺序关系,这时可将"因"称为一般表现,将"果"称为特殊表现。所谓因果都是相对的,在作用效应链的链条上,汉语的"前因后果"这个成语可视为对因果相对性的绝妙说明。上面例句中的"怕"对前面的"发脾气"是果,对后面的"不敢参加任何活动"是因。同样,"提拔"与"当公关部主任";"警告"与"不得草率从事"也是因果关系。显然,也可仅把表现中的动词部分"不敢参加"当"从事"叫做果。然而,这只是枝节问题,问题的本质在于,在一个句子里,如同可以有两类对象一样,也可以有两类表现:因表现和果表现。表述对象的一般是名词,表述表现的一般是动词,从这个意义上来看,所谓"一个句子只有一个中心动词"的语法规范与语言表达的需要并不协调。也许下面的陈述方式更符合语言深层的结构规范:一个句子至少由一个对象语义块和一个表现语义块构成,但更为常见的结构是:两个对象语义块加一个表现语义块,一个对象语义块加两个表现语义块,两个对象语义块加两个表现语义块。这四种语

句构成方式可简称为 1 + 1,2 + 1,1 + 2,2 + 2 句式。这里我们借用了“句式”的术语,但赋予了新的涵义,它表达一个语句的语义块构架信息。

句式规定了一个语句主语义块的数量和类型,这个信息对于语句分析的重要性不言而喻。问题是,我们能在进行分析之前或之初,获得这一信息么?答案是:在语义块感知处理[11]以后,多数情况都能得到。这一信息的主要来源在词汇层面,可通过语义结构方程予以表示,详细论述见【7】。在概念层面也能取得这一信息,它是句类知识的重要组成部分,在【14】到【20】中结合具体句类有具体说明。

关于句式,这里需要补充四点。第一,在出现两个对象时,可以两个都是 B,或一 A 一 B,但绝不可能两个都是 A,同样,在出现两个表现时,可以两个都是 E,或一 E 一 C,但绝不可能两个都是 C,这就是对 B 和 E 冠以一般,而对 A 和 C 冠以特殊的原因。第二,句式的“2”可以扩展成“多”;“2”实际上包含“多”。第三,2 + 2 句式的两个表现通常是上述一般的顺序关系,而不是因果关系。对第二表现是果表现的情况,将专门命名为作用效应句,并当作基本句类之一来处理。第四,存在最简单的 1 + 0 或 0 + 1 句式。

上面我们从句子整体结构的角度引入了因果表现的概念,并将果表现纳入 C 语义块。果表现就意味着“C 语义块可扩展为另一语句”,或简称 C 的语句扩展性。因为,所谓果表现,就是“新的效应又会引发新的作用”这一基本观念的具体体现。也就是说,语句表达时,将作用效应链再循环的功能交给 C 语义块来承担,而再循环的表达当然又需要一个语句。上面的例句体现了 C 语义块这一特点。

因果表现的表达需要,是引入两类表现概念的第一个来源。还有第二个来源,就是语义块的构成方式。

上面的论述似乎给读者一个错觉:对象和表现语义块是截然分开的,要么是对象语义块,要么是表现语义块。所以,这里需要申明:除了单纯描述对象和表现的语义块之外,还有同时描述对象和表现的复合语义块。这种复合语义块可以采用偏正结构的形式,如上面例句中的“李小姐的脾气”,也可以采用句子的形式,如“李小姐一发脾气”。这就是说,语言的表达对象及其表现可以融合在一个语义块里。应该把具有这种融合性的表现与不具有这种融合性的表现区别开来,我们把前者叫做内容 C,把后者叫做特征表现 E。

两类对象,两类表现,表现与对象的融合性,果表现的语句扩展性,这四点,是形成 E、A、B、C 四种主语义块概念的理论依据。融合性意味着 A、B、C 实质上是广义对象语义块的构成基元。

但是,上述四点,只是语义块的一般特性,要对具体语句的具体语义块进行具体分析,还必须把四种主语义块作为句类的函数来处理。RB1 和 RB2 是关系的双方,两者在关系句里都是不可或缺的。TB1 和 TB2 是转移的起点和终点,但它们在转移句里并非不可或缺,仅在转移句的子类——自身转移句里,必须有两者之一,或在少数情况下同时涉及两者。简单作用句可以不涉及内容,但信息转移句必须具有内容,而且具有语句扩展性。这些是句类的基本知识,且以标准格式的形式给出。

把 EABC 语义块作为句类的函数来处理是一个重要的发展,但绝不是因此而万事大吉。句类之间仍然存在一定的模糊性,这就会造成语义块类别的模糊性,例如作用对象和效应对象的区分就是一个比较复杂的问题,对象和内容的区分,在许多情况也相当复杂。这些专门问题在〈问答 32〉中有详细阐述。

AEBC 四类语义块的提法,在形式上似乎与传统语言学的主谓宾补的提法相对应,可以大体给出下表所示的对应关系:

E	谓语	(补语)
A	主语	宾语
B	宾语	主语
C	补语	宾语 谓语

显然,这种对比无实质性意义,E、A、B、C 是语言深层的语义描述量,是句类的函数,但与句类的格式无关。主谓宾补恰恰相反,它是语言表层的语法描述量,不管句类,但与句类格式息息相关。质言之,E、A、B、C 是语义层面的概念,主谓宾补是语法层面的概念。两者从不同层面或角度对句子的结构提出分析的模式,不能相互代替。

上面给出了主语义块表示式的一般规则,这些表示式可称为语义块的物理表示式,这些物理表示式按一定顺序‘相加’,就构成语句的物理表示式,在上面给出了 4 个示例,这些物理表示式就是所谓语言深层结构的具体表达,因为它们提供了语句的全局联想脉络,能形成文【1】引言中所说的预期和判断能力。在功能上是对大脑语言感知过程的适当模拟。

按照 HNC 理论的作用效应链框架,自然语言语句物理表示式的完备性在理论上已十分清晰,即 7 个基本句类和 36 个混合句类的基本模式。按文【1】所范定的自然语言概念体系,基本句类一级子类的物理表示式是可以穷尽的,我们正在为此而努力。

最后,简单叙述一下 E、A、B、C 概念的形成过程,这对于加深对这一概念的理解或许有所裨益。与主谓宾补相联系,语法学还有动词的及物和不及物以及双宾语等概念。但及物性的具体表现,仅在语法层面进行研究十分困难,它涉及宾语的分类问题,有的及物动词要求双宾语,有的不仅要求宾语,还要求补语。这些问题都必须进入语义层面,才能给出明确的答案。从理解来说,仅有及物的概念是远远不够的,重要的是:‘它’及‘什么样的’物”?开始的时候,曾以为这只是词汇层面的特性,后来才发现不是这样,它也是概念层面的重要特性,这一发现导致‘语义块函数’概念的形成。但应该说,是格语法理论的创立者菲尔墨最先想到了这一点,他是对宾语和主语进行语义分类的第一位先行者。可惜他的理论匆忙出台,在理论的总体性和层次性方面都十分欠缺。现在看来,主语和宾语的语义分类必须用 A、B、C 函数的概念,即将语义块作为作用效应链的函数来处理才能给出完善的表述,才能彻底消除格语法理论关于语义角色完备性的困惑。至于双宾语,它一定是转移型概念。而同时要求宾语和补语一定是作用效应型概念,这些信息现在都能通过层次网络符号和语义结构方程[7]给出比较精确的表述。

2.3 辅语义块

句子除了主语义块之外,还有辅语义块。辅语义块如何分类?怎样对它进行总体研究?

本文采用的方法是:对汉语的全部语言逻辑概念进行对比分析和综合归纳,在剔除主要素指示符以后,由剩余指示符的分类得出七类辅语义块。它们是:

手段　Ms(Means)
工具　In(Instrument)
途径　Wy(Way)
参照　Re(Refer)
条件　Cn(Condition)
因　Pr(Premise)
果　Rt(Result)

当辅要素在句子里出现时,通常都带有相应的语言逻辑概念指示符予以标志。但辅要素的指示符与主要素指示符有两个重要区别,一是它不仅仅是一个指示符号,自身还有独立的逻辑意义;二是这个指示符的有无与语句格式无关,即与主语义块的排列顺序无关。l 语义网络的设计,可以说是以辅要素及 E 要素说明部分的语言逻辑概念为核心,这两者才是语言逻辑概念的真正代表,主要素的指示符仅居于从属地位。

在这七个辅要素中,条件、手段和途径处于特殊地位。条件和手段在句子中出现的频度最高,相应的词汇十分丰富。这当然不是偶然的语言现象,任何作用效应的完成都必须具备一定的条件,而一个句子无非是对广义作用效应(包括判断)的描述,因而离不开条件。语言描述的主要对象是人类活动,而人类的活动不仅必然受到条件的制约,还必须采取适当的手段和途径。这三项辅要素所涉及的概念具有三大类抽象概念的综合特征,为此专门设置了综合类概念 s。

s 类概念高层层次符号的前两级如下表所示:

s1	途径
s11	具体途径
s12	策略
s2	手段
s21	方式
s22	方法
s3	条件
s31	时间条件
s32	空间条件
s33	社会条件

s34　　一般物质条件
s35　　前提条件
s4　　工具
s41　　具体工具
s42　　材料
s43　　原料
s44　　能源

s 类概念的设计经历了十分曲折的过程,因为它的综合特征颇令人困扰。其高层节点的排序也很难抉择,上面表中的排序乃基于下列考虑:途径是战略性的,手段是战术性的,故途径先于手段。条件是战略和战术决策的依据之一,故列为三者之末。最后的工具是广义的,包括人类创造的一切手的延伸物。

具体的语言词汇往往不严格区分途径、方式和方法,这是语言概念从脉络性模糊走到了违例性模糊边缘的不良表现,但概念节点的设计不能跟随自然语言的这种不良倾向而"随波逐流"。s 类概念的设计贯彻了这一原则。

s 类概念的条件与基本概念的联系最为密切,时间、空间、数量和质量显然是最基本的条件。所以,曾有过将条件与基本概念挂靠的设想。但条件不同于通常的物和人,挂靠的近似性很差,甚至不可能通过多次挂靠而逼近真值或改善近似程度,因此必须独立设计。时间条件概念"时机"就是一个很好的例子。

辅要素的另外三个概念,分属三种不同情况。作为语言逻辑概念的因与果,都有可挂靠的基元概念,挂靠对象分别是 121,9121 和 122,b0。参照这个概念,是语言逻辑概念中唯一挂靠基本逻辑的概念,挂靠对象是 j100 和 j102 之并,因此,挂靠层可以省略。

辅语义块的辨识一般比主语义块简单。虽然汉语的辅要素逻辑指示符多数借用于实词,但由于指示符与指示对象之间的语义关联性很强,通过语义距离的计算不难消除这一多义性模糊。主辅语义块的辨识处理在文【11】中讨论。但这里要谈一下主辅语义块的划分标准问题。

在一个句子里,主辅语义块的划分由表达的总体需要所决定,这个内在标准,软件很难直接把握。在语义块辨识时,判断的唯一依据只能是语义块标志符。到句类分析的后期,才有可能根据各语义块的内容作出实质性的判断。

这里还应该强调一点,综合类概念虽然与辅块内容有天然的联系,但不能孤立地把它们的反映射词汇作为辅语义块的判据,因为这些词汇和概念照样可以构成主语义块的内容。但是,如果这些词汇或概念与相应的语言逻辑概念搭配在一起,那就成为双保险的可靠判据了。

2.4　句类分析概述

句类分析同层次网络符号和概念组合结构[1]一起,构成概念层次网络理论的三极。简

单地说:符号是基础,结构是手段,分析是目的。

上面已经指出:所谓语句的深层结构,就是对作用效应链各特定环节的一种描述。描述的基本方式就是 EABC 四种主语义块的适当配置,即语句的物理表示式。不同的环节要求不同的配置,句类是不同环节的标志。作用效应链有六个基本环节,所以,相应的基本句类也是六个,而 EABC 是句类的函数。

六个基本句类主语义块的具体配置及具体的句类知识将分别在文【14】到【20】中作专题讨论。这里只作一般性阐述。

自然语言理解处理的基本使命是消除言语中的五重模糊。这五重模糊依次是:语音模糊,音词转换模糊,词的多义性模糊,语义块切分组合模糊,缺省、重复及指代性模糊。模糊的五重性只是形式表现,本质上都可归结为多义性模糊,因此,解模糊处理实质上是多义选一处理。人在听口语或看书面语的过程中,大脑里最基本、最频繁的操作就是进行多义选一处理,这是毫无疑义的,虽然目前我们对它的具体运行机制所知甚少。

多义(包括歧义)选一处理的一般原则是众所周知的,就是依靠上下文的联想。但是,如何进行上下文联想的处理?概念层次网络理论的答案是:上下文联想处理有近程、中程和远程之分,近程联想是指语义块内部的联想,中程联想是指语义块之间的联想,即句子内部的联想,远程联想是指句子之间的联想,包括基于要点主题分析的篇章级联想。中近程联想处理的基础是进行概念之间语义距离的计算,把最小语义距离作为多义选一的基本判据。概念层次网络符号及基于这一符号体系建立的概念知识库[6]和语言知识库[7][8]为语义距离的计算提供了必要的知识。具体的计算方式将在【3】中讨论。但是,语义距离计算或不同层次的联想处理都必须建立在句类知识的基础上。我们将把句类知识的运用叫作句类分析。

这里,让我们来看一个具体的例子。汉语的"去"字,不计其语言逻辑意义的独立义项有下列四个:v22b,v22b2,v00 # (v382,v312) &w,v382。在"去上海""扬长而去""去皮吃""去伪存真"里依次取上列义项。这里的多义选一不都是简单的四选一问题,因为其他的词汇也可能有多个义项。但不论选一处理简单或复杂,只要把相应词汇的义项转换为层次网络符号,这里的多义选一问题都可以解决。"上海"的映射符号是 fpwj2,它与自身转移 v22b 优先搭配,"扬长"的映射符号是 uv22b2,与离开 v22b2 是狭义同行优先。"皮"的映射符号是 jw61 - ,依据这一高层表示也可以判定,它应与对象优先于 w 的作用型概念第三义项优先匹配。"伪、存、真"的映射符号分别是:j812,v381 和 j811,它们与舍弃 v382 又是同行优先。

此例的多义选一处理似乎都是迎刃而解。但实际上知识运用的层次很不相同。"扬长而去"最为简单,是语义块内部的模糊消解问题,这里只需要使用同行优先准则。即使如此,首先也得建立起语义块内部的认识。"去上海"和"去伪存真"则复杂一些,它们都涉及语义块之间的模糊消解处理,尽管语义距离计算不难消除多义模糊,但起码的理解仍需要用到自身转移句[17]和效应句[14]的句类知识。对"去伪存真",还需要运用汉语语义块构成的对仗性知识。"去皮吃"则更为复杂,这里的复杂性不仅在于要消除"去"和"吃"的多义模糊,还必须进一步明确"去皮"和"吃"的组合结构,这就需要物转移句的一个二级子类——生理活动

物转移句的句类知识。

总之,多义选一处理在形式上只是语义距离的计算,但知识的运用则涉及多个层面:从词汇层面到语境层面。不同层面知识的运用需要一个总策划,一个调度中心,这个中心就是句类知识,这个调度过程就是句类分析。

多义选一处理贯穿于理解处理的全过程,预处理的"分词层选"[9]和语义块感知[11]就进行了大量的模糊消解处理,包括多义选一处理。但是,这两步处理不可能消除全部模糊,剩余的模糊只能留给句类分析来处理,这就是说,句类分析面对的模糊将是一些"难啃的硬骨头"。

那么,句类分析有什么新的招数?这是"坐井观天"与"高屋建瓴"的区别。前两步处理对知识的运用是局部性的,句类分析对知识的运用是整体或全局性的。但是,句类分析并不对"层选"和"语义块感知"处理搞"包办代替",有关问题仍反馈回去由它们自行处理[3]。在这个新的背景下,两预处理模块对知识的运用也就冲破了局部性的限制。

虽然多义选一处理对模糊消解具有重大的实用价值,句类分析在其中担负着最后拍板的关键角色。但这终究只是句类分析的初级内容。从某种意义上说,通过语义距离的计算作出多义选一的处理,仅仅是在功能上模仿了大脑的思维效果,根本还谈不上真正的理解。

要向理解的殿堂真正跨进一步,就必须能够对语句的合理性作出一定的判断。这是句类分析的第二项内容。语句合理性分析不同于多义选一处理,后者是机械式的和死板的,因为语义距离总有一个最小值。而你很难制定一个不能搭配的语义距离阈值。但基于语义距离仍能对合理的程度给出某种评估。像本文开头所引的乔姆斯基例句"无色的绿色思想在狂怒地睡觉"和邢公畹先生的例句"所有的石头都死了",层次网络符号可以给出有关概念毫不关联的明确信息,因而句类分析不难对这类语句的不合理性作出判断。

结 束 语

本文是【1】的姊妹篇,试图在语义层面提出一个语句分析的理论模式。

我们把这个理论模式叫作句类分析。

句类分析的起点是语义块的辨识和句子类别的辨识,后者简称句类辨识。

语义块的辨识包括语义块的切分、组合和类别辨识三方面的内容,统称语义块感知处理,将在【11】中作专题论述。但这里可以预告一个要点,就是语义块感知处理的基本武器是lv准则,这个准则里的 l 就是指文【1】所阐述的 l 类概念。

语义块是句类的函数,由语义块物理表示式构成的语句物理表示式就是语句的深层结构。这些语句表示式是句类分析的立足点,而且,自然语言的物理表示式是可以穷尽的。这是本文的基本论点。

因此,语义块感知处理实际上离不开句类辨识。

句类辨识的信息则来于【1】所阐述的基元概念和组合结构,特别是其中的主体基元概念

和作用效应型组合结构。

这就是说,本文所提出的语句分析模式的理论基础是概念层次网络理论。关于基本句类的内涵,包括它的格式和句式,本文只作了最高层次的简要说明。

每个基本句类都分成若干一级子类,一级子类之下,又分二级子类。这些问题不可能在本文展开讨论,它们是文【14】到【20】的基本内容。

对句类分析的立足点,即语义块,本文则作了详尽的说明。在四种主语义块中,C 语义块扮演了特殊的角色,文中着重阐述了它的对象表现双重性、扩展性及融合性,其中的融合性又形成语义块构成的重要特征,对此,在【14】的第 2 节中有进一步的说明。

1994 年冬

后　记

在〈论文〉系列预定的 21 篇论文中,本篇是最早动笔的一篇,这次作了一些文字上的修改。文中关于基本句类一级子类物理表示式可以穷尽的预测已经实现,这是 HNC 联合攻关组在 1997 年的重大成果之一。本书附录中以刘志文为首的论文“自然语言语句的 HNC 表示”和苗传江的论文“HNC 理论的句类”对此有更详尽的阐述。关于基本句类的详细清单和说明,读者还可通过因特网参阅《HNC 句类知识手册》。

本文后来的重要发展有:句类变换;广义对象语义块的分离;两可(主辅两可)类语义块的定义和处理;对句类非标准格式引入了规范和违例两种格式的划分,并据此穷尽了句类格式的各种变化。这些都在〈HNC 理解处理的 52 个论题〉中有所阐述。

1998 年 8 月 18 日

HNC 理解处理系统的基本框架

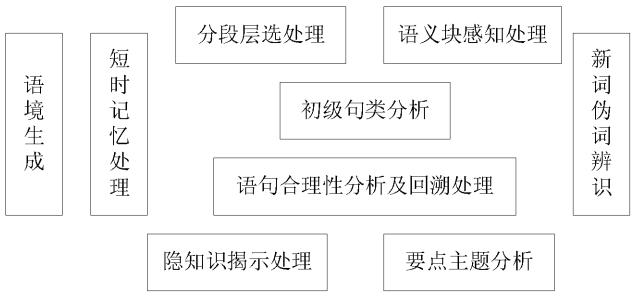
引　　言

本文讨论概念层次网络理论[1][2]的应用,但不涉及技术实现的细节。主要是说明:语义层面自然语言处理基本框架的构成及其主要功能模块,它们所需要的知识及知识运用的方式和难点,概念层次网络理论及 HNC 知识库可提供的信息和知识的局限性。

本文先给出基本框架的模块框图,随后各节对其中的某些模块作理论性说明。

3.1　基本框架的一般性说明

基本框架的构成:



这个基本框架由四个层次的处理构成。共九个基本模块。

第一层次的两个模块属于预处理,本论文集的【9】【11】两篇对它们作了专题论述。第一个模块"分段层选处理"是汉语的特殊需要,对于文字文本,它确实是一个额外的负担,但对于语音文本,它却是汉语的"秘密武器"。这一点将在文【4】中说明。

第二层次是初级句类分析,将全面运用句类知识,消除剩余的模糊,并对语义块感知处理的结果进行检验。

第三层次是语句合理性分析,亦称中级句类分析,是整个框架的核心。

第四层次有两个处理模块:隐知识揭示和要点主题分析,亦称高级句类分析。

上述六个模块是主体模块。

此外,还有三个配套模块,一是左边的语境生成和短时记忆,二是右边的新词伪词辨识,

最后一个模块对汉语尤为重要。

为便于说明这一框图的意图,下面先给出一组例句。希望通过这些例句对语言理解处理,特别是汉语理解处理面临的基本问题给出一个较为明朗的轮廓。这些例句模拟声调完全模糊的输入文本,其模糊度接近语音识别输出的语音符号阵列,但远大于汉字拼音输入的模糊度,因为,以拼音方式输入汉字文本时,可以人为地部分给出单音词、双音词和多字词的标志信息,并且可以无模糊地输入一些极高频度的单音词。这将大大减轻理解处理的负担。

对每一例句给出分段处理[9]后的结果,音段间隔用‘…’标记。三音节以上的音段通常是两层,出现多字词时,将形成第三层,用( )号表示。汉语的理解处理从层选开始。

1 dang ran…zhu yao…wei xian…bu zai bo hei…fang mian

当然 主要 危险 * 不在 波黑 方面

载波

ye…bu zai…mo…si…ke

不在

(莫斯科)

2 guan jian zai…bei er ge…lai de…he…bo hei…sai…zu…de…fan ying

关键 贝尔 来得 波黑 反应 *

健在 儿歌 *

(贝尔格莱德)

3 sai…er…wei…ya…qiang ren…mi…luo…she…wei qi…

强人 围棋 *

(塞尔维亚)

zai…kong xi qian yi tian…jiu…de dao…xiao xi

空隙 * 迁移 * 得到 消息

喜钱 一天

4 xi fang zheng zai…zhun bei…qian…suo wei you de…fan ji xing dong

西方 * 正在 准备 所谓 有的 反击 * 行动

方正 惟有 畸形 *

(前所未有)

5 ta…mei you wei ci ti chu…kang yi

没有 *为此 * 提出 * 抗议

有为 * 磁体

ye…mei you qu xiao di er tian he mei guo…te shi de…hui tan

没有 * 取消 * 第二 天河 美国 特使 会谈

有趣 * 小弟 二天 褐煤 使得

6 bo hei…jun fang xiang mi…luo…she…wei qi quan li…tiao zhan

波黑　军方　详密　围棋＊　权利＊　挑战

方向＊　齐全＊

7 qi tu…tong guo…zhan zheng qu dai…ta… cheng wei…sai…zu…de…shou ling

企图＊　通过　战争　取代　成为＊　首领＊

争取

8 xi fang…xue ruo…jun fang…li liang…dui…ta…lai shuo

西方　削弱　军方　力量　来说

bu shi yi jian…huai shi

不是＊意见＊　坏事

适宜＊

9 xing dong…kai shi hou…ta…yi mian xing shi shang…kang yi

行动　开始　一面＊　形式＊　抗议

时候　免刑　时尚

yi mian…he mei guo…te shi mi tan

一面＊　褐煤　特使　密谈＊

美国　失密＊

10 ta…xi wang…yi jian…shuang…diao

希望　意见

（一箭双雕）

ji…bao liu…sai…zu…shou ling di wei

保留　首领＊　地位＊

领地

you…neng…huo de…lian he guo jie chu…jin yun

获得　联合　国界＊　禁运

杰出＊

（联合国）

wan jiu ri yi e hua…de…jing ji

挽救 日益恶化　经济＊

旧日 衣蛾

现在先来对这一组例句作四项考察。一是层选的可行性，二是语义块切分组合处理的可行性，三是句类辨识的可行性，四是解模糊的可行性。

——层选考察：

层选处理先要区分奇音段和偶音段[9]，两者的情况如下表所示：

奇音段情况

例句序号	内容	优先选取的单音词及其位置序号	
2	guan jian zai	zai	3
3	kong xi qian yi tian	qian	3
5	te shi de	de	3
9	he mei guo	he	1
	kai shi hou	hou	3
	yi mian xing shi shang	shang	5
10	lian he guo jie chu	三字词	

偶音段情况

例句序号	内容	优先选取的单音词及其位置序号
1	bu zai bo hei	无
4	xi fang zheng zai	无
	fan ji xing dong	无
5	mei you wei ci ti chu	无
	mei you qu xiao di er tian he mei guo	tian，he
6	jun fang xiang mi	伪四音段
7	zhan zheng qu dai	无
8	bu shi yi jian	无
9	te shi mi tan	无
10	shou ling di wei	无
	wan jiu ri yi e hua	无

奇音段至少有一个单音词或一个三字词。偶音段通常无单音词，或是下列两情况之一：两个单音词，一个三音词加一个单音词。这是汉语音段的基本规律。加上单音词的位置信息[9]，单音词的语音信息[8]，层选处理不难得到表中所给出的正确结论，除了第六句的伪四音段。这里应该指出：例句奇音段的层选都十分简单，甚至都不需要进行语义距离的计算。

——语义块切分组合考察：

根据 lv 准则，可得到下表所示的切分组合信息。对其中的不确定者用‘？’标志。

例句序号	切分标志	组合标志	E 标志	E 要素
1				不在
	ye			不在
2		he，de		zai
3	zai		jiu	得到
4		de？	正在	准备（反击）
5			没有 *	提出 *（抗议）
	ye，he？	de	没有 *	取消 *（会谈）

6				挑战
7	通过？	de		企图＊，取代，成为＊
8	dui			削弱，不是＊
9	he？一面＊			开始，抗议，密谈＊
10				希望，保留
	ji，you		neng	获得，杰出＊，禁运
		de	日益	挽救

由这些信息可得到下表所示的语义块初步切分方案：

例句序号	语义块数 量	E 要素映射符号	语句类型	层选关联	上下文关联
1	3	¦jv01；¦jv21；v146	势态判断句		
2	3	jv01	势态判断句	加强 zai	继承
3	3，1	v3a1；v218	信息接收句	加强 qian	消息待补
4	3	v539115	过程势态句		
5	3	vgc022	反应句	澄清	
	2	v900 # v312	作用-信息交换句	tian he	
6	3	vrb30	竞争句	澄清伪四音段	继承米氏
7	5，1	vgb0，v24a	追求-替代		继承军方
		v5330a	-效应句		米氏
8	3，1	j1v112	基本判断句		继承米氏
9	1，1	vgc022	反应句		继承米氏
	2	vgc249	信息交换句	加强 he	
10	3	vg1121	反应句		继承米氏
	2	v381	效应句		
	4	v3a1，v903	效应-免除句		
	2	v00 # v13a1	作用句		

此表列举了语义块感知处理的初步结果，也给出了句类辨识及解模糊处理的一些结果。

表中给出了各句的语义块个数，其中第二个数字表示辅语义块个数。对于 1、2、3、7、8、9 句以及第 10 句的 1、2、4 小句，这只是 1v 准则的简单应用。对于 4、5 两句，则需要运用 E 语义块的构成知识[14]作相应的并合处理[9]。如第 4 句的“正在”与“准备”并合，第 5 句的“没有＊”与“提出＊”及“取消＊”并合。

表中还给出了句类辨识的结果，如“语句类型”列所示。辨识的判据主要是该列左边所给出的 E 要素层次网络符号。关于这些判据，本论文系列的【14】到【21】篇有详细论述。

语义块切分组合处理与句类辨识不可能截然分开[11]。语义块感知处理就包括语义切分、组合、句类辨识三项内容。分段—层选—语义块切分组合—句类辨识是相互依赖的处理

过程,前者为后者准备条件,后者又为前者消除遗留的疑难。把他们统称语义块感知处理亦无不可。第 6 例句就充分表明这种相互依赖性。层选处理在这里毫无作为,语义块感知处理直到最后一个词才找到第一个感知信息“挑战”。但是,“挑战”这个词的层次网络符号 vrb30 给出了句类的充分信息,由此可以假定,这是一个特殊的混合句类—竞争句[21],此句类必须有关系的双方,在双方之间应有逻辑指示符 102。根据这一句类知识,就可以返回去找出竞争句两语义块 RB1 和 RB2 之间的标志符 xiang,从而确定 RB2 为“mi luo she wei qi”,与此同时也就消除了伪词“详密”和“围棋 * ”的干扰。在进行这些推理时,竞争句句类知识的运用是关键所在。到此为止,只剩下双音词“权利 * ”的模糊未解。“权利 * ”的集合是(权利,权力,全力),它应该代表竞争的内容,因此排除了“全力”,但“权利”与“权力”的选择则比较复杂,涉及隐知识的揭示。

这里应该对“mi luo she wei qi”的判定有所交代。它属于新词辨识中比较简单的一类,即人名、地名或其他名称的确定。这类新词的征兆之一是单音段的连续出现,“mi luo she”正是如此。但这一名称具体属性的判定,例句 6 则远不如例句 3 方便,后者的前面有明确的指示信息“强人”。而由例句 6 的句类知识则不能得出一定是人名的结论,因为它也可以是一个社会组织的名称。在第 6 句的“上下文关联”栏里的注释“继承米氏”,就是表示应该继承第 3 句对“mi luo … wei qi”的判断。当然,这就需要“短时记忆”处理模块的配合。

对句类辨识的考察下面不再列表说明,因为没有这个必要。层次网络符号体系总体设计的目标之一就是为句类辨识和分析服务[1][2],E 要素的层次网络符号通常可以给出句类的充分信息。在 E 要素本身存在多音词的模糊或词的多义模糊时,由于还有一系列其他的辨识手段[9],通常仍不难作出句类的确定性判断。万不得已时,可采取“先假设,后求证”的办法,正是“天无绝人之路”。句类辨识是理解处理关键的一步,因为只有确定或假设句类以后,才能运用相应的句类知识,而只有依靠句类知识,计算机才有可能模仿人的思维方式,有效地激活语句要素之间的联想,立足于概念之间的关联性作出各种各样的判断,并从初级联想处理跨入中级联想处理,即从初级句类分析进入语句的合理性分析。但应该指出:语义距离或概念关联性的计算要区分语义块内部的核心部分与说明部分之间、语句不同要素之间、前后语句之间等不同情况,这将在下文作简要讨论。句类知识的运用是句类分析的关键,也是消解模糊的强有力手段。但也应该指出,模糊的消除,有的需要句类知识,有的并不需要。把握这一特点,是语义层面处理程序必须具有的功能。

上面对例句 6 的具体说明不过是句类知识应用的沧海一粟。关于句类知识,本论文系列的【14】到【21】篇有专门论述。但应该指出,句类知识仍只是知识海洋的一小部分,计算机即使完全掌握了句类知识,并能充分运用【6】到【8】中所阐述的概念层面和语言层面的知识,它对语言理解的深度和广度仍难以与人相比。但是,也应该指出,面对理解处理当前最迫切需要解决的大量模糊问题[4],通过初级及中级句类分析,计算机的解模糊能力有可能接近甚至达到人类的水平。当然,目前这只是一个理论上的预测,这一预测的证实或实现,需要在知识库的建设以及理解处理的软件设计和开发方面作出重大的努力。本文希望能对这一预

测的前景多作一点理论上的说明或探索,所以下面对每一例句都给出句类辨识及句类知识运用的简要说明。

例句 1:E 要素多义,它的三个义项可分别相应于判断句、状态句和过程代谢句,但不论是哪一种句类,E 的前面都应该是 B 语义块。具体的句类决定于 B 要素的概念类别性:抽象概念优先于判断句,w 类概念优先于状态句,p 类概念优先于过程代谢句。此句的 B 要素"危险*"有四个待选(危险,为限,纬线,胃腺),后两个属于罕用的专业性词汇,可依据'语境原则'暂时予以排除;"为限"是优先带前搭配'以'的动词,这里既无此搭配,又与前面的'主要'排斥,且不符合 B 要素的要求,因而也可排除。这样只剩下双字词'危险'可选,从而完成了双音词'危险*'的四选一解模糊处理。危险的映射符号是 r53322,是一种有害的势态,从而得出该句为势态判断句的结论。

例句 2:此句的单音词 zai 和双音词'反应*'都可能是 E 要素,但后者由于前面有组合逻辑指示符 de 而失去了这一资格,zai 成了唯一的 E 候选者。这样,单音词 zai 极大的多义模糊也已基本消除,基于与第一句相同的推理过程得出判断句的结论。在表中标明势态判断句,是参考了上一句的结果。此句中尚有未定的新词 sai zu,但不影响上述判断。

例句 3:此句的双字词'得到'由于前面有单音词 jiu 而加强了它充当 E 要素的优先性,它同后面的'消息'一起确定了这是一个信息接收句[17]。其标准句类格式是 T1B + T1 + TC。现 TC"消息"在 T1 之后,表明此句符合标准格式。在 T1 之前的两个语义块,必然是一主一辅。这就是句类知识的运用,它加强了单音词 zai 是辅语义块指示标志的判断,也加强了层选处理对五音段 kong xi qian yi tian 的处理,与此同时,还能确定'强人'之后的"mi luo she wei qi"是人名,从而消除了伪词'围棋*'的干扰。

例句 4:此句有两个动词'准备'与"反击",它们既可分别独立充当 E 要素,也可以合在一起构成复合 E 要素(因为'准备'可充当 vv 类概念),如果'反击'不构成 E 要素,此句是过程势态句,如果'反击'构成 E 要素,此句是关系作用句。两者的句类格式分别是 PB + P1 + PC 和 A + XR + RB2。用这两种句类格式对该句进行句类检验,很容易肯定前者而否定后者。因为,第一,"行动"一词不符合 RB2 的要求,但符合 PC 的要求;第二,如果选用四字词"前所未有",则余下的 de 只能充当语义块组合符号 l41。于是,"正在准备"成为 E 要素 P1,句中各音节各得其所,一切模糊迎刃而解。"反击行动"将成为 PC 的核心部分,反击与行动构成偏正结构。这一初级句类分析过程在句类及其格式知识的引导下,难道不是顺理成章、畅通无阻么?

从句间关系来说,本句是对上一句'消息'的具体说明,这就是处理模块'隐知识揭示'的内容之一。

例句 5:包含两个同样格式的句子。对两句中的双音词'没有*'首先应根据'语境原则'暂时取消'煤油",而仅保留双字词'没有"。这是一个多义词,但其多义模糊由于它后面分别有双音词'提出*'和'取消*'而可以消除。这里有一个巧合,就是这两个双音词集合全是动词。于是,可以确定两句的 E 语义块分别是'没有提出*抗议'和'没有取消*"。这

两个 E 语义块都需要对象和内容,这一知识来于双字词‘抗议’和‘取消’的句类知识,它们分别属于关系反应句和作用句。根据这项知识分别对两句进行检验,找到后者的对象和内容分别是‘特使’和‘会谈’,都符合合理性要求,从而加强了层选处理对三音段 te shi de 和十音段中的 tian 与 he 的判断。前者的剩余部分‘为此 * ’只能充当内容,因此,这里需要确定被省略的对象及‘此’的指代,这也属于隐知识的揭示。

上述处理过程中还有双音词‘提出 * ’和‘取消 * ’的模糊问题。前者由于双字词‘抗议’的跟随而消除,对(提出,剔除)—抗议的映射符号

vg902　vg9382　vgc022

进行语义距离计算就可作出选定‘提出’的判断,是否进一步运用(提出,剔除)的语义结构方程知识已无关紧要。

双音词‘取消 * ’的解模糊过程比较复杂,但颇有示范价值,故详述如下。这里有关词汇“取消,取笑,特使,会谈”的映射符号分别是:

vg00 # v312　vg7111 + j862　j731/pa14　vgc249b

由于作用型概念 v900 以 pa 为优先对象、以 gc 为优先内容是复合基元概念集群‘9—a—c’的基本特性,作者曾在〈语义学日记选录〉中称之为‘第一号概念联想脉络’。所以,这里实际上并不需要查询概念关联性知识库,就能通过语义距离的粗估,得出‘取消’优先于“取笑”的判断。但是,这里还有一个更强有力的证据,它蕴涵在‘取笑’的映射符号 vg7111 的第三个‘1’中,它表示对他人、他事的态度,而这里‘会谈’的一方应该是省略的作用者自己,所以,“取笑”不满足合理性要求。至于‘会谈’的必有双方,是词汇和概念层面的确定性知识。而省略的作用者,则属于语法层面的基本知识。

例句 6:已如前述,不重复。

例句 7:由动词‘企图’与“取代”可以假设,这是混合句类的作用替代句。其句类格式是

A + XT4a + T4B2 + T4C

这里,

A:省略;　XT4a:企图 * 取代;　T4B2:ta;　T4C:成为 * sai zu 首领 * ;　Wy:通过战争。基于语句物理表示式的句类检验过程就是如此简单明了,当然这里有一个小麻烦,就是 E 块核心要素‘企图’与‘取代’出现了分离[2],中间插入了辅块 Wy,但由于‘企图’具有 vv 特征,这点麻烦是不难克服的。

例句 8:这是一个基本判断句。不过,判断对象 DB 又是一个语句。

例句 9:与例句 5 类似,包含两个句子。例句 5 用‘ye’作两句的连接标志。这里用两个“一面 * ”作连接标志。应用‘一面’作为连词的语法、语义知识,既消除了双音词‘一面 * ’的模糊,也消除了双音词‘密谈 * ’的模糊,确定前者是‘一面’而不是‘以免’,后者是‘密谈’而不是‘密探’。完成了语义块的初步切分之后,就可以确定第一句为关系反应句,第二句为信息交换句。此句随后的语句合理性分析包括下列四项内容:检验反应句和信息交换句各要素之间的协调性;验证层选处理对三音段 kai shi hou,xing shi shang,he mei guo 的预处理;验

证解模糊处理对各双音词的选定,找出各项省略的出处并验证其协调性。最后一项处理比较复杂,需要较强的句间处理功能,需要隐知识揭示的配合。

例句 10:与例句 5、9 类似,这又是一个"1 + 2"句式,标志信息是"ji"和"you"。表述"希望"的两方面内容。由于"希望"要求"2 + 2"句式,因此出现了前者对后者的嵌套,在形式上显得比较复杂。第一项希望是"保留 se zu 首领 * 地位 *",第二项希望是"联合国杰出 * 禁运",从而达到"挽救日益恶化的经济 *"之目的。

由"希望"这个概念构成的语句,是反应句中的复杂子类[15],例句 10 包含了这一复杂性的典型表现。【15】指出:反应句意愿子类中的 7121 子类,如"希望,祝愿"等概念,优先"2 + 2"句式。该文指出:"对 7121 构成的 2 + 2 句式,有一个特殊情况,就是对自身的希望,这时 XBCB 可以省略,这一知识目前尚未给出表示的手段"。这里正好碰上了这个麻烦情况,第一句就是对自己的希望。问题不仅在于这个麻烦本身,还在于它引发了另一个更大的麻烦,就是第二项希望的"自身化"。本来第二项希望的对象 XBCB 是联合国而不是自身,如果没有第一项希望,语句简明规范的表达形式是:"他希望联合国杰出 * 禁运",而现在由于第一项希望以及"ji"与"you"的介入,第二项希望的表达方式只好向第一项看齐,把它的全部构成作为"获得"的扩展内容来处理。"获得"的 B 要素仍是"他",这属于句类格式转换的课题,这个课题对于机器翻译极为重要,我们尚未进行研究,只好就此打住。

第二项希望之后还有一个作用句,如果在两句之间加一个介词 117"以",则两者的关系就明朗了,但这里缺省了"以"。这是与前面多次遇到的主语义块缺省性质不同的另一种缺省,是复合句类的一种特殊情况。

尽管例句 10 出现了上述一系列的复杂问题,但并不影响语义块的切分和句类的判断,因为与此有关的动词"希望、保留、获得和挽救"都是无模糊的。句式的判断,可依据三项相互加强的信息,一是单音词"ji"和"you"的知识,在音节感知库中有明确表示,二是两者后面跟随的"保留"与"neng 获得",三是四字词"一箭双雕"的语义。

最后应说明本例的一个有趣情况,就是它的解模糊处理表面上弱依赖于句类知识。双音词"杰出 *"的模糊集是"杰出,解除,接触,戒除",它们与"禁运"的语义距离计算比较简明,能以较高的置信度选出"解除"。同样,从"首领 *"的模糊集"首领,守灵","地位 *"的模糊集"地位,低微,敌伪,堤围",也不难得到"首领"与"地位"的语义距离最小的结果。前者在判断过程可以应用免除句以 04 类概念为优先内容之一的句类知识[15],但由于 03 与 04 概念的优先搭配在概念关联性知识库中有明确表示,实际上不必上升到句类的水平上来思考。后者在语义距离的计算过程需要排除"敌伪"对"首领"的排序干扰,这属于语法知识的运用。

但从上面的分析过程可以看出,如果运用反应句和效应句的句类知识,上述局部解模糊处理过程将上升到居高临下而不是坐井观天的境界。

上面以就事论事的方式,讨论了各例句在句类分析过程所遇到的各种问题和解决这些问题的途径。问题的中心是各种各样的模糊,而解决问题的基本立足点是句类的假设和句类知识的运用,具体操作涉及语义距离计算,以达到消解各种模糊的目的。同时,也要用到

若干语法知识,主要是 HNC 命名的语言逻辑概念知识。这里的叙述方式“只见树木,不见森林”,因为目的在于为下面的论述提供依据,或者说提供感性认识。

3.2 关于语义距离及其计算

“语义距离”这个概念试图对概念之间关联性的强弱给以定量表述。关联性的定量表述有“相关函数”或“相关系数”这样现成的术语,所以直接采用“概念相关函数”或“语义相关系数”之类的术语比较自然,“语义距离”这个术语的引入在理论上并无必要,不过是取其表述简明而已,实际上语义距离的计算就是计算概念之间的相关系数。

不同概念之间的关联性有明显的强弱之分,这是毫无疑义的。但如何量化和如何计算,则需要新的思路,不可能照搬信号处理中求相关函数的统计方法。统计方法的出发点是将待考察的系统视为“黑箱”,但语句不是“黑箱”,即使是语音识别系统的输出语音阵列,也只能说是一个“明暗相间、明为主导”的箱子,弃“明”而不用,显然是不明智的。

概念之间的关联性需要通过多重层面予以表达,有概念层面的关联性,有词汇层面的关联性,有语法层面的关联性,有语义块内部的关联性,有语义块之间的关联性。不同层面相关系数的量化和计算方法都应该有所不同。对这些不同侧面的辨识是进行语义距离计算的先决条件,以语句物理表示式为立足点的句类分析,是判断这些先决条件的强有力武器,在上一节针对十个例句作了示范性说明。这是语义距离计算的基本特点。

相关函数是一个条件概率,语义距离的条件性更为突出,在某种意义上,条件的把握是计算语义距离的关键。下面将对条件进行具体的说明,上一节对十个例句的分析都是侧重对条件的阐述,从中可以看到,句类知识是最基本的条件。

当然,在某种情况下,对条件可以弱化。语法学所概括的词性约束规则:即形容词与名词、副词与动词和形容词、数词与量词的搭配规则就是明显的例子。这些搭配实际上是有条件的,但作为语法规则来陈述,可以不管条件。

概念关联性或语义距离的概念,在某种意义上是对上述词性约束规则的扩展和深化。扩展表现在它力图表述语义块之间或语句要素之间的约束,深化表现在它力图尽可能给出条件。

语句要素之间的约束就是【1】中所阐述的链式关联,这一知识分别从概念层面和词汇层面进行表达。前者的表述是概念关联性知识库的内容[6],后者的表述是词语知识库的内容[7]。

条件则通过句类知识、交式关联和“同行优先”三条途径来表述,后两条实际上就是词性匹配的具体条件,第一条是运用链式关联知识的条件。

从上面的说明可知,语义距离的计算首先要区分语义块内部和语义块之间两种情况。

语义块内部语义距离的计算主要是运用“同行优先”准则、概念关联性知识库中“交式关联”知识[6]、语义结构方程所给出的语义块构成知识[7]。

语义块之间语义距离的计算主要是运用概念关联性知识库中的基本句类知识和概念节点的链式关联知识,语义结构方程所给出的搭配知识。

所谓"同行优先"准则,是对层次网络符号天然属性的一种简明陈述,正式的陈述是:同行的五元组概念及挂靠的(w,p)类概念优先相互搭配,在【1】中曾对此详加阐述。从应用的角度来看,这不过是用数字符号表达概念关联性的一个简单技巧。在具体应用这一准则于语义距离计算时,要区分四种不同的搭配方式,因为每种搭配方式各有自己的约束准则。四种搭配方式是:修饰型搭配,补充型搭配,并合型搭配,对象内容型搭配。前三种是语义块内部的搭配,第四种则表现为语义块之间的搭配。

下面就来对这四种搭配作较详细的说明。

——关于修饰型搭配

修饰型搭配大体上相应于语法学的上述词性约束规则,"同行优先"准则不过是对此规则的运用条件给以表述。"衷心的祝福"、"衷心的石头"、"衷心的消息"都是形容词与名词搭配,但后两者不合理。"衷心地希望……"、"衷心地诅咒……"都是副词与动词的搭配,但后者不合理。把这些词汇映射成层次网络符号,通过语义距离计算,计算机不难得到"衷心的祝福"、"衷心地希望"语义距离最小的结论,不难作出"衷心的诅咒"绝对不合理的结论。同时,也不难得到"衷心的石头"、"衷心的消息"不合理的结论。

对修饰型搭配的语义距离计算,就是将两概念的层次符号从高往低逐层匹配,"相同得分,相异不计",它类似于在极性重合相关处理时期对相关系数的简化计算。但计算前必须进行约束性检验,对修饰型搭配来说,需要作两项检验:一是词性及其顺序的检验,二是对偶性检验。

两概念组合的合理性或合法性,可从关联性和排斥性两个角度进行考察。排斥性可视为反关联,相应于相关函数的负值。但在语义距离计算时,仅取正值,负值一律视为相斥。从这个意义上说,约束性检验就是互斥或正负检验。不满足约束条件,就意味着互斥,表示两概念不能组合,这一结果对于解模糊或纠错处理最为实用。

词性约束是常规的语法知识,无庸赘述。需要说明的只是它的两条顺序约束:一,gu 类概念作为形容词使用时,在顺序上可前可后,但 ug 类概念优先于前;二,u 及 vu 类概念作为副词使用时,在顺序上可前可后,但 uv 及 uu 类概念优先于前。上面例句 10 中的"日益恶化"就是一个典型的同行修饰搭配,"日益"这个 uv 类概念和"恶化"这个 vg 类概念满足词性约束条件,其相关系数等于 1。

对偶性约束指对偶性概念的正负双方不能互相修饰,此理不言自明。"衷心"与"诅咒"虽然高层层次符号同行,但前者不能修饰后者,因为它们违背了对偶性约束。

"同行优先"准则有狭义与广义之分,即本行与交式关联行之分[1]。在本行里又有 0 分行和非 0 分行之分,这就不来细说。

广义"同行优先"准则的应用,目前就是将交式关联的级别指数[6]转换成相关系数,这时不是匹配层次符号,而是依据层次符号查询概念关联性知识库。这里的数值转换,类似于层

选处理时从音节感知库的独立性指数换算单音词的位置置信度[9]。

量词与表述对象的搭配也属于修饰型搭配,对这一搭配的语义距离计算可不作任何约束检验,而计算结果本身就是一种检验,因为两者必须狭义“同行”,相关系数应等于1。汉语量词之烦琐令人生畏,但由于现在赋予了“同行”特性,就理解处理来说,反而成了一笔意外的“财富”,可作为解模糊的一项手段。

——关于补充型搭配

补充型搭配有两种类型,一是高层概念与低层概念的搭配,二是泛指概念与特指概念的搭配。

第一类搭配又分两种情况,一是动词的高低搭配,二是名词的高低搭配。第一种情况仅出现在E语义块内部,是造成E语义块分离的原因之一[2]。这种高低搭配和分离现象不是概念表达的内在需要,而是语言表达多样性和艺术性的需要。在一般情况,语言的这一特性只会带来理解处理的困难,但高低层概念的搭配则相反,它带来的是机遇。原因在于相互匹配的高低层概念必须满足“同行”的条件。前节例句五中的“提出*抗议”就是高低层概念的“同行”搭配。如前文所述,对双音词“提出*”的解模糊处理就利用了这一信息。

名词的高低搭配是包含性概念的特性,这种搭配也满足狭义“同行”条件。

对高低搭配也需要进行顺序约束检验,顺序准则是:高层在前,低层在后。这一准则对动词似乎普遍适用,名词则不然,与语种有关,汉语遵循这一准则,而英语则相反。

泛指与特指的在许多情况也属于“同行”,这是由于对泛指和特指的人或物均采用挂靠表示方式,两者的层次符号一样,从而也能对两者进行语义距离的计算。由于这个计算非常简单,并不是一项负担,而应视为灵敏性反应的一种手段。这里不妨用一个例子来说明这一点。假定输入语音流中出现了yue fei,则从词库中将找出“岳飞”这个词,如果该文本实际指的是原苏联物理学家“约飞”,计算机能觉察“岳飞”是一个伪词么?回答是肯定的。“岳飞”的层次符号是pa4,而该文表述的内容应主要涉及a6。线索就在这里,语义距离的计算本身非常简单,但关键在于要运用专业活动的句类知识:专业活动aj的A要素优先于从事该项专业的人paj。在一般情况这类判断需要很多的常识性知识,但这里是不是“岳飞”的判断,似乎可以绕过常识,仅从层次符号就能得到。当然,这样“绕过”的适用范围也许非常有限,但终究是有胜于无吧。

对泛指与特指的语义距离计算,可暂不作顺序约束检验。

汉语里数词与量词的搭配属于广义“同行”补充型搭配。顺序约束条件是:数词在前,量词在后。但汉语的数词并非一定要与量词搭配,成语里的“五湖四海”“三令五申”“百孔千疮”“百炼成钢”都省略了量词,其中的数词都是虚用,表示“多”或“全”的意思。现代汉语的“五讲四美”“十大新闻”“三好学生”也省略了量词。关于数词的运用,需要建立一个专用的小知识库,特别是“一”字的语义语用知识。

——关于并合型搭配

并合型搭配之间通常加逻辑指示符,这样的指示符有四类,现将它们和相应的汉语和英

语符号列表如下(表中顺便给出了“的”的另一义项):

符号	汉字	英语	意义
l41	的 de		偏正
l41461	的 de	's	偏正
h□ug	的 de		词性转换
l42	得 de	of	反偏正
l43	和同与及并跟	and	逻辑并
l44	或 hu	or	逻辑选

前两种并合将称为“修饰”并合,后三种并合将称为“逻辑”并合。

修饰并合与前述的修饰搭配不同,两者的差异在于“同行”性的有无,修饰搭配具有“同行”性,修饰并合不具有。英语不仅对这两种组合方式在表达形式上给予了明确区分,对修饰并合的三种类型也加以区分,汉语则一律不加区分。仅用符号“的”表示它们的共性,而模糊它们的个性。对前两种修饰并合,曾有过用“的”和“底”加以区分的建议,但未得到广泛响应,说明这一模糊并不影响人的理解。

从理解处理来看,对修饰并合和修饰搭配的语义距离计算,都需要进行对偶性检验和词性检验,虽然词性检验的内容略有不同,但并不影响语义距离的计算。因此,汉语在这里的模糊表示似乎无损于理解处理,其实不然。问题在于两种情况的合理性阈值差异甚大,修饰搭配的阈值很高,而修饰并合的阈值很低,人在理解过程中能自动调节这一阈值,计算机很难做到这一点,因为这不仅涉及概念和词汇层面的知识,还涉及常识性知识。但是,理解处理的途径是阳关道与独木桥并存,解模糊处理更是如此,此路不通,可置之不理而另觅它径。在前一节的十个例句中,有五处以 de 为标志的修饰组合,而且都是修饰并合,但需要利用并合前后概念关联性知识的只有第七句,这一句又恰好具有足够的关联性。当然,十个例句不能代替统计,汉语的这一模糊对理解处理造成的不利影响需要利用语料库作深入的研究。

对于逻辑并合,需要进行类别符号的对仗性检验,即检验并合前后两概念的类别符号是否相同或相当。“相当”是模糊的说法,有待给出具体的规则,这是不难做到的。这一规则的制定也有赖于语料库的建设。汉语常省略逻辑并合标志,这一省略与修饰搭配符号的省略将模糊两类组合,由于这两类组合的约束准则不同,将影响到语义距离的计算,因此,必须先消除组合模糊,这确实是汉语理解处理的一项额外负担。

但是,像上述两类修饰模糊一样,对这一负担应采取灵活反应策略,因为许多情况可以置之不理。而在无此模糊时,从对仗性检验及语义距离计算结果常能取得消除模糊的关键性信息。

——关于对象内容型搭配

前面已经说明,对象内容型搭配是语义块之间的搭配。具体的说,就是 E 要素与 B 要素或 C 要素的搭配。这种搭配,一般说来,并不具有“同行”特性。但应该指出,上述修饰型和补充型搭配的天然“同行”特性乃来于概念层次网络符号的知识表示方式,没有这种表示

方式,也就无所谓‘同行’。层次网络符号由于在五元组中引入了 r 类概念,并对具体概念采用了以挂靠为主的表示方式,使得对象内容型搭配大大增加了‘同行’的机会。

对象或内容‘同行’显然是一个非常宝贵的信息,这一信息分别在概念和词汇层面予以明确的表达。前者用概念关联性知识库 B、C 栏目的第一项表示[6],后者用语义结构方程“1—3”规则的 k = 7 表示[7]。

当然,对象和内容属于‘同行’的情况,即使我们着意作了尽可能多的安排,仍然只是少数。对居于多数的不‘同行’情况,采取三条途径提供关联信息。一是概念关联性知识库中的 A、B、C、M、Pr、Rt 栏目,二是词义表示中给出的关于对象和内容的层次符号,三是语义结构方程的‘1—3’规则。这些知识的表示方法在【6】和【7】中有详细说明。

上列第一项知识实际上就是概念相关系数的一种表示方式,已如上述。第二和第三项知识如何转换成相关系数,尚有待深入研究。初期可采用 1 比特量化的简单方式:满足预定条件,即取相关系数等于 1,否则等于 0。

以上所述,都是两两相关。而一个句子,甚至一个语义块有多个两两相关,把它们综合起来,对整个语义块或句子作总体评价是初级和中级句类分析的任务,这将在下一节讨论。

3.3 句类分析的初级处理

在【2】中对句类分析作了初步论述,提出了三级处理的划分。在上述语言理解处理框架的四个层次中,第二到第四层次的核心内容实际上就是句类分析的三级处理。上面对十个例句理解处理的逐一说明,是以句类分析为主线。本节将对句类分析的初级处理从理论方面作进一步的阐述。中级句类分析在下一节讨论。

1. 判定句子的句类、句式和格式

这里应再次说明一下,本论文集中的术语‘句类,句式’的意义与传统定义不同。这里的句类是论文【2】中所定义的基本句类和混合句类以及【14】到【21】中所阐述的各个子类。这里的句式是【2】中所定义的‘1 + 1 " 1 + 2 " 2 + 1’和‘2 + 2’四种句式。格式是【2】中所定义的语句物理表示式的标准格式及其各种非标准句式。

2. 判定各语义块的个数及其构成

语义块的构成分为简单、复合、扩展及句蜕四类。简单构成只有核心部分,即只有语句要素;复合构成有核心和说明两部分,但说明部分可以是另一个语句,这是句蜕的一种形式;扩展构成指该语义块本身又是一个语句,扩展通常是含有 C 要素语义块的特征;句蜕是指一个语义块由另一个语句蜕化而来[2],通常是将一个语句的某一语义块或块素变成句蜕块的要素部分,同时将原语句的其他语义块变成句蜕块的说明部分。

3. 判定缺省及指代

这里的缺省包括语义块的整体缺省和语义块构成的缺省。

这三项处理内容概括了初级句类分析的广度,但对分析的深度则未作硬性规定而保留

一定的灵活性,这主要基于以下两方面的考虑。

第一是在分析的深度方面,初级与中级处理的界限不像广度方面那么清晰,实际上很难对此作出硬性规定。例如作用句 B 语义块的构成,可以复杂到作用对象、效应对象和效应内容三者齐备,但也可以简单到只有由一个词汇表达的对象或内容。这时,对象的构成可能与 E 要素语义结构方程给出的要求不符或矛盾,如‘提高水平’‘保证质量’‘加快进度’‘搞好关系’这类的词汇组合就产生了这种矛盾现象。从信息的及时利用来说,当然最好是立即加以处理。但对多数情况,这是不必要的时间浪费。这里就出现了处理时机和信息保留的复杂问题。

第二是知识库的建设未能先行。全方位满足初级及中级句类分析需要的知识库建设是一项复杂而浩大的系统工程,目前仅有一个实验性或示范性的雏形。由于经费的拮据,不可能在近期内取得满足理解处理全面需要的进展。理解处理软件的研制过程必须适应这一非技术因素造成的巨大困难。在初级句类分析的深度方面,先采取‘见机行事’的策略带有迫不得已的性质,然而是必要的。

下面就来对初级句类分析的三项或三步处理逐一加以说明。每一项说明由两部分构成,一是信息来源,二是运用这些信息时可能遇到的各种问题。信息有静态与动态之分,静态信息由知识库提供,动态信息由上述框架的两项预处理——分段层选及语义块切分组合提供。

——第一步:句类及其格式

句类信息有三个来源,一是 E 要素的层次网络符号,二是 E 要素的语义结构方程,三是主语义块的逻辑指示符。前两项 E 要素信息,在性质上可合称内部信息,而第三项信息在性质上是间接的旁证,可称外部信息。

内外信息可同时出现,可能有内无外,也可能内外俱无,这都是正常情况。此外,还可能出现有外无内的不正常情况。这四种不同情况将引发不同的后续处理。

当内外信息同时出现时,必须进行一致性检验。两种信息必须一致,利用这一特性,可消除 E 要素或指示符的可能模糊,这一解模糊手段当然要加以利用。如果出现不一致的情况,则与对待‘有外无内’一样,按反常情况处理。这时,初期理解程序应请求人工干预,而不要推给中级句类分析去处理,因为这可能是知识库错误造成的反常结果。

内外信息同时出现一定相应于非标准格式的语句,反之,有内无外则相应于标准格式,这一结果应记录在案。

内外俱无则一律按特殊情况处理,它可能是不含 E 要素的状态句,也可能仅仅是一个广义对象语义块。

上列各种处理途径,除了请求人工干预外,最后都汇聚到第一步处理的主要事项:确定句类。

这时,按内部信息的是否模糊和是否充分,有四种不同的情况,但初期处理程序并不设置四种走向,仅作简单的登记处理即转向第二步。所谓简单的登记处理,就是在无模糊时登

记语句的类别,有模糊时登记可能的类别,在信息充分时给出充分性标记,不充分时给出不充分性标记。

信息不充分是指句类的子类不能明确判定。这有两个原因,一是语句因继承上文而省略了分离结构 E 要素的低层表示,二是词知识库的表示不完善。

E 要素模糊是经常遇到的情况。在初级句类分析之前虽然进行了语义块切分组合处理[11],但上列例句中以符号“ * ”标志的 E 要素模糊的大部分并未消除,如第五句的“ 提出 * ”和“ 取消 * ”,第七句的“ 成为 * ”,第八句的“ 不是 * ”。这是第一类模糊。第二类模糊是词或音的多义模糊,如第一句的“ 不在 ”和第二句的 zai。这两项解模糊处理不仅是第一步处理的中心内容,也是整个初级句类分析的中心内容。显然这一句类模糊越早消除,越有利于后续处理。但反过来说,后续处理的内容越明确,越有利于句类模糊的消除。这种相互依赖性使得处理顺序最好是见机行事,但这样的“ 见机行事 ”将使程序陷于无所适从,因此,宁可采用“ 先易后难 ”的低效然而稳妥的步骤,而放弃“ 擒贼先擒王 ”的高效策略,先转入第二步。

——第二步 :语义块的个数及其构成

根据 lv 准则切分出来的语义块[11],会产生误切,个数通常偏多,当然也有偏少的情况。切分错误只有在确定句类的前提下,才有可能得到纠正。因为句类规定了主语义块的个数。

第二步处理的中心内容,实际上是对预切的语义块作一次“ 点名 ”处理。“ 点名 ”处理的前提是已知或假定句类及句式,根据两者提供的名册,核对预切的语义块,所谓“ 核对 ”,就是进行“ 多求扩展,少求缺省 ”的处理。意思是 :如果预切主语义块数量多于名册,按 C、B、A 顺序进行语义块扩展的判断。如果预切语义块数量少于名册,则先考虑缺省。例如前一节例句 8,预切结果是五主一辅,其中有两个 E 要素,一个是作用句的 X“ 削弱 ”,另一个是基本判断句的 jD“ 不是 * ”。前者无模糊,因而关于作用句的前提是完全确定的,后者有模糊,因而关于基本判断句的前提是假定。在此前提下进行“ 点名 ”处理,发现 :作用句的 A、B 俱全,但基本判断句缺 DB,由此推知,该作用句是基本判断句的扩展 DB,整个语句是一个作用判断句。

在上述“ 点名 ”处理过程中,作用句和基本判断句的句类知识是关键因素。

“ 点名 ”处理不仅施于语义块,也施于复合语义块的构成。这里的复合是广义的,包括扩展语义块和句蜕。

复合语义块的构成信息来于两方面,一是语义结构方程[7]的“ 1—2 ”规则。二是语义块核心的特性,该特性蕴涵在该核心概念的层次网络符号里,如前一节例句 5 中的“ 会谈 ”和例句 9 中的“ 密谈 * ”。

复合语义块构成的“ 点名 ”处理远比语义块“ 点名 ”处理复杂,这一点下文就要谈到。

——第三步 :缺省及替代的指定

这是“ 点名 ”处理必然的后续步骤。“ 点名 ”过程自然就会发现缺省,也找到了指代符号。

对于缺省和替代,语言学和计算语言学都已做了大量研究工作。西语丰富的语法信息对此大有裨益。汉语相对贫乏的语法标志会带来一些不便,但对以句类分析为纲的语义层

面处理影响不大。

这里说的缺省包括语义块的缺省和语义块构成的缺省。

主语义块的数量由句类完全确定,它的缺省通过“点名”不难发现。

语义块构成的缺省则不然,情况要复杂得多。语义块的构成本身是不确定的,E、A、B、C四种主语义块的构成方式差异甚大。所以,“语义块构成缺省”的提法本身并不严谨,但如果将缺省限定在B、C两种语义块,则大体上有章可循。这个“章”就是语义结构方程所给出的B、C构成信息和语义块物理表示式中蕴涵的结构信息。这两种信息在性质上有所不同。从理论上说,前者所给出的信息是确定的,是否缺省可据此敲定,而后者所蕴涵的信息具有两可性。但对一个具体语句来说,确定的缺省可能无关紧要而可以置之不理,而两可缺省反而必须加以澄清。对这种复杂情况的应变处理,初级句类分析不可能胜任,也不应该提出这种过高的要求。但是,对缺省的情况必须一律详细登记,以备后用。

语义块的构成缺省具有很大“随意”性。这里说的“随意”性实际上是指下列诸因素的影响:口语或书面语;语句的体式;全局语境或局部语境;表达的艺术性需要;说话人或书写人的风格及意向。这就是说,缺省本身含有丰富的信息,句类知识和语义结构方程在很大程度上就是为缺省的发现和处理提供依据。

缺省的理解处理实质上属于隐知识揭示处理的内容,是HNC理解处理的中期目标。

与缺省相对应,语言中存在丰富多彩的冗余和重复现象,口语中更为突出。理论上,层次网络符号应能提供明确的冗余和重复信息,但对这些信息如何处理则尚未作具体研究,这里只能略而不谈了。

从上面的叙述可以看出,初级句类分析的任务主要是就上列三项内容逐项在一张表格上登记。这张表格的设计实际上就是软件的方案设计,但已超出本文的预定范畴了。

3.4 语句合理性分析及回溯处理

上一节已经说明,初级句类分析的结果是针对每一种句类假设生成一张关于语句深层结构的表格。语句合理性分析就是对这张表格的有关项目进行合理性或协调性检验。如果检验失败了,就依次对下一个句类假设进行检验。如果初级句类分析的全部句类假设都宣告失败,表明语义块感知处理出现了重大失误。这时,就需要进行回溯处理,从语义块感知开始另寻处理途径,实际上不过是动用原来备存的一些置信度较低的信息。

上面说到的回溯处理显然是全局性的,它对初级句类分析的结果全盘否定。这种情况比较少见。回溯处理也可能再次失败,这时,就只能请求人机交互或人工干预。除了全局回溯以外,多数情况是局部回溯,它包括语义块切分点的调整,不符合语境原则的伪词的辨识和拆分,连见动词中E要素的再确定,不带E要素前后标记的单音词E块的发现和确认等。

基于上述,衡量HNC理解处理发展水平的标准不能只考查音词转换正确率,这只是一

个无实质性内涵的统计标准。机器翻译和语音识别多年来采用类似的标准,对两学科的发展并没有起到实质性的推动作用。

为了推动 HNC 理解处理,应该建立符合语句合理性要求的特定衡量标准,这个标准应包括下列五项内容:

1. 句类辨识玄度
2. 语义块辨识玄度
3. 全局回溯玄度
4. 解模糊玄度
5. 交互玄度

这里不得不引入一个新词"玄度"。"玄"具有"知其然"和"知其所以然"的双重含义,它取自《老子》第一章,"此两者(指哲学意义的有无),同出而异名,同谓之玄。玄之又玄,众妙之门。"可见,它包含康德所刻意区分的知性(understanding)和理性(reason)的意义。按五元组的术语,玄度就是"知其然",并"知其所以然"的值。与通常意义下"正确率、虚警、漏报"等概念相比,它不仅表明了统计结果,而且表明产生这个结果的条件和原因。玄度是多维变量。玄度表示函数将作为一个专题另文讨论。

上列玄度标准规定了语句合理性分析的下列基本内容:

1. 能否对句类作出正确判断。包括复杂情况下的两可判断,对显而易见的错误句类假设作出迅速决断等。

2. 能否对各种语义块作出正确判断。包括:(1)对特定句类特定语义块的特定要求能否作出灵敏反应;(2)对语句标准格式和非标准格式能否作出人类语言感知水平的相当判断;(3)对常见的语义块省略特别是承上省略能否作出灵敏反应;(4)对扩展为语句的语义块能否有效辨识;(5)对由语句蜕化而来的语义块能否有效辨识。

3. 对语句中的每一个词和每一个音节都能否从概念关联性的要求或语法要求使之各得其所,不允许出现前后无关联性的孤立词或孤立音节。

4. 对两逗号之间的语音串或文字串能否作出是句子或非句子的判断。

5. 能否依据"语境原则",而不仅仅是依据词频统计知识对双音词模糊集的抉择作出灵敏反应。

合理性分析的依托是上一节初级句类分析所给出的关于语句深层结构的综合表格,基本判据就是本文第 2 节所反复说明的句类知识运用;同行优先准则的运用;最后是语境原则的运用。

3.5 短时记忆和语境生成(兼【1】【3】小结)

HNC 理论预定建立以下五个层面的自然语言理论模式:

1. 自然语言概念体系的理论模式；
2. 自然语言语义块和语句的理论模式；
3. 句群关联性及篇章要点的表述模式；
4. 短期记忆和长期记忆的形成及其相互转换模式；
5. 基于文字文本的知识学习模式。

没有这五个理论模式的建立，自然语言理解这一人工智能的分支学科不可能摆脱当前的低水平状态，长此以往将对信息时代空前膨胀的语言信息财富陷入“望洋兴叹”的困境。虽然信息的生成、传送和接收，当前显得热闹非凡，但应该清醒地看到，语言信息财富的利用水平仍然极不协调地滞留在相当于物质财富利用的农业时代。改变这一状态的唯一出路是自然语言理解的突破，HNC 理论试图为这一宏伟目标尽一份力量。

【1】和【2】是前两个理论模式的概要说明，但后三个理论的模式探索则有待于前两个理论模式的技术实现。本文试图对此有所推动，但这一历史的重任只能依靠年青一代来承担。

短期记忆与语境生成处理模块密切依赖于后三个理论模式的建立，但这不等于说，当前就只能束手等待。

短期记忆和语境生成都可以从简单方式做起。在第 2 节中已给出了一些示范性说明。

词知识库中未登录的人名、地名和物名是短期记忆模块应优先关注的对象。这里的人和物都是广义的，包括 HNC 符号体系所定义的 pe、gw、rw 和 vC 类概念。它们在现代传媒信息中占有过于特殊的地位，不能不运用短期记忆模块予以处理。通过句类知识的运用，这是不难做到的，当前可仅作为一项技术设计问题来处理。

语境生成似乎茫无头绪，其实不然。复合基元概念局域网络的设计已为此作了充分准备，不难据此列出一张语境类别的清单。所谓语境生成，就是对这一清单的具体认定。这可以通过对要素词语 HNC 映射符号的简单统计作出判断。

关于复合基元的设计在【1】中未详细阐述，但在它的续篇【6】中，从语境生成角度作了较详细的说明。

1995 年冬

后　记

近三年前写的这篇论文，大体上形成了 HNC 理解处理技术实际发展的基本思路。这里应说明两点：第一，文中关于初级、中级和高级句类分析之说已废而不用，改为语义块感知—句类假设与句类分析—语义块构成处理三部曲的提法，并将这个三部曲所体现的处理策略概括为“中间切入，先上后下”的八字诀。其次，文中多次提到的语义结构方程也已弃而不

用,它所提供的知识改由汉语 HNC 语言知识库给出更完备的表述。因此,原来以“汉语非单音词知识库及语义结构方程”命名的【7】已用 1997 年秋写的“关于汉语 HNC 知识库的建设”一文来代替。

由于上述两项大的变动,本文中的一些提法和用词已不够确切。但为了保持历史原貌,一律未加改动。

1998 年 8 月 18 日

概念知识和语言知识

引　　言

人工智能早期一系列的挫折,使人们认识到知识的重要性。要使计算机表现出智能,唯一的办法就是使它拥有并运用知识。正是这一认识促成了 20 世纪 70 到 80 年代的“专家系统热”,并取得了引人注目的成就。但这些专家系统的知识,都是局限于特定的领域,而一般自然语言理解(这里不包括特定领域的简单语言应用系统)所需要的知识则完全不同于通常的专家系统。它需要各种各样的知识,但可分为三大类:概念知识、语言知识、常识及专业知识。

前两类知识库的本质区别在于:语言知识的内容与具体语种有关,而概念知识与语种无关。把概念知识从语言知识中独立出来,从概念层次网络理论看来,是势在必然的发展。

本文仅具体讨论概念知识,两种最重要的语言知识另在【7】【8】中讨论。所以,本文题目似乎有点名不副实。但这两类知识不可能截然分开,混合形态必然存在,而对后者的说明是本文的内容之一。另外,本文还会谈及三类知识库的一般特征。这样,本文的题目大体上可以通过了。

由于本文以概念知识的阐述为主,所以,它应视为文【1】的续篇。文【1】对复合基元概念略而不谈或言而未尽的部分将在本文的 6.2.3 中作系统说明。

6.1　语言知识与概念知识

语言知识如何表达?这个问题到今天,也就是写这篇小文的时候,依然不能给出完善的答案。也许可以说,这个答案完善之日,即自然语言理解接近大功告成之时。而这个时日,似乎不是一代人的努力可以达到的。但本文将预示一个重要的进展:语言知识的表达将从以语法知识为主导的时期,全方位地转移到以语义知识为主导的新时期。这一转移的得以实现,从理论上说,是由于概念层次网络理论的建立。从知识库的角度来说,则有赖于把概念知识从语言知识中独立出来的举措。

在中国传统语言学的术语里,语言文字学及其有关学科统称训诂学。黄侃先生对训诂的定义是:“训诂者,用语言解释语言之谓”。黄先生在这里说的是语言知识的传统表示方法,词典就是用的这个方法。这个方法适用于人,但不适用于计算机。一阶谓词逻辑、框架理论、语义网络、概念从属理论、功能合一语法等都是对语言知识传统表示方法的突破,使之

更适合于计算机使用。其中,概念从属理论第一个深刻认识到语言知识和概念知识的区别,但还没有明确产生在语言知识库之上建立概念知识库、在语言知识库之下建立常识及专业知识库的思路。

语言是概念的外壳,思维过程表观上是对语言的操作,本质上是对概念的操作,特定情况甚至可以不利用语言这个外壳。不过在操作过程中通常是把概念依附在思考者最习惯的一种语言外壳之上。语言知识有其语种个性和语用个性的侧面,概念知识应尽可能排除这些个性。为什么要用‘尽可能’这个修饰词?因为这些个性的范定并非易事。例如,西语动词的形态依赖于主语的人称和数,汉语对这一语法规则置之不理,这是语言知识中语种个性的典型表现。西语的这一语法知识实际上是冗余知识,因为这项知识已包含在主语的词汇里了,所以汉语才能置之不理。但动词形态依赖于时态这一语法知识则不同,时态知识在西语有关词汇里原来并不存在,所以西语用形态变化的手段予以表示,汉语虽然没有这一手段,但它通过其他方式达到时态表达的同样效果。时态知识本身并非冗余,这与人称和数的形态表示不同。由此可见,对语法知识语种个性的分析,不能仅着眼于形式,也要着眼于内容。不过,总的说来,对语法知识语种个性的把握还不算十分困难。更困难的是语用个性的表达,而语用个性更是同语种个性交织在一起的。

把复杂的问题作类别性和层次性的双重分解,化为一连串比较简单的问题,这是解决复杂问题的一般途径和方法。对语言知识的表达也应遵循这一法则。排除语种个性和语用个性以后的语言知识,显然是比较干净和比较单纯的知识,它的表达自然也就比较简单。这个设想似乎迹近疯狂,但现在看来,这是语言知识表达的必由之路。具体的实现方案就是用概念层次网络符号去近似表达语言概念,并将自然语言理解所需要的知识分成三个层次:概念层次,语言层次和常识层次,并分别建立这三个层次的知识库。

概念知识库里的知识是与语种无关的最高层次的知识。其基本内容是;

1. 概念类别知识,主要是三个超级语义网络的节点配置知识。

2. 句类知识,包括格式知识、句式知识和语义块构成知识。

3. 概念关联性知识,主要是概念节点之间的关联性知识。包括交式关联和链式关联,后者表现为对语句要素之间概念搭配的约束。

语言知识库的基本内容,按传统分类法,应有语法、语义、语用三个层面的知识库,但这三类知识是交织在一起的,各自独立建库不符合经济原则。这三类知识以语义为天然核心,语法和语用知识实际上都是依附在语义知识之上,因此。语言知识库的主体应该是语义知识,表达对象主要是词汇,相应的知识库也可称为词义库。对汉语,还必须加上字知识库和音节感知库。

在词知识库中,词汇的意义用层次网络符号体系近似表示。这里的‘意义’不仅包含语义知识,也包含语法和部分语用知识。语义知识用层次网络符号表示,即将词汇的各义项映射成层次网络符号,因此词义库有时也称映射库。语法知识用类别符号、层次符号和独立性指数等多种形式来表示。实际上,l 网络和 f 网络所表达的知识都可视为语法知识。部分

语用知识则用结构方程来表示。结构方程是层次网络符号的重要补充,两者一起构成层次网络符号体系。结构方程提供语义块(主要是 B 语义块)构成知识、语句要素的类别优先性或概念优先性知识以及句式知识。当然,结构方程所提供的信息,直接用层次网络符号也能表达,在字知识库中就是这么做的。从这个意义上讲,结构方程可视为句类知识和语用知识的一种简化表示方法。

以上所述,在专门讨论语言知识库的两文【7】【8】中有详细说明。两文虽然都冠有“汉语”二字,但它们所描述的两类汉语知识库的数据结构和知识表示方式是普适的,对任何语种都适用。

上述词知识库中所包含的语法和语用知识,其语种个性的表现是局部性的。语种个性的全局性表现则应另行建库。就汉语来说,是音节感知库,这在文【8】中专门讨论。就西语来说,是语法规则库。如上述动词随人称和数的形态变化,就是语法规则库的内容。汉语是否需要另建语法规则库是一个十分敏感的问题,这里不来讨论。关键不在库的形式,而在它的内容。

常识及专业知识可定义为概念知识和语言知识的补,也就是说,它包括这两类知识以外的所有知识。计算机的数据库实际上就是一种常识或专业知识库。

这三类知识对自然语言的理解处理都不可或缺,但从模糊消解这一当前的首要目标来说,常识性知识居于次要地位,重要的是概念知识和语言知识。

6.2 自然语言概念符号体系的补充说明

本节是文【1】的续篇。该文已对自然语言概念符号体系作了总体性的概要说明。本节先说明类别符号的连用和节点配置所体现的综合性知识,随后对复合基元概念的各种局域网络作系统说明,揭示它们的语境特征,并给出相应的概念节点表。最后,补充说明物表示的特殊约定。

6.2.1 类别符号的连用知识

概念层次网络理论将概念分为抽象和具体两大类。前者又分为基元、基本、语言逻辑三类,用类别符号 ϕ,j,l 表示。后者又分为物和人两类,用类别符号 w 和 p 表示。那么,抽象和具体这两个概念如何表达?这就属于类别符号的综合知识,层次网络理论并未对此给出精确表达的手段,只是把它们映射成:

抽象 (ϕ;j;l)/vgu 具体 (w;p)/gu

这个符号里蕴涵的综合知识很有点“可意会而不可言传”的味道。一般来说,映射符号只是语言概念的近似。我们在所有的论文里都是把“映射”作为“精确表示”的近似来使用的。但这里不是近似程度的问题,而是对综合知识的表达似乎存在手段不完备的缺陷,这一缺陷如何弥补尚有待研究。

类别符号连用知识的表达似乎比较简单。如果假定类别符号的连写仅代表“并或”一种组合形式,而且对“并”与“或”不加区分,那就几乎不存在连用知识的问题了。但实际的连用远非如此简单。首先连写的顺序问题就很值得推敲。这个顺序必须包含附加信息,但需要附加的信息很多,存在一系列不易处理的矛盾。

下面就来分别介绍一下六种连用的有关约定。

(1) 五元组的自身连用

五元组是抽象概念多元性表现的基元。在五元组的设计里,形容词和副词的概念是通过组合来体现的,gu 和 ug 都含有形容词的词性,uu、uv 和 vu 都含有副词的词性。gu 与 ug 的区别是:前者兼有名词和形容词的词性;后者是“纯粹”的形容词,相当于语法学定义的“区别词”。没有词性兼类或“形容词为主、名词为辅”的意思,这是约定,符号本身只是暗示而不能明确规定这一点。uu 表示一般副词,uv 表示专用于修饰动词的副词,vu 表示兼有动词和修饰词(可副可形容)的词性,这也是约定。兼词性汉语极为常见,单纯一种词性的词汇是少数,西语反是。vu 类词汇更可认为是汉语的“特产”,如“间断、健全、讲究”等等。

五元组符号自身的连用,除 uu 之外,还有 zz 和 vv。前者是量词的定义,后者则表示该动词具有与另一动词连用的特征,两者联合构成 E 要素,该 E 的句类由后面的动词决定。vv 类动词本身也可以连用,它们后面必然还有另一动词,从而构成三动词连用的复合 E 要素,这都是汉语的特殊语法现象。

五元组符号可以多级连用,这时只表示词性的兼类。

五元组也可视为“词性”的基元,它们可以组装,这就为“词性”的表达提供了以不变应万变的强有力手段。

(2) 抽象概念类别符号的自身连用

这里只考虑抽象概念的基元、基本和语言逻辑三大类,它们的连用情况远比五元组简单。抽象概念的自然顺序是:j,ϕ,l。有三种非自然顺序的连用:ϕj,lj,lϕ。前者仅限于基元概念状态 500 对基本概念的挂靠。后两者就是所谓逻辑概念对基元或基本概念的挂靠。这里的所谓挂靠,其实就是两类概念的并,说穿了,不过是省掉一对括号、一个逗号加若干五元组符号而已。当然,在概念的层次级别上,基本概念俨然“高高在上”,基元概念略低一等,而语言逻辑概念显然具有对两者的依附性。所以,挂靠的说法在这里也有其理论依据。

抽象概念类别符号的自然顺序连用目前只有 jl。它是“基本逻辑概念”的定义,与组合结构的“并”毫不相干,但可视为偏正的“近似”:jl≈j/l。是概念节点的一种特殊配置。

(3) 具体概念类别符号的自身连用

具体概念仅两个类别符号,只有一种组合,两种排列。目前仅用了 pw 连用形式,它是“人造物”的定义,可视为组合结构“效应”的近似:pw v6500□w,是概念节点的又一种特殊配置。

(4) 抽象概念类别符号与五元组的连用

五元组本来就是抽象概念的多元性表现,所以,抽象概念类别符号与五元组的连用是天

然的‘搭配’,先后顺序更应顺乎自然,不容颠倒。但具体表达仍引入了下列约定:第一,基本概念以静态为主,故将 jg 简写为 j,但是,如果是含 g 的连用,g 不省略;第二,对仅起语义块指示作用的语言逻辑概念仅用类别符号 l,一律不带五元组符号,但这不是省去 g,应理解为特殊定义。再说得细一点,不应该出现符号 jg,如果出现,可当做 j 处理。但可以出现 lg,而且不能当做语言逻辑指示符使用。

(5) 具体概念类别符号与抽象概念类别符号的连用

总共有六种连用形式:wφ,wj,ws;pφ,pj,ps。

这六种连用才是真正的挂靠,具体概念挂靠于抽象概念。前者必须挂靠,因为它没有自身的层次符号定义。

这里不能不简单回顾一下设计层次网络符号时对这个问题的思考过程:概念的抽象性和具体性显然不可能截然分开,概念本身就是抽象的,具体概念之说在逻辑上就不够严谨。另一方面,具体的概念必然涉及多方面的细节,很难‘净化’。所以,语义网络的层次符号按抽象概念、而不按具体概念设计;这就是说,w 和 p 无自己的层次符号,因此,对它们的精确表达只能借用抽象概念的层次符号。这就是挂靠一词的起源,或者说,采取挂靠的表达方式乃理所当然的步骤。

抽象概念类别符号之后一般跟五元组符号。具体概念是否有此必要?或者说,具体概念是否具有五元组的多元性表现?答案应该是肯定的,因为物和人也有动态与静态之分,也有属性、值和效应。但问题在于具体概念的严格动静之分很难把握,不加区分的模糊表示似乎更为妥当。将行人、老人和死人分别映射为

pv22b pg10bc55 pg146

未必更为精确,将五元组符号去掉也许更为简明,因为老人和死人也都处于变化的过程,而行人也需要休息而暂时转入静态。因此,这类连用大多数情况直接跟层次符号。意思是:由该节点定义所描述的人或物。一组层次符号只能描述一个方面。如果需要从多方面描述,就采取多重挂靠,用多组层次符号予以描述。各组之间用展开符号‘+’[1]连接。

上列六种连用可理解为相应类别的物和人。这里应该说明的是,w 和 p 的挂靠层未引入语言逻辑概念 l 和语法概念 f,因为,这两类概念无助于对 w 和 p 的特性说明。

以上所述,是对具体概念的一般表示方式。但具体语言概念的表达不能全部采用挂靠方式。把空气、阳光和水这些生命赖以生存的基本物作为基本定义来处理,显然有助于概念的联想,这就产生了以 jw 为类别符号的一类概念。与 jl 类似,它可视为偏正组合结构的近似,即 jw j/w。这又是一种概念节点的特殊配置。

到此为止,已经引出了三种特殊配置的概念节点:jl,jw 和 pw。前两种近似于偏正组合结构,是两个根节点,有自己的层次符号。pw 不同,它相应于效应组合结构,没有自己独立的层次符号,对它的精确描述依赖于挂靠的层次符号。

(6) 五元组与具体概念类别符号的连用

目前仅引入了 gw、rw 和 rvw 三种配置。分别定义为‘信息产品’、‘静态效应物’和‘动态

效应物”。例如,“雨风流”用 rvw 表示,而“霜冰虹”则用 rw 表示。

(7)物性符号 x 的引入

它是五元组符号 u 的物化。通过它把具体的物性与抽象的属性区别开来。例如,温度、色彩、体积、重量这一类的具体物性就分别映射成 jx00、jx10、xj20 - 、xj518,分别与“热”jw00、“光”jw10、“体”j20 - 、物理学的“质量”j518 相对应。不言而喻,面积和长度的映射符号是对“面”和“线”的挂靠,其映射符号分别是 xj20 - 0 和 xj20 - 00。但 x 是一个兼有抽象具体双重特征的概念类别。几乎所有的物性概念都可以加以引申,而作为抽象概念的 u 来使用。

总之,类别符号的连用或组合可以反映多方面的知识。跨类组合可产生新的概念类别,由此引发的联想是全局性的。从语义网络设计的角度来说,跨类组合属于抽象概念与具体概念之间、三大语义网络之间的模糊边界问题。这个问题虽然十分复杂,但上面的阐述和文【1】的概念类别定义表明,这个问题已得到了妥善处置。

6.2.2 抽象概念的综合知识

抽象概念一、二级节点的设计围绕着句类、语义块和语境这三个“主题”。基元概念为句类辨识和语境生成提供基本信息;语言逻辑概念为语义块感知提供基本信息;基本逻辑概念为基本判断句提供基本信息。而基本概念所提供的是一切信息的基本属性。

三大超级语义网络的一级节点都有各自的“集群”特性,这一特性在文【1】中已有阐述,这里将结合对基元概念设计过程的回顾和反思,作一点补充。

主体基元概念以作用为首,构成基元概念矩阵的 0 行。0 行的概念一定是作用型概念。但作用型概念也可从其他概念组合得到,这是作用效应链的基本观点之一,概念组合结构符“#”的引入即源于此。

主体基元概念以效应为果,构成基元概念矩阵的 3 行。3 行仅概括了效应的典型代表。但效应型概念也可从其他概念组合得到,这是作用效应链的基本观点之二,概念组合结构符“□”的引入即源于此。而且,从作用效应链的两极观点来说,过程、转移、关系和状态都是广义作用和广义效应。但是,从语言表达来说,作用效应链的六个环节是相互独立的六个表达角度。这就是六个基本句类的理论依据。

因此,六个主体基元概念二级节点的设计,不仅必须反映每一基元概念的总体特征,还应该考虑到每一基本句类的子类分类特征。这是主体基元概念二级节点设计的基本依据。

所谓基元概念的总体特征是指与全局联想有关的内容。如作用就要联想到作用的承受和反应;效应就要联想到变化和结果;过程就要联想到开始与结束、因与果、趋向与转化以及进展性与重复性等;转移就要联想到发和收、入和出、起点与终点、转移的内容、工具和途径等。

但是,并不是每一总体特征都具备构成句类子类的资格,我们只选择那些具备这一资格

的特征充当二级节点,而将其他的特征放到 0 分行里。过程的内容(指运动、演变和生命过程)在 0 分行,而转移的内容(指物和信息)却构成转移的二级节点,道理就在这里。

0 分行加二级节点,内容甚多。大脑思考的具体过程绝不会是对这些内容进行‘循环搜索’,而是以所谓‘数据驱动’的方式激发联想,但计算机的初期处理方式可以考虑从‘循环搜索’的笨办法着手,因此,给出搜索的顺序知识是必要的。这个搜索顺序的排定也许是概念知识库中可以实现(指库的内容和格式容易明确)的第一个项目。

综上所述可知,主体基元概念的二级节点配置包含了句类的基本信息。这就是它成为最基本的挂靠体的原因。除了 w 和 p 以外,l 类概念,复合基元概念的 6、9、c 行以及基元概念的 52 和 53 分行都向它挂靠。这一挂靠约定是主体基元概念的综合知识之一,应成为概念知识库的第二个项目。

关于挂靠的约定,现以表格形式综合如下:

表达对象	挂靠对象			
	基本概念	基元概念	主体基元概念	综合概念
语言逻辑概念	+	+		+
物与人	+	+		+
6m,9,c			+	
52,53		+		
500	+			

仅用挂靠方式显然不能表述人类所有的日常活动。有些日常活动需要复合型表示,我们抽取了三类活动,分别命名为‘劳作与服务’、‘交往与娱乐’、‘记忆与想象’,它们构成人类活动的第二类语境网络。第一类语境网络是【1】中所定义的‘a - b - d’行。关于语境网络将在下一节阐述。这三类活动也具有【1】中所阐述的 6、9、c 层次性表现,分别用层次符号 6i,9i 和 ci 表示,i 的取值从 6 到 8,以与上列三种人类活动相对应。例如,66,96 和 c6 分别表示本能、理智和社会层次的‘劳作与服务’,其他类推。所以,6、9、c 三行不仅是单纯的挂靠,还另有复合方式。人类日常活动的这三项特殊内容应成为概念知识库的第三个项目。

在本能活动的 6 行和一般智能活动的 9 行之间,插入了心理活动的 7 行和思维活动的 8 行。这个安排只是反映进化过程的自然顺序,无其他用意。

基本概念的集群特性比较简单,时间和空间为第一组,量、质、度为第二组,一切事物的基本属性为第三组。各组之间的关联性很弱,但组内的关联性很强。这就是基本概念集群性的主要特征。在基本概念之首的‘序’里,引入了广义空间和广义距离的概念。这些概念对于诱导概念的联想极为重要,在〈语义学日记选录〉中有详细阐述。

6.2.3 复合基元概念与语境

上一节曾提到,“基元概念为句类辨识和语境生成提供基本信息”。本节将对后一点作具体说明。为此,这里需要对 φ6 到 φd 的复合基元概念再作一次概括性说明。

(1) 心理活动及精神状态(φ7)

在所有的语义网络中,φ7 高层层次符号的层数最多,约定为 4 层。这一特殊约定当然与 φ7 类概念的特殊复杂性有关。自然语言理解对 φ7 类概念的隐知识揭示当前应取回避态度,HNC 也应如此。

φ7 类概念的相继出现,通常是表述一个事件或一项活动的起因;或是表述它们引起的后果;或是表述人们的状况或特色,从而构成一个句群的一类语境,HNC 将把它命名为 7 号语境(句间联想脉络)。这一联想脉络的激活不难从 φ7 类概念的简单统计结果而产生。为表明这一论断,下面给出 φ7 类概念节点的清单。

φ7 概念节点表(表中略去符号 φ,下同)

70	心理与精神活动
71	心理反应
710	一般心理状态
711	态度
7110	以理性为主要依托的态度
7111	感情色彩较浓的态度
7112	涉及自身素质或利益的态度
7113	对事业的态度
7114	对利害关系的态度
7115	对人际交往的态度
7116	对一般事物包括公益活动的态度
7117	对亲近关系者的态度
712	愿望
7120	一般愿望
7121	待实现的愿望
7122	已实现的愿望及其效应
7123	要争取实现的愿望
713	情感
7130	一般情感
7131	喜
7132	忧
7133	惧
7134	外触发为主的情绪表现
7135	爱

7136 恶
7137 恨
7138 内因为主的心理反应
7139 悔愧咎
713a 憾
713b 窘
714 情感与精神状态
72 精神状态
720 一般精神状态
7201 意志
7202 注意
721 能动性
7211 先天能动性
7212 后天能动性
722 气质
7211 性格
7212 秉性

对 7 号语境子类的划分,上面的节点表仅作了要点提示,当然需要更系统和深入的研究。

2. 思维活动(ϕ8,高层层次符号约定为 3 层)

ϕv8 是构成一般判断句的唯一依据,但 ϕv8 只是一般判断句的必要而非充分条件,句类的判定通常要依据具体词语的句类代码。对汉语 ϕv8 类的新词不难依据相应字义自动作出句类判断,但对于西语则几乎不可能,惟有请求人工干预之路了。

ϕv8,jlv1 和 jlv0 类概念的相继出现,通常是对一个事件、一种现象或一种状况进行预测或分析判断,从而构成一个句群甚至篇章的一类语境,HNC 将把它命名为 8 号语境(句间联想脉络)。同样,这一联想脉络也可以通过相应的简单统计而激活。jlv1 和 jlv0 类概念节点表已在文【1】中给出,下面仅给出 ϕ8 的概念节点表。

ϕ8 概念节点表

80 一般思维活动
81 认识与理解
810 一般认识与理解
811 分析与综合
812 演绎
813 判断
82 探索与发现
821 探索
822 发现
83 策划与设计

831 策划与计划
832 设计与规划
84 评价与决策
841 评价
842 决策
843 规定
844 约定
845 认定与承认

3. 专业活动与追求活动(ϕa,ϕb,高层层次符号约定为 3 层)

ϕa 类概念是语境特征最强的一类概念,他们通常构成作用句或含有作用的混合句类。这些句类的作用者或参与者不仅必须是人,而且优先于相应分行的 pay 或 peay 这些句类的 YC 优先相应分行的 gay 或 ray 这些句类的 XB 或 YB 优先挂靠相应分行的 pway 或 gway。由此可见,ϕay 类概念的相继出现,是表述相应 ay 子类专业活动的可靠信息。这些句类的检验信息简明而可靠。因此,由 ϕay 类概念形成的语境,将命名为 1 号语境,也称 1 号联想脉络。

ϕb 类概念与 ϕa 类关联性很强,两者互为因果。也可以说 ϕb 是对 ϕa 的另一综合观察角度。下面分别给出两者的节点表。

ϕa 概念节点表

a0 一般专业活动
a00 常规专业活动
a01 一般组织活动
a02 实施
a1 政治
a10 制度
a11 组织
a12 治理与管理
a13 政治斗争
a14 外交活动
a15 征服与反征服
a2 经济
a20 一般经济活动
a21 工业
a22 商业
a23 服务
a24 金融
a25 国家经济
a26 技术经济
a27 自然经济

a3 文化
a30 一般文化活动
a31 文学
a32 艺术
a33 技艺
a34 信息文化
a35 大众综合文化
a4 军事
a40 一般军事活动
a41 组织
a42 战争
a43 战争效应
a44 军事行动
a45 军事技术
a5 法律
a50 一般法律
a51 立法
a52 法律界用法
a53 法律与道德
a54 一般执法
a55 检察
a56 判决
a57 施行
a58 人与法
a59 违法
a5a 关系人用法
a5b 关系人反应
a6 科技
a60 哲理探索
a61 科研活动
a62 技术活动
a63 自然科学
a64 人文科学
a65 理论科学
a66 实验科学
a67 科技与生产力
a68 学科
a7 教育
a70 一般教育
a71 教

a72	学
a73	考
a74	学校教育
a8	卫生
a80	生命卫生
a81	查
a82	治
a83	养
a84	环境卫生
a85	全局性卫生
a86	局域性卫生
a87	局部及个人卫生

φb 概念节点表

b0	追求
b00	理性追求
b01	行动追求
b02	伴随追求的战斗
b03	对命运的抗争
b1	改革
b10	一般改革
b11	整体或根本改革
b12	局部改革
b2	继承
b20	一般继承
b21	发展继承
b22	消极继承
b23	局部继承
b3	竞争
b30	一般竞争
b31	攻
b32	守
b33	挑战与应战
b4	协同
b40	一般协同
b41	战略协同
b42	战术协同
b43	积极协同
b44	消极协同

从上面两概念节点表可以清楚看到 φay 与 φby 的交链式关联性和互为因果性。在下一节说明的概念“词典”中，φa 是最重要的部分。作用句和作用型混合句的句类知识库也要针对 φay 子类作重点说明。这些概念层面知识将对句类假设检验和解模糊处理产生立竿见影的明显效果。

4.（φy6，φy7，φy8）类概念（y = 6，9，c）

这三类概念在文【1】的复合基元概念说明中隐而未谈。那里只说到，观察人类活动需要引入本能—理智—社会三层次的概念，因为人类活动的这个三层次表现比较明显。这一特征用本体层 + 挂靠层的方式来表示，本体层用符号 φ6m，φ9 和 φc 来表示，挂靠层用主体基元概念表示，如 6.2.2 节中的挂靠表所示。

但是，人类活动的上述三层次特征不能只用挂靠方式来表达，还应有自身定义的方式。这就是引入（φy6，φy7，φy8）类概念的依据。在层次符号上，这三类概念不会与挂靠方式的（亦称扩展型）主体基元概念混淆，因为，主体基元概念的 y_1 和本体层 6m 中 m 的取值范围都是 0 5。而 φym 类概念 m 的取值范围是 6 8。

这三类概念的节点表如下：

（φy6，φy7，φy8）概念节点表

y6	劳作与服务
y60	一般劳作
y61	主劳作
y62	辅劳作
y63	专业劳作
y64	服务劳作
y7	交往与娱乐
y70	一般休闲活动
y71	交往
y710	会见
y711	接待
y712	访问
y713	别
y714	赠
y715	交往言辞
y716	交往动作
y72	娱乐
y720	生活娱乐
y721	欣赏
y722	艺术
y723	技艺
y724	健身

y725	狩猎
y73	比赛
y730	一般比赛
y731	智力比赛
y732	体力比赛
y733	对抗性比赛
y734	水平性比赛
y735	个人比赛
y736	团队比赛
y737	赌
y738	比赛科目
y74	行
y741	出差
y742	探亲
y743	游览
y744	探险
y8	记忆与想象
y80	忆
y81	想象
y82	信念
y83	红喜事
y84	白喜事
y85	法术

(ϕy6, ϕy7, ϕy8) 的设计,如果不计说明层 y,其高层层次符号预定为 3 层,这里只列出了 ϕy7 的全貌。通过这个概念节点表可清楚看到,(ϕy6, ϕy7, ϕy8) 皆各自形成相对独立的语境。三类概念的顺序是按人类活动‘进化’的观点来排列的,从联想脉络的清晰性来看,并非如此。但为了记忆的便利,仍按上列顺序安排句群联想脉络的序号,即将 (ϕy6, ϕy7, ϕy8) 概念形成的句群依次命名为 2 号、3 号和 4 号联想脉络或语境。这些联想脉络的二级或三级子类与相应概念节点层次符号的 2、3 层对应。

(ϕy6, ϕy7, ϕy8) 类概念通常形成混合句类,即使是与转移最相关的 ϕy74 也是如此。细心的读者会注意到,ϕy74 未设置 ϕy740 节点,因为它就是自身转移节点 ϕ22b 。

2、3、4 号联想脉络的显著特征是它们的二维性,另一维就是人类活动的本能、理智及社会性,由 ϕy 的取值提供信息。

5. ϕd 与 ϕ72 类概念

ϕd 与 ϕ7 特别是 ϕ72 强交链式关联,这两类概念的相继出现必然是对人类行为及其精神面貌的表述。在传记和小说,叙人和叙事的文章或谈论中,经常出现这样的表述段落。我们将把这一类表述定义为 5 号句群联想脉络或语境。下面给出 ϕd 的概念节点表。

ϕd 概念节点表

d0	行为
d00	一般行为
d01	行为与理智
d02	大公行为
d03	自私行为
d04	对他人的行为
d1	观念
d10	一般观念
d11	根本及全局观念
d12	观点
d2	准则
d20	一般行为准则
d21	社会性准则
d22	交往准则
d23	束己准则
d24	对他人的准则

本节到此为止,已全部补充说明了文【1】略而未谈的复合基元概念网络。设计复合基元概念各种局域网络的基本出发点就是对语境作出计算机可操作的分类,这种语境类别也可称为句群的联想脉络特征。为了便于语境生成模块[3]的软件实现,我们将这些句群联想脉络按其重要性和清晰性进行了编号。在第一阶段预定仅建立 8 类语境或句群联想脉络,上面已介绍了 7 个。

关于 5 号语境,这里顺便表达一项希望或一个观点。国外心理学早就建立了智商的概念,近来认知心理学又建立了情商的概念。从 HNC 的 5 号联想脉络来看,仅有智商和情商的概念是不够的,还应该加上‘德商’的概念。我们期望,这一观点能得到有关领域专家的响应。

最后要补充的是 6 号语境,它由 ϕ6m 类概念构成,其理自明,在〈语义学日记选录〉中有所阐述,这里就不重复了。

6.2.4 物表示的特殊处理

物表示的挂靠方法仅提供最基本的联想知识,这里‘最基本’的含义是从物与人的关系来确定的,主要是指物的功用。这一点对语言表达最重要。要让计算机理解物,只能也必须从这里入手,至于物本身的其他特征,不论它多么丰富多彩,都是次要的。例如,运输工具的一级近似表示都是 pw22b,首先把它同自身转移这一基元概念关联起来。二级近似再挂靠 jw5k,以表示水陆空的区别。这是物的知识表示的基本原则,其层次符号的设计一般不宜照搬自然科学的分类方法。

挂靠只是物表示的手段之一,不可能完全代替物本身的层次符号设计。层次符号设计的基本原则是“共性分层”。但共性的层次性依赖于观察的角度,对抽象概念可定位于某一观察角度,使共性的分层特性呈现稳定的序列。但具体物的表达往往需要多个观察角度,如果面面俱到,必然繁琐不堪。如果灵活排序,又必须付出另加说明的代价,这是一个很难处理好的矛盾。

对于这一矛盾的处理只能是见机行事。

以“国家”这一概念为例,面积和人口、发达程度、时空性、政治制度等都是它的基本特征,当然可以通过常识知识库详细表述这些特征。但是,在理解处理过程中,碰到一个具体国家就进入常识知识库去查寻,显然不是明智的做法。应在层次符号上给出最基本的信息,但又要适可而止。例如,没有必要在层次符号上把特立尼达和多巴哥与巴巴多斯区别开来,然而,把它们与卢旺达在地域上区别开来是有益的。

总之,物表示的层次符号设计不同于抽象概念,共性分层的原则基本不适用。对每一具体情况都需要作特殊约定。这些约定是概念知识库中一个大的类别。

物表示在挂靠层之后,还可以有自身独立定义的层次符号。在两者之间用“*”隔开。文【1】引言中概念组合结构符号集的挂靠结束符“*”即用于此项目的。wj 类概念已大量采用这种方式。

6.3 概念节点关联性知识

对概念关联性的表达,我们引入了两个术语:交式关联和链式关联。

从理论上说,交式关联是不同观察角度的重叠,是作用效应链各环节的交织性表现,链式关联是作用效应链各环节的因果性表现。例如,过程的生与死 gv14 交式关联于效应的产生与消除 vg31;效应的合分 vg390 聚散 vg394 交式关联于关系的结合与分离 v41。效应的扩展与缩小 vg34 链式关联于量与范围 j4;立与破 vg35 链式关联于内容与质 j40 和 j41;推动与抑制 vg36 链式关联于过程 g1 和度 j6。而 03 和 04 则与 d0 交链式关联,04 的关联性更强,因为行为的理性特征是约束而不是放纵。

从组合结构来看,交式关联表现在偏正、并或及内容结构中。链式关联表现在作用、效应、对象、内容及主谓结构中。两者只有一个内容结构的交叉点,由此可见:交式关联将主要表现在语义块内部,而链式关联将主要表现在语义块之间。当然,这里所说的语义块是指狭义的或简单的语义块,不包括具有句子特征或包含句子的复杂语义块。关于复杂语义块的构成,请参看【11】。

上面所说的交链式关联在组合结构中的表现,就是交链式关联性的二级分类。而交式和链式是概念关联性的一级分类。所谓概念关联性知识,就是指概念节点、概念集群、概念类别之间关联性的各级类别表现。这是一张非常烦琐然而脉络分明的关系网。其内容就是

概念关联性知识库。下面就来讨论编织这张网的一般技巧。

首先要从建库的立场对关联性的类别加以简化。把两个层次简化为一个层次,重新定义下列四种主关联和八种辅关联。

四种主关联是:

1. 狭义交式关联:概念搭配体现为语义块内部的广义同行优先。
2. B 关联或 B 函数:E、B 两种语句要素的概念搭配优先。
3. C 关联或 C 函数:E、C 两种语句要素的概念搭配优先。
4. A 关联或 A 函数:E、A 两种语句要素的概念搭配优先。

八种辅关联是:文【2】中所定义的七种辅要素,加一种互为因果关联。

1. 方式关联或 Ms 函数
2. 工具关联或 In 函数
3. 途径关联或 Wy 函数
4. 比照关联或 Re 函数
5. 条件关联或 Cn 函数
6. 因关联或 Pr 函数
7. 果关联或 Rt 函数
8. 互为因果关联或 M 函数

这一简化的主要优点是可以把概念关联知识同语义块感知处理及句类分析更紧密地结合起来。

每个基元或基本概念的节点构成概念关联知识库的"项目",每一"项目"下根据情况设置若干上列关联性"栏目",每一"栏目"下填写相关的概念节点,并标明其关联性的强弱。

这个知识库可称为概念词典,但这里的"词汇"是层次网络符号。

可考虑不同版本的词典,以适应不同处理的需要。

对 jl 和 jw 概念,应建立专门的概念词典。

6.4 句类标准格式知识

句类标准格式是句类知识中最基础的知识,简称格式知识。

关于七个基本句类的概念层面知识,或简称句类知识,将在文【14】到【20】中分别说明。上一节所说明的概念关联性知识实际上就是句类知识的一部分。

这里所说的句类标准格式知识,简单地说,是指关于一个句子的主语义块配置的知识。详细地说,则包括文【2】所定义的语句物理表示式和语义块的构成知识。例如,反应句的标准格式有以下四种:

X2B + X2 + XAC

X2A + X2 + XBC

XAC + X2B + X2 + X2C

X2B + X2C

第一种格式表述的各要素关系是:反应者 X2B 对作用者 XACA 及其表现 XACC 产生反应 X2。第二种格式实际上是第一种格式的翻版,只不过强调反应者处于主动地位。第三种格式表述的要素关系是:作用者及其表现使反应者 X2B 产生反应 X2 并导致后续反应 X2C。第四种格式是第一种的简化。

这四种格式中,第三和第四种相应于"2 + 2"和"1 + 1"句式。但前两种格式的句式是不确定的,他们可以是"1 + 2"或"2 + 1"句式,也可以是"2 + 2"句式,因为复合语义块 XAC 或 XBC 可分可合,这就是上面所说的句类格式中隐含的语义块构成知识。显然,句类格式及句式并不能在概念层面完全确定。但是,有些概念节点具有明确的句类格式及句式要求,例如概念节点"意愿"7121 就要求上列格式 2 的"2 + 2"句式。这样的概念层面知识显然是十分宝贵的。

反应句格式和句式的多变,为各句类之首。然而仍能在概念层面提取一部分确定的格式及句式信息,其他句类这样的信息就更为丰富。这方面的详细情况在阐述句类知识的【14】到【20】中说明。把所有在概念层面可提取的格式及句式知识,包括上述反应句格式 1 和 2 所表达的模糊性知识,综合起来构成一个知识库,是一项极为重要但不难实现的工作。

当然,句式及语义块构成知识主要是通过词汇层面的语义结构方程[7]来给出。但也有不能在词汇层面完全确定这两类知识的情况,而且语言常有"违例"表现。因此,决不能仅仅依靠词汇层面的知识。在知识的运用方式上,概念层面与词汇层面知识的配合是理解处理软件设计中的重大课题,这里就不来多说。

6.5 层次网络符号到自然语言词汇的反映射知识库

此知识库的数据结构也采用词典方式,但词典项目不仅是概念基元,还包括所有的复合概念,因此,此词典的"词汇"量远大于前述概念词典。

此词典的内容是各语种的词汇,不是每一语种一部词典,而是各语种合在一起,如同多语种对照词典,但对照的共同标准是概念层次网络符号。

此词典用于是人机对话系统和机器翻译的语言生成。

显然,两种应用对此词典的要求差异甚大,应考虑分别建库。

结 束 语

本文实质上是文【1】的续篇。

文中着重介绍了四类概念知识：

1. 概念类别知识及概念节点配置知识；
2. 复合基元概念与语境的关系；
3. 概念节点关联性知识；
4. 句类格式知识。

这四类知识都需要建立相应的概念知识库，当然，某些知识可直接纳入 HNC 理解处理程序自身的库函数。这些知识与语种无关，是理解处理最基础的知识。句类分析，或模拟人的思维过程对自然语言进行理解处理，必须从这一知识的运用入手，也可以说，它是语义层面理解处理的起点。

1995 年冬

关于汉语 HNC 知识库的建设

引　　言

没有语言知识库,就没有自然语言理解处理技术。同样,没有 HNC 知识库,也就没有 HNC 技术。但 HNC 知识库的建设经历过非常艰难曲折的历程。没有现在的 HNC 联合攻关小组,就不会有 HNC 知识库。这篇短文首先向联合攻关小组的知识库建设者致敬。

本文将讨论四个理论性问题,希望对 HNC 知识库的建设能起到一定的指导作用。但许多具体问题的解决还有赖于知识库建设者自身的创造性。然而,可以肯定的是,HNC 知识库不会是怀疑论者所说的那种无底洞,它是一定能够胜利完成的,并将对句类分析技术的形成作出决定性的贡献。

四个问题的题目是:

(1) 语言知识库建设的历史回顾

(2) HNC 知识库建设回顾

(3) 概念层面知识

——“先高后低,两步到位”是 HNC 词知识库建设的必由之路

(4) 句类知识

7.1 历 史 回 顾

人工智能(AI)必须以知识为依托,自然语言处理(NLP)必须以语言知识为依托,这是基本常识,没有人对此提出过异议。但是 AI 和 NLP 最需要什么样的知识?这些知识如何表达,又如何获得?这是知识库建设的基本问题。对这个问题的认识自 AI 诞生以来,已有了巨大的进步,但从 NLP 的需要来看,这个进步是远远不够的。

AI 的早期发起者几乎将知识混同于规则,这是不奇怪的,因为规则易于为计算机所把握。利用规则进行推理的过程,可利用产生式给以形式描述。这样,计算机的程序就可以在形式上模拟大脑的思考。如果大脑的思考过程仅仅是逻辑推理,那么,知识等同于规则的认识就是正确的。但是,大脑的运作过程不仅是推理,或者说,它的运作过程主要不是采用产生式规则的形式。但推理终究是大脑运作的基本表现之一。因此,规则的运用(不称“规则派”,本文一概不采用对前人贴标签的办法,因为这种办法在大多数情况下有失公允)仍然可以取得显著的效果。20 世纪 70 年代崭露头角的专家系统就是规则运用的巨大成果。最

近,IBM 的‘深蓝’计算机在与国际象棋世界冠军卡斯帕洛夫的人机大战中赢得了胜利,应该说体现了这一运用的顶峰成就。

逻辑推理对自然语言处理、语言学和词知识库的建设都有重大影响。在语言学上的近期突出表现是蒙塔古语言学的兴起,在知识库建设上的集中表现是美国的 CYC 计划。至于 NLP,应该说,到目前为止,所有的 NLP 处理系统从早期的 LUNAR 和 HEARSAY 到最近的 LeMICON 都是规则系统。尽管后者的知识获得是自学习的,但知识的运用仍然是规则的。

以产生式形式表现的规则就是逻辑学的蕴含关系。它是推理的基本形式。按照逻辑学的观点,知识就是一系列的命题,命题之间存在推理关系。规模空前、推理规则达 100 多万条的 CYC 知识库就是基于这一思路花了 10 年时间(1985 1995)建立起来的,当初其主建者曾宣称,到世纪之交,CYC 知识库将成为计算机的基本配置之一。但是,到 10 年届满时,这个梦想完全落空,CYC 被一些人视为失败的典型。主建者虽然不承认失败,并雄心勃勃地宣称,将以 CYC 为基础开展英语和西班牙语互译系统的研制,但从他的一句带有慨叹性质的话‘A word is a world’中可以看到,主建者对词知识库的建设显然有些心有余悸。

CYC 建设的 10 年期间,正是语料库语言学和功能语法大发展的 10 年,但主建者对前者似乎置若罔闻,这成了批判者的基本论点,但主建者心里明白,他所追求的知识不是简单的统计可以得到的。那么,CYC 的根本问题何在?

根本问题在于该知识库的目标和知识表示方式。

CYC 知识库主建者将 CYC 的目标定位在建立一个万能的 NLP 系统上,以弥补领域专家系统的不足。例如,一个心血管疾病的诊断专家系统并不能辨认患者年龄与体重的填写错误,CYC 系统可以帮助它解决这类问题。显然这涉及浩瀚无边的常识性知识,如果对这类知识采用一阶谓词加自然语言的方式加以描述,数以百万计甚至千万计的规则也难以包容,因此,CYC 含有 160 万条规则是不奇怪的。但是,问题的要害不在于一阶谓词,而在于以自然语言充当命题的概念表述符号,这是规则膨胀的根本原因。

上述 CYC 的目标应该说是 NLP 的天职。主建者在语料库的呼声压倒一切时不逐时流,按既定方针坚持到底,值得钦敬。问题在于 CYC 的目标不可能一蹴而就,主建者犯了 70 年代山克先生同样的错误,在沼泽地上建立高楼大厦。

NLP 的基础是语言知识,在语言知识里既包含与语言形式无关的概念知识,又包含与语言形式有关的纯语言知识。在概念知识里,又有高层共性知识与低层个性知识之分,我们将把前者简称为概念知识,而将后者称为常识性知识。

将知识划分为概念知识、(纯)语言知识、常识性知识,并分别建库,这应该是知识库建设的第一条根本原则,CYC 及迄今为止的所有知识库都没有遵循这一原则。

知识库建设的第二条根本原则是应将服务目标首先定位在自然语言五重或三重模糊的消解。口语五重模糊和书面语三重模糊的消解是理解的前提。但模糊消解的具体办法多种多样;消解的过程与理解的过程既有同步性又有异步性;模糊消解的深度是可测定的,而理解的深度是不可测定的(至少在目前);对模糊消解进行假设检验可以是无条件的,而理解是

有条件的。如果说,NLP 的最终目标——如同大脑一样理解自然语言——过于遥远,那么能否把模糊消解作为近期目标,集中兵力予以突破?

计算语言学界对此并未形成共识。从理论上这个问题很难阐述明白,但从语言信息产业的角度来看,则可以说是一目了然。语音识别、文字识别、全文检索、机器翻译、文字校对等方面都已有应用软件投放市场,这些软件的共同弱点何在?就是在模糊面前无能为力,而用户对此又十分敏感。因此提高语言信息产品的市场信誉,从而提高市场占有率的根本出路在于提高消解模糊的能力。这一点,不应存在任何疑义。

明确词知识库建设的这一中心目标十分重要,因为它关系到知识项的选择,关系到人工建库方式与语料库运用方式的分工等重大决策。

知识库建设的第三条根本原则是应以句类知识为核心。

这是 HNC 理论的必然推论,本文将围绕这个问题进行阐述,这里先从历史回顾的角度予以引言式说明。

谈到语言知识,人们首先想到的是语法。语法这个词汉语原来是没有的,是从西方引进的,但这不等于说汉语传统语言学没有语法的概念,只不过表明语法对汉语传统语言学所面临的问题不十分重要罢了。

语法有狭义与广义之分,狭义语法是指以形态变化和虚词搭配为依托的语言法则。这些法则里本来包含语义信息,但语法学从自身研究的便利出发曾长期有意脱离语义而自成体系。先叔祖父季刚先生正是基于这一点而将‘马氏文通’戏称为‘狗屁不通’的。这个状况直到乔姆斯基的转换生成语法和菲尔墨的格语法出现以后才发生了变化,随后的功能语法继承了乔姆斯基和菲尔墨的传统,这些语法应称为广义语法,它包含了语义甚至语用。以广义语法为基础的复杂特征集表示方法是语词知识表示的一个巨大进步。它吸收了广义语法的最新成果和明斯基的 AI 框架知识表示思想,为新一代 NLP 处理技术作出了重要贡献。

应该指出,广义语法学虽然融入了语义知识,但并未对语义表述给出完善的理论框架,因此,现有复杂特征集样式的词知识库从理解处理的需要来说,还远不够完善。

HNC 知识库希望从根本上改变这一状况。“根本”的具体表现就是以 HNC 的概念表达符号体系和语句表达符号体系进行词知识的表达。

最后应该顺便提一下,与 CYC 同时进行建设的大规模语言知识库还有美国的 WordNet 和日本的 EDR,这两类知识库存在的根本问题与 CYC 相同,但主建者总体思路还不如 CYC,这里就不加评述了。

7.2 HNC 知识库建设回顾

服务于汉语理解处理的语言知识库应划分为以下三种:

〈1〉非单字词知识库(以下简称词知识库)

〈2〉字知识库

〈3〉音节知识库

后两者是汉语的特殊需要。音节知识库仅用于语音文本的理解处理,它不是同音字知识的简单堆积,而是音节感知意义的提炼。字知识库只能充当音节知识库的后盾,而不能代替它,换句话说,音节知识库必须是独立的。

汉语的字兼充语素和词的双重角色,汉语的词以双音词为主,这一状况使得词知识库和字知识库不存在天然的界限,但两者又不能相互代替。

从技术实现的角度来说,以字知识库为主体比较合理。由同一汉字构造的不同双字词,有相应的共性知识和个性知识,如果把共性知识放在字知识库中,词知识库只存放个性知识,显然可以大大节省知识库的存储空间。在 HNC 处理技术的发展初期,主要是运用词的共性知识,这样,字知识库为主体的方案就很容易与急功近利的客观需要不谋而合了。

HNC 知识库的建设在"字为主体"的框架下"挣扎"了三年之久,直到 1996 年底,在张全博士的推动下,才决定放弃,改用词为主体的方案。

字主体和词主体两种方案的本质区别在于:前者需要进行词知识的共性与个性分离,而后者不需要。这一分离工作原则上十分简明,实行起来却相当复杂,只适合于个人的精工细作,而不利于开展"兵团作战"。

如果对字或词的知识表示像词典那样,以词义的解释为主,那么,字主体方案的合理性对汉语是天经地义的。但是,"义"的注明只是词知识的基础部分,从知识表示需要的空间来说,它只占很小的比例。词知识应主要由语用知识构成,HNC 把必须表示的语用知识浓缩成句类知识(见第 4 节)。对句类知识可以有两种表达方式:启发式和规则式。字主体方式适合于采用启发式,但过多采用启发式,对 HNC 处理技术的初期成长极为不利,而词主体方式则可以不受限制地采用规则式。

上述两点终于使笔者"回头是岸"。

这里应该指出,字主体方式必须仔细考虑字、词知识库的恰当分工,因而两库自然毕其功于一役。但词主体方式可以不考虑两者的分工,走先词后字之路。这正是当前 HNC 词知识库的建设之路。

词主体的 HNC 词知识库完成以后,可为字知识库的建设奠定良好的基础,但字的语用性有别于该字构成的词,字知识不是相应词汇知识的逐项迭加(词典正是采用这种方式),需要对相应词知识项进行删除、合并、增设的复杂处理,才能形成相应的字知识项,这实际上就是进行词知识共性与个性的分离处理,不过在程序上与字主体方式"反其道而行之"而已。

经过知识共性与个性分离处理的字知识库是一个理想的字知识库,它反过来可以用于词知识库的数据压缩,但这些不是当务之急。当前 HNC 处理技术最急需的字知识是字独立使用的语义和它形成组合词时的语义语用特性,后者更应该是字知识库关注的焦点。

知识库的建设通常是一个严格的"照章办事"的过程,"章"决定知识库的水平,"照"决定知识库的质量,"照"严格依附于"章"。但 HNC 知识库的建设有所不同,它的"照",既有严格的一面,又有灵活的一面。这是 HNC 知识库建设中的根本特点,它来于 HNC 符号体系的特

点和现状、HNC 处理技术的发展步调和当前急需。

前阶段词知识库建设中的主要问题是“规章不全”,经过几个月的实践,现在可以制定较为完备的规章了。这个规章里包括了语义块分离及变换、句类转换、内外混合句类的表示方案,这些表示方案的得以制定,应归功于几个月的兵团作战方式,是词知识库建设的重要进展。

7.3 概念层面知识

HNC 词知识库建设者当前遇到的困难之一是 HNC 符号的填写。发生这一困难是非常自然的,因为从文字符号到 HNC 符号的转换(或映射)不是简单的映射规则所能说明的,需要对两套符号的内涵及外延意义都有深刻的领会,而且映射方式并不唯一,许多情况不存在标准答案。基于这一情况,提出了逐步到位的填写方式,即先填高层或主要的 HNC 符号,并允许使用变量符号 y 和 t,y 用于高层,t 用于底层。HNC 符号的每一个字母或数字都有其独立和确定的意义。因此,HNC 映射符号具有“残而不废”的特点,文字只能在有限的情况下作有限的分解,“局部字”是有条件的,无“半字”、“1/3 字”或“1/4 字”之说。HNC 映射符号则不同,它可以全面分解,局部映射是无条件的,“半映射”、“1/3 映射”或“1/4 映射”都无条件成立,都有确定的含义。因此,逐步到位映射方式的运用不仅是临时的客观需要,也有其深刻的科学性。事实上,HNC 映射符号在词知识库全方位完工以后,还有一个综合调整和优化的过程。

当然,这并不是说 HNC 映射符号的填写不需要精心推敲,更不应该误会“残而不废”的提法而漫不经心。HNC 同样会有错别“字”,而错别“字”是绝不允许的。对 HNC 映射符号的深刻领会必须有一个“愤启悱发”的过程,这一学习方法的提出,主要是针对 HNC 符号体系的,因而只适用于这一环节。HNC 符号目前还没有形成《HNC 符号体系手册》这样的完备资料,使得这一学习方式显得更有必要。

HNC 符号是句类辨识的基础,而句类又是词知识框架的基础。这两项转换都处在概念层面。转换中的某些“细节”可意会而难以言传,而意会的关键在于对概念层面知识领会的深度。因此,本节专门讨论概念层面知识。

第 1 节说到,概念知识是语言的共性知识,与语种无关。“语言”知识是语言的个性知识,与语种有关。这一说法可作为“概念知识”与“语言知识”这两个特定概念的定义。对每一个具体的词,其知识都有这两个层面的表现。例如,一个动词,它必然蕴含句类信息,并与特定的句类代码相联系,这是概念层面的知识。这两项信息是一类动词(不一定同义或近义)的共性,是一个概念集合的表现。对这类信息 HNC 用简明的句类代码来表示,代码所表示的详细内容存放在概念知识库里,例如句类代码:X22,XR011,XR0110 所对应的句类标准格式分别为 X22J = X2B + X22 + XAC,XR011J = A + XR + RB2,XR0110J = A + XR + RB2 + RC。代码附属于具体的词,存放在词知识库里,代码所对应的句类格式则存放在概念知识库里,

这就是概念层面知识与语言层面知识的分工,也是概念知识库与词知识库的分工。

知识除表现在概念和语言层面之外,还表现在常识层面,后者往往与具体概念相联系。而对于具体概念,往往不必作概面层面与语言层面的区分,可统称为语言层面。例如,“北京”这个词,其语言知识首先是‘中国首都’,其次是‘经纬度’。至于北京的历史、人口、面积、经济文化现状及设施、名胜古迹、交通、气候、特产等,都应纳入常识层面。

这就是说,概念、语言、常识三层面知识有不同的观察角度,不同观察角度所看到的景象也就不同。知识总体是一个角度,词汇是另一个角度,在词汇里,还有动词与名词,体词与谓词之分,名词里又有抽象与具体之分,等等。

本节将着重阐述从知识整体角度看到的三层面表现。这首先需要说明一下 HNC 的观点。

概念层面的知识是启发性的,高度浓缩的,居高临下的。这是概念知识的基本特征,规则性知识则恰恰相反。CYC 知识库无概念知识库为先导,一头扎进规则性知识的汪洋大海,付出甚大而收效甚微,不能不说是一个沉痛的教训。HNC 词知识库的建设,当然应该吸取这一历史教训,加强概念知识库的建设。

从知识库建设的实际需要出发,明确提出三层面知识的划分,HNC 是始作俑者。但是,类似的提法在语言哲学里早有讨论,不过,那些讨论都未深入到概念知识范畴的具体范定,而这是建设概念知识库的关键。

从 HNC 理论来看,概念层面知识的范畴是明确而清晰的,其主体构成是下列四类知识:

(1) 概念层次网络符号体系(下文有时简称节点表)

(2) 网络节点之间的交式及链式关联

(3) 句类格式及其转换规则

(4) 特定句类特定广义对象块素的特定概念优先性要求

在这四项主体构成中,与词知识库建设密切相关的是第一项和第三项。下面就来对这两点做要点说明。同时也旁及第二点及有关问题。

概念层次网络符号体系是对语言概念的总体性表述,或者更准确地说,是对自然语言概念体系进行总体描述的尝试。这尝试二字十分重要。因为,设计者的目的不仅是希望这个符号体系成为表述语言概念的数字化符号,而且希望它们成为概念联想的激活因子。这是一个理想,一个必须为之奋斗的目标,在笔者看来,也是 NLP 的必由之路。

这一理想的彻底实现,也许不是一代人的努力可以完成的,但 HNC 终究迈出了关键性的第一步,这一步的集中标志就是节点表和句类格式。

节点表和句类格式知识能否在解模糊处理中发挥关键性作用呢?

首先,这取决于 HNC 理论的合理性、完备性和明确性。其次,它取决于 HNC 符号体系的软件可解释性。需要指出的是:HNC 符号体系本身只蕴含概念联想的启发性知识,而不是规则性知识,软件对启发性知识的运用,即把启发性知识与现场信息结合起来,形成判断,是一个复杂的转换过程,这一转换的机制是大脑的奥秘之一。HNC 对这一转换机制的形式

化表述,当前仅归纳为三个要点:(1)同行优先;(2)交式及链式关联;(3)句类格式。

下面就来对这三点分别进行说明,说明中不可避免地要用到许多 HNC 特有的术语,不熟悉这些术语的读者,请重温【1】【2】的有关说明。

7.3.1 关于同行优先

"同行"是 HNC 的基本概念之一,意思是"高层层次符号相同"。"同行"有本义和扩展义之分,本义同行简称同行,指概念类型符号相同,五元组符号可以不同的同行。扩展同行指挂靠层的同行,简称挂靠同行。本义同行概念的关联性是概念所固有的,HNC 妙手偶得之,通过五元组符号予以凸现。挂靠同行的概念关联表现是人为的,但同样反映关联性的客观实际。

同行优先的含义可概括为以下六点:

(1)五元组不同的同行抽象概念优先语义块内部的偏正或反偏正结构;

(2)五元组相同的同行抽象概念优先语义块内部的联合结构,特别是 E 语义块的高低搭配结构;

(3)挂靠同行具体概念与相应的抽象概念优先动宾或主谓结构;

(4)部分五元组不同的同行抽象概念优先动宾结构;

(5)x 与 w 的同行概念优先偏正或反偏正结构;

(6)x 与 x 或 w 与 w 的同行概念优先联合结构。

同行优先所体现的概念关联性当然只是概念关联性的一部分,但它是最常用、最明朗的部分,HNC 符号通过五元组、挂靠和概念内涵的层次化数字表示,对这部分关联性给出了极为简明的联想激活信息,并使关联程度(或语义距离)的计算成为简单的"与或"运算,笔者认为,至少在这一点上,HNC 处理的效果和效率应与大脑相近。

这里应特别指出,挂靠同行是一个很有应用价值的课题。但各挂靠项的安排尚待深入研究,从而制定出明细的准则。

7.3.2 关于交式关联

交式和链式关联是 HNC 必须引入的两个概念。这两个概念是在节点表设计过程中形成的,开始时只是对面临的诸多矛盾的一种解决方案,最终才发展成为表征概念节点关联性的两个基本特征量。

"交"的概念来于集合论的"交集",逻辑学的"合取",粗浅地说,两个概念节点交式关联就是指"两概念有交叉",深入地说,交式关联是同一概念本体从不同观察角度看到的不同映象。这里说的"不同观察角度"就是作用效应链的六个环节。例如,"死亡"这一"概念",从过程来看,它是代谢的"谢"142;从效应来看,它是"消失"312;从状态来看,它是"减少"500412,因此,过程节点 142,效应节点 312 和状态节点 500412 是交式关联的。同样 141,311,500411 也是交式关联的,其上层的概念节点 14,31,50041,也就必然具有交式关联性。"合与分"这

一对偶概念有效应的 39 和关系的 41,因此 39 与 41 交式关联“得与失”这一对偶概念有效应的 3a 和关系的 46,因此 3a 与 46 交式关系。“运动过程”109 与“转移”22 交式关联“约束”04 与“抑制”362 交式关联,过程的“趋向与转向”13 与效应的“变化”309 交式关联等等。总之,交式关联性在概念节点之间普遍存在,从概念节点去考察关联比直接从自然语言更简明更纯净,也更为抽象和深刻。这类关联性不可能从现有的熟语料库直接得到,只能依靠人工方式先建立一个框架,但这一关联性知识库(概念层面的)的完善,则必须利用大规模语料库。

当然,概念层面的交式关联不能替代词汇层面的搭配研究,因为它排除了关联性的个性或语用特征,后者对语言生成极为重要。虽然从解模糊来说,共性知识的作用大于个性知识,但知识库的建设决不能忽视个性知识。

交式关联节点具有共享 u,z 概念的宝贵特性。这样,“同行优先”的概念就可以从狭义扩展为广义,从而扩大了该准则的应用范围。

对广义同行优先需要进行深入的研究,但这一研究安排在全方位的知识库和句类分析软件全面完成之后,才较为恰当。

7.3.3 关于节点表(兼及工作单填写策略)

概念节点存在交式关联意味着,节点表的每一个概念节点具有概念基元与概念集合的二重性。所以,不能将 HNC 的概念基元混同于义素。义素学所追求的义素是如同化学元素一样的基本语义单元,HNC 所追求的则是反映语言概念总体框架的结构单元,义素未考虑概念总体的构架特征,而概念基元是承上启下的构架,同时又是一个可独立存在的单元。

抽象概念基元都具有五元组特征,而义素不具有这一特征。这突出表现了两者的本质差异,义素有动词义素、形容词义素、名词义素之分,概念基元则没有这种区分。无动词节点、名词节点或形容词节点之说。

节点表的每一个节点都附有汉语命名,但必须说明,这只是一个命名,不能把这个节点的内涵混同于命名的词义,更不能把以动词命名的节点误认为是动词节点,或名词、形容词命名的节点,误认为是名词、形容词节点。命名的词义只反映节点内涵的概要,其全部含义需要一个细则说明,但写这项说明的工作量十分浩繁,只得暂付阙如,HNC 理论应编著多语种合一的《HNC 反映射符号体系手册》和《HNC 句类知识手册》。这将是两部历史性的“手册”,HNC 联合攻关组应承担起这一历史的重任。但当前只能参阅文【14】到【21】,略资弥补。

但应该指出,目前的节点表作为语言概念总体表述的框架,是完备的,并且是基本合理的。我们可据此先完成汉语的反映射符号体系手册。换句话说,第一部完善的映射手册只可能与第一套 HNC 词知识库同步完成。

上述说法似有循环论证之嫌,但这正是 HNC 符号体系设计者的基本思路之一,是辩证法的体现。所谓完备性,是指概念类(即 ϕ,j,l,f,s,w,p,x,jw,jl 十大类)及其高层节点的完

备性。对完备性的具体含义,可作如下陈述:对主体基元概念,完备性是指句类的完备,或全局联想脉络的完备;对表述人类活动的复合基元概念,是指语境信息基本依托的完备;对语言逻辑概念,是指语义块辨识信息的完备;对基本概念,则是指基本哲学意义上的完备。上述完备性是通过各类概念的高层节点配置来实现的。至于高层节点以下的底层节点可以也必须在 HNC 词知识库全方位完成以后才能全面展开设计,"可以"是因为底层节点已不蕴含基本句类知识,"必须"是因为底层概念的个性表述有赖于对相应概念集的综合分析,而词典不可能提供这个条件,但一个仅有高层表示的全方位 HNC 词知识库却不难提供。质言之,就是第一个 HNC 知识库可以也必须走"先高后低,两步到位"之路。

当然,"先高后低,两步到位"策略的实现,还必须拥有技术上的充分保证,而这是可以实现的。因为,第一,HNC 符号"残而不废"的特点使得仅有高中层表示的 HNC 词知识库不影响前期 HNC 处理软件的使用;第二,体现"残而不废"的高度层次化的数字表示方式使得 HNC 词义表达获得了极易于扩充和修改的特征。

"两步到位"的策略不仅适用于词义表示,也适用于块素概念优先性和句类变换等项目的表达。但不适用于句类代码、语义块构成和语义块分离三项目。格式代码也应力求一步到位,因为,这四个项目的错误将直接造成 HNC 处理软件的"方向性"错误。

综上所述,不难明白,HNC 工作单的填写应遵循下列基本守则:

〈1〉词义表达的高层 HNC 符号应力求准确。未定义的中层符号可用 emn 表示,底层符号可用变量 t 表示。

〈2〉句类标记和相应的句类代码应力求准确和完备。但是,极为罕用的句类代码也可以暂时置之不理,以免过分增加句类分析的负担。

〈3〉句类格式代码应贯彻"抓两头"的原则,即仅表示禁止的和极常见的格式,不要贪求无所不包。

〈4〉语义块的良性构成、隐含内容的 A、B 类语义块和语义块的多种构成方式,必须按规定给出表示。

〈5〉语义块分离现象必须按规定给出表示。

〈6〉句类变换按"最低要求"填写。

〈7〉块素概念优先性按配套原则填写。

〈8〉类别符号要力求完备。

为了实施上列基本守则,应尽快制定 5、6 两项的有关标准。

本节最后需要对节点表作两点补充说明:

第一,部分节点表未完成高层设计,这包括 ϕ 类的 714 行,f 类的 1、2、41 行,l 类的 5、6、7 行。对这些节点,可用变量 y 代替未设计的层次符号,其他节点则不允许。

第二,jw 已完成的设计应严格遵守。其未完成部分及 w,p,gw,rw 的挂靠准则应于近期进行一次专题研讨。

7.4 句类知识

以句类知识为中心进行词知识库的建设，是 HNC 词知识库与传统词知识库的本质区别。20 世纪 80 年代末期以后的词知识库吸收各种广义语法的最新成果，运用复杂特征集的知识表示方式，在知识范畴方面，也包含了主辅语义块数量、各块的语义角色、块素的概念类别等信息，其知识范畴的广度几乎与 HNC 无异。那么，两者究竟有何区别？

本质区别可归结为两点：

〈1〉一个是提纲挈领，一个是有领无纲，这个纲就是句类。

〈2〉一个是以 HNC 符号表达意义，一个是仍然以自然语言的词汇或词汇的某种类聚表达意义。前者能充分激励概念的联想，后者不能或基本不能。

本节就来从纲的作用谈起。

各种广义语法的基本思路是给句法分析的旧框架披上格语法的新衣。但句法框架始终处于究其然而不究其所以然的本暗末明状态，格语法则始终未能突破只见树木不见森林的困境。如果问句法分析，为什么动词有及物与不及物之分？有“单位、双位、三位”之分？如果问格语法，自然语言需要多少个格才达到完备？他们都会瞠然不知所答。为什么？因为他们都没有抓住句类这个纲。

从句类分析来说，上述问题都是一目了然的，动词的及物与不及物、“单位、双位、三位”之分完全取决于它所隶属的句类，格数量的完备取决于全部基本句类的主块数量加上八种辅块及其子类。

句类分析是 HNC 自然语言理解处理技术的核心，是 HNC 理论的必然技术归宿。句类分析不排斥句法分析，但仅把它作为适应不同语种需求的辅助环节。就汉语的解模糊处理来说，这一环节的作用甚微，远远小于印欧语系。

句类知识是句类分析的主要知识来源。直接服务于句类分析、必须在词知识库中加以明确表示的句类知识有以下八项：

〈1〉句类代码

〈2〉格式代码

〈3〉句类变换信息

〈4〉语义块构成及其分离变换信息

〈5〉块素概念优先性知识

〈6〉句类外混合信息

〈7〉主辅语义块变换信息

〈8〉名词的角色信息和语义块构成信息

这八项知识的填写是 HNC 工作单的主要内容，它们的填写细则另有说明。这里仅就三方面问题作一些要点说明。

7.4.1 关于句类代码和格式代码

这里说的句类代码是基本句类和混合句类的代码。

句类代码是八项句类知识的纲。

从形式上看,代码只是压缩了数据,句类代码不过是一种信息编码而已。但句类代码不是一般的信息编码,它所编的是语句语义构成在概念层面的全部信息。它是 HNC 理论所发现的“7 个基本句类和 36 个混合句类可以构成语句语义空间的完备表述”这一重要结论的具体体现。完备性就表现在每一自然语言的动词都存在相应的句类代码。或更准确地说,就是自然语言的任何语句都有确定的相应句类代码。或再换一句话说,目前的句类代码表是完备的句类表示。

每一种句类代码对应着多种格式代码,现在给出的格式代码表穷尽了自然语言可能出现的全部排列形式。看起来非常复杂,但其内在规律极为简明。倘若完成了华罗庚先生所说的“由厚而薄”的过程,就不会有复杂之感了。

句类代码代替语句物理表示式的标准格式。格式定义为主语义块的排列顺序,标准格式定义为“第一对象语义块排在第一位,特征语义块排在第二位,随后的排列顺序约定对象语义块在前,内容语义块在后”。按照这样一套约定,就不需要在各语义块之间给出标志,或者说不需要对各对象或内容语义块给出标志。标志之说不适用于 E 语义块,因为它“鹤立鸡群”,自成标志。汉语及 VO 型语言存在这里所定义的标准格式,OV 型语言(例如日语)则不存在。但这并不影响标准格式的理论意义。

从句类代码表可以看到,一个基本句类最多只有四个主语义块,这不是人为约定,而是基本句类的内在特征,是考察了全部基本句类之后的结论。当然,这个说法不包括转移句和效应句中含 $\sum$JK 的情况。

基本句类中有三种特殊情况应该说明。

一是表中所给出的作用效应句和因果句实质上是复合句类。因为它们不只一个特征语义块,但我们把它们作为基本句类来处理。

二是两对象效应句和单向关系句,它们的对象语义块有“1,2”之分。“1”相应于广义作用方,“2”相应于广义效应方,含主次、因果、源流之分,填写时要细心思考。

三是基本状态句,它具有表中所列举的三种格式,这三种格式都符合标准格式的定义。因此,它可以说是一种有三种标准格式的特殊子类。例句如“主席团坐(在)台上”、“主席团(在)台上坐(着)”、“台上坐(着)主席团”这里虚词“在”与“着”的取舍,不属于语义逻辑意义的范畴,由“在”字和“着”字的 HNC 符号给出。

7.4.2 关于混合句类和复合句类

混合句类仍然只有一个 E 语义块,因而仍属于基本句类。复合句类则至少有两个 E 语义块,因而不属于简单句类。

混合句类代码形式为 E1E2 * kmn,k 表示混合句类 JK 的总数,m 表示取自 E1 的 JK 个

数,约定从 JK1 起,n 表示从 E2 所取 JK 的起始序号。复合句类定义为至少有两个独立的 E 块 E1 和 E2,分别来自不同的简单句类,并至少共用一个 JK 语义块。复合句类代码的简单形式为 E1 * E2 * mn,m、n 是共用语义块对 E1 和 E2 句类的 JK 序号。

7.4.3 关于语义块构成、分离及变换

这是三个互相关联的复杂问题。

由于语义块有复杂构成,形成了分离的土壤,分离又往往伴随着变换。

复杂构成定义为语义块有多个块素,分离定义为一个语义块的块素之间插入其它语义块。变换一般定义为主辅语义块之间的整体变换,这里则指分离出去的主块块素变换为辅块。

语义块的复杂构成应区分 E 块和广义对象块两类。两者的构成方式和分离方式都不同。对 E 块的分离,不必考虑变换。

E 块的复杂构成一般可表达为

$$E = QE + EQ + EH + HE$$

详细论述见【14】。

广义对象语义块的复杂构成,按内涵总是先分解为对象和内容两部分,按形式总可以分解为 KQ,KH 两部分,这两种分解形式都可以延拓。对主语、宾语、状语的复杂构成,统一作对象内容之分是 HNC 语义块构成理论的要点。

从复合语义块中分离出去的块素有两种归宿,一是被别的语义块吸收,二是独立出来自成一块。无论是哪一种情况,都会给句类分析带来麻烦。因此,对于这些可能出现的分离情况,必须在词汇层面给以说明,为计算机提供引导信息。

在某些情况,分离出去的块素不仅独立成块,而且变换成辅块,对这种情况,也必须加以说明。

1997 年 8 月

附录 给苗传江、庄咏璆的便笺

苗、庄:

昨夜梦中想到,混合句类代码应有反结构,符号为:E1E2 * kmn。“反”的约定是:先取 E2 的 JK,后取 E1 的 JK,其他不变。此举,可彻底消除年来就此事之困扰。节点 y80 多属 X20Y * 21 混合句类。

1998 年 6 月 18 日凌晨

语义块的切分组合处理

语义块的切分组合处理是自然语言理解处理必不可少的预备步骤,任何语种都不例外。就汉语来说,它是分段层选处理[9]的后续步骤。

语义块是语义层面的语句结构单位,是句类分析的基础[2]。语义块的切分组合处理,顾名思义,就是把一个句子分成若干个语义块,并对每个语义块再分出它的核心部分和说明部分。前者是语义块的切分处理,后者是语义块的组合处理。

像分段层选处理一样,语义块的切分组合处理只是优先级的判断,把优先考虑的语义块组合方式及其他有关信息通知随后的句类分析系统,最终的确认则有待句类分析之后。

本文分为四节,分别讨论:语义块切分的基本准则;语义块切分处理的策略分析;语义块的组合处理;主辅语义块定义的相对性及其他。

11.1 语义块切分的基本准则:lv 准则

我们曾将语义块切分的基本准则概括为 lv 准则。这里的 l 和 v 就是概念层次网络理论所定义的类别符号,l 代表语言逻辑概念,v 是抽象概念五元组的动态表达。不能将 lv 准则里的 v 理解为动词,首先,五元组的 v 概念本来就与动词的概念有区别。第二,uv 和 vu 类概念,特别 jlvu12 和 jlvu13,都是语义块切分的重要标志。第三,动词这个语法概念,在西语可作为 E 语义块的切分标志,但不适用于汉语,因为汉语是通过词在语句中的表现来体现它的词性。词性在西语是词的因属性,在汉语是词的果属性。西汉的这一差异,对 E 语义块的形式辨识会产生一定的影响,但从下面的讨论可以看到,若采用句类分析的路线,汉语词性的模糊并不构成 E 语义块辨识的根本障碍。

lv 准则也可分别称为 l 准则和 v 准则,l 准则可用于所有主辅语义块的辨识,v 准则则仅用于 E 语义块的辨识。本节只讨论 l 准则,v 准则放在下一节讨论。

在文【1】中曾指出:语言逻辑概念 l 的设置和具体设计,是围绕着语义块的切分和组合这一中心目标而展开的。具体说来就是:l0 和 l2 是主语义块指示符,l1 和 l3 是辅语义块指示符,l4 是语义块组合指示符,l8 是辅语义块说明符,l6 和 la 是 E 语义块的说明符,lb 是句间说明符,l9 虽不是语义块的直接指示符,但有间接指示广义对象语义块的作用。

下面从信息指示的角度对 l 类概念和 l 准则作进一步的说明。

首先是 l0 到 l3 子类,下文将简称切分类。当句子偏离标准格式时,必须加入这个子类的有关词汇,对有关语义块给出标志信息。当然有例外情况,如诗词里"红旗漫卷西风"之类

的倒装句。如果自然语言的句子都死板地采用句类标准格式,那么可以说,广义对象语义块的切分标志就没有引入的必要。换句话说,标准格式的语句里不存在主语义块切分信息,l准则仅用于判定辅语义块,对主语义块的判定只有 v 准则可以利用。所以,标准格式语句的主语义块切分通常反而比较困难。句类分析的一项统一操作就是将非标准格式转换成标准格式,同时将句子里原有的切分标记词汇撇开不管。但这里应强调指出:含切分标志信息的词汇里不仅包含语义块的切分信息,也含有语义块的类别信息和句子的类别信息。语义块类别信息由 l 类概念本体层的第二层提供,语句类别信息由挂靠层提供。但这种类别信息并不一定确切,因为词汇的映射符号往往是多义的;语句类别信息还可能不存在,因为映射符号可能是高层的,即不带挂靠层。切分处理的使命是将所有这些信息都收集起来,并转换成适当的表示,供后续的句类分析使用。

其次是 l4 子类,下文将简称组合类。从理论上说,语义块的组合标志,应该包括概念组合结构的全部内容,因为概念的组合结构同时也是语义块和句子的组合结构。从语言深层来看,语义块和句子都是复合概念。但复合的方式则各有侧重。语义块侧重偏正、并、或三种结构,句子侧重主谓、对象和内容三种结构。偏正结构用于语义块核心部分和说明部分的分隔,并、或用于核心或说明部分的内部组合。所以,组合类逻辑概念主要是组合符号 / , ; 的等效物。这类组合结构理应与句类无关,因此,似乎没有必要像切分类那样引入挂靠层,其实不然。对切分类,挂靠层的作用仅在于指示句类,而在一般情况则用于意义的精确表示。就偏正结构来说,其内涵十分丰富,英语的 of 和 's 就代表两种不同意义的偏正结构,使用挂靠层就能给出精确的表示。

组合类语言逻辑概念的语种个性很强,但语种个性的表达又十分复杂,目前只好采取模糊方式,因为这个问题显然不宜"独家包办"。对汉语表示偏正结构的字"的"和英语表示反偏正结构的词"of"分别采用 l41 和 l42 来表示,而西语的关系代词则采用 l40 表示,以显示它们的语法功能区别。这种表达方式显得烦琐,但容许改变,权作引玉之砖吧。

机器翻译的目标语生成过程需要进行语义块组合结构的变换,我们希望将这一变换过程所必需的基本信息纳入 l4 的本体扩展层。但目前尚未作具体设计。

第三是 l8 子类。从语义块切分标志的意义上说,它是 l1 的补充。这就是说,由 l8 标志的概念一定是辅语义块的切分标志。但此类语言逻辑概念有自己的独立意义,它们必须有挂靠层。这里"独立"的意思是:其逻辑意义不是相应挂靠层意义的简单逻辑对应,而是有所扩展。例如,汉语"根据"义项之一的映射符号是 lv80121,而"从"和"到"的义项之一则分别是 lj80425 和 lj80426。

第四是 l9 子类。这个子类是唯一不能直接作为语义块切分或组合标志的逻辑概念,但它常用于广义对象语义块的开头,而且绝不会出现在一个复杂语义块的末尾(除了头尾合一的简单语义块),这个特性在许多情况下对于语义块的切分是一项重要的参考信息。这一特性仅为指示代词所有,而为其他代词所无。正是由于这一点,我们仅将指示代词纳入语言逻辑概念 l9,而将其他代词作为一般抽象概念或 p 概念来处理,目的是突出语义块的标志信

息。我们约定：l 后面直接跟层次符号，一定是语义块的切分组合标志信息。如果 l 后面加五元组符号，则是一般语言逻辑概念，是否充当标志信息由语句的具体情况决定。

第五是 l6 和 la 子类。l6 用于时态说明，la 用于逻辑态说明，是从副词中抽出的两个子类，目的是为了突出 E 语义块的标志信息。这两个子类的概念一定是 E 语义块的说明部分。在四种主语义块中，E 语义块最难辨识，因为它可以分离，而汉语尤为困难，因为它没有中心动词的形态标志。专门设置这些子类，有助于减轻这一困难。它后面的第一个 v 概念可作为 E 语义块核心部分的优先候选。

最后是 lb 子类，它是句间的切分标志，当然也是语义块的切分标志。

上面曾谈到，非标准格式的语句通常要给出广义对象语义块的切分标志。辅语义块的排序则与格式之标准与否无关，那么，辅语义块是否一定给出切分标志？答案是：一般如此，但有特例。特例主要是条件辅语义块。

l 准则的运用，从语义块切分组合的角度来说，仅涉及对 l 类概念层次符号第一层的操作。对第二层及其后面的挂靠层的操作，已超出切分组合的范围，其作用是提取语义块的类别信息和句类信息。所以，严格地说，l 准则不能仅理解为语义块切分组合准则。

11.2 语义块切分处理的策略分析

上一节讨论了语义块的切分准则——lv 准则，由这一准则不难制定相应的软件运行规则。如果一个句子里支持这些规则的信息齐备无缺，既不模糊，又无冗余，则语义块切分处理可以说是轻而易举，几乎不必多置一词。但实际处理时面对的语句往往是不仅信息不齐备，而且是模糊与冗余并存。因此，语义块切分处理不仅要利用语言逻辑概念提供的信息，还要利用其他的知识，才能取得一定的应变能力。这就是说，语义块切分处理不仅需要一系列规则，还需要运用规则的规则，也就是策略。

策略的要点是随机应变。对当前面临的问题，就是要对待切分语句的特征表现进行剖析，然后确定相应的处理步骤。

语义块切分分类处理是从分段层选处理到句类分析的过渡，它面对的是分段层选后的句式：

$$J(i,k) = \sum W_{ikn} \qquad (1)$$

J 代表句子，J(i,k) 表示一个待处理句子的各种可能形式。此式对下标 in 求和。W 代表“音”或“词”，i 代表段的序号，n 代表多音段某一特定层选中的音词序号。这里的“音”一般指单音词，也代表双音词甚至多音词。这里的“词”指双字词或多字词。下标 ik 代表多音段 i 的 k 号层选的意思。而左式的自变量 (i,k) 则代表多音段层选的各种组合。设 Σ(i,k) = N，则此式表明：原始文本的一个句子经分段层选处理以后变成了 N 个句子。即使原始文本是无模糊的文字文本，这个表示式依然成立，不过各分项有“词”无“音”，N 较小，甚至 N = 1。

语义块切分分类处理的任务是把分段层选句式(1)变换成下面的语义块句式(2)：

$$J(i,k) = \sum K_{ktm} = \sum K_{kt} \quad (2)$$

此式对 t 求和。式中的 K 表示语义块，t 表示语义块的形式序号，m 表示句类编号。这里提前说一句，句类分析的第一步操作是将上面的形式语义块句式变换成句类表示式：

$$J(i,k) = \sum K_{mk} \quad (3)$$

此式对 k 求和，k 的意义同上。

从表示式（1）经过表示式（2）到表示式（3）的两步变换都是一一对应的。就这两步变换而言，原始文本的模糊与否只表现为量的区别，即使是语音识别输出语音阵列的巨大模糊也不过是增加工作量而已。但句类分析的最终目标是对表示式（3）的多个样本作出唯一性判断。虽然这个“多选一”处理已转变为简单的极值求解问题，但原始文本的巨大模糊可能造成假极值，因此，模糊性的最终影响也会表现在质的方面，即可能造成误判。

现在回过来讨论从式（1）变换到式（2）过程中可能面临的具体问题。

式（1）中的分项 W 有模糊集和安全岛两种情况。在原始文本模糊度较大时，虽然层选处理过程自然地（顺带地）进行了双音词到双字词的解模糊处理[9]，但残余的双音词仍然存在。单音词的模糊度虽已大大减小，但依然比较严重。下面的讨论以这一复杂背景为出发点。

依据 lv 准则，不言而喻，语义块切分处理的起步操作是从式（1）的分项序列中依次找出 l 类或 v 类概念。于是，面临的第一个困扰是：l 类概念模糊时如何处理？第二个困扰是：v 类概念多次或连续出现时如何处理？

第一项困扰实际上就是一个典型的多义选一问题。概念相关系数或语义距离的计算是消除这一困扰的根本途径，但同时也要利用某些先验知识。因此，从方法学来说，语义块切分时对 l 类概念的多义选一，同层选处理时的单音词位置判定完全一样。这两个问题形式上差别很大，实质上是同一个内容：“在若干种概念组合方式中，哪一种组合的置信度最高？”这就都变成了语义距离计算问题[3]。

当然，语义距离的具体计算过程对每一个具体问题有所不同，先验知识的利用也有差别。这个不同和差别主要表现在知识利用的深度上。层选处理时，语义距离的计算和先验知识的利用主要在段内进行，必要时才向外延伸。而语义块切分的着眼点是在待处理音节的前后两个方向进行联想处理，根本不考虑段的内外界限。下面，让我们通过具体的句例来印证上述论点。

例 1：

yi ding… yao… ba… zhe ge… xiao xi… gao su li… xiao jie

一定　　这个　　消息　　告诉 *　　小姐 *

肃立

告诉 *　高速

小姐 *　小节　小结　消解　小解

这个句子经分段处理后分为七个音段,其中只有一个需要作层选处理的三音段,对三音段必须作跨段处理,处理后不难(确实不难)得出如下结论:

第一,此三音段应选上层。

第二,上层的双音词‘gao su’应取双字词‘告诉’。

第三,最后一个双音段里的双音词也应取双字词‘小姐’,这是对双音词‘告诉*’进行跨段处理的副‘产品’。

不带调的拼音输入经层选处理后,绝大部分模糊已经消除,此例 S(i,k) 的具体形式有 S(6,1) 和 S(6,2)。但后者的置信度很低,只需要考虑一个句子 S(6,1)。这个句子总共八项,五个双音词都变成了双字词,模糊项只是三个单音词‘yao,ba,li’。

对这个句子依次扫描过去,头两项都含有语义块切分信息,但暂时按下不表。第三项“ba”含有 l0,这是一个非常重要的广义对象语义块切分信息,虽然它的类别信息是模糊的,既可以是作用对象 l0200,也可以是转移内容 l0320,但这对切分无关紧要。麻烦的是‘ba’中还有其他的常用独立义项,其中作为工具量词‘把’和基本数字‘八’的两义项在这里必须考虑。那么,这些重要信息从哪里来?答案只能是音节感知库。这里顺便说一句,汉语必须建立音节感知库的观点据说不容易得到认同,但这个例子应该是一个颇有说服力的证据。

对‘ba’进行双向联想处理,扫到第四项‘这个’lgu91 时,即可肯定它必须取义项 l0,扫到第五项‘消息’和第六项‘告诉’时,可进一步肯定它必须取义项 l0320。其中的‘告诉’是本句扫到的第一个也是唯一的一个 v 类概念。到此已切出了两个语义块:转移内容(T3C)语义块转移特征要素(T3)语义块,扫到最后一项时,也就完成了第三个语义块信息转移对象(T3B)的切分。

到此为止,不仅完成了对例句 1 的语义块切分,同时也完成了语义块和句子的类别辨识,一举两得,在讨论层选处理的文【9】中对此有所说明,软件设计应该适应这一情况。

按照从表示式(1)到表示式(2)的严格变换要求,式(1)的头两项还没有着落,这将在第 3 节讨论。至于单音词‘li’的模糊,超出了本文的范畴,这一类的模糊问题,将在文【10】中集中讨论。

对于第一个困扰的对策已如上述,下面来讨论第二个困扰的对策,但先从具体的例句谈起。

例 2:

jiao yu bu jiang zhao kai quan guo zhong dian da xue xiao zhang hui yi

教育 … 召开 … 全国 重点* 大学 校长* … 会议*

预卜 过重 电大学校

tao lun shen hua jiao yu gaige wen ti

… 讨论 … 深化* 教育 … 改革 … 问题*

花椒*

重点* 终点 钟点

校长 * 嚣张　销帐　小帐　消长
深化 * 神话　神化
会议 * 回忆　会意

全句分为九段,其中的八音段和四音段经层选处理后不仅可确定仅取上层,而且八音段里的双音词 zhong dian 和 xiao zhang 已转化为双字词“重点”和“校长”。这里将假定三音段的层选未定,双音词 hui yi 和 shen hua 未消。语义块切分处理时面临的是两个含有双音词的分段层选句式 S(1,1) 和 S(1,2)。

现在,让我们越过语义块切分处理,先对这两个句式作一点综合分析。我们已得到六个双字词:“教育,召开,大学,校长,讨论,改革”。它们都属于基元概念的“社会性活动集群”,这个集群具有最清晰的联想脉络,在作者的〈语义学日记选录〉中曾写道:“从人类的万千活动中分出一项专业性活动,谈不上什么学问,但对于理解十分重要,我倾向于把它列为第一号联想主脉络。这一联想主脉络的特点是:二级联想脉络的界限最为分明,概念 ai,pai,peai 之间的搭配优先性几乎是绝对的,这项知识对于解模糊及纠错处理极为宝贵”(1994.11.8)。这里剩余的三音段和双音词模糊都可以通过此联想脉络的先验知识予以消除,但这一知识的运用最好以句类分析为依托。下面就来进行具体的分析,同时也是对上述论点的论证。

在上列双字词中,“召开”的信息含量最大。虽然从词性来说,它是汉语里常见的 vg 型概念。但它又是一个作用效应型概念,必须有内容的补充。“会议”是它的天然搭配。因此,从“召开”可以把双音词 hui yi 转换成双字词“会议”,这个转换可完全基于“召开”的上述概念特性和概念组合的同行优先准则,而两者死搭配知识的有无应无关紧要(当然,在词库里应该有这一项知识)。

其次,从“召开,教育,大学,校长”诸词可知此句是关于教育活动 a7 的作用效应句,作用者 A 优先于 pea7 和 pa7。这个结论是由语言知识和概念知识[6]的综合运用而得到的,这是一项极为重要的信息。作者愿意不厌其烦地再次指出,句类分析的基点就是千方百计地提取这一信息:语句类别及各语义块之间的关联性约束。一旦掌握了这一信息,句类分析的中级和高级处理就有了坚实的基础,而层选及切分处理中遗留下来的模糊也就不难解决。具体到这里的三音段层选,显然,“教育 bu”与 pea7 的语义距离远小于“jiao 预卜”与 pea7 的语义距离。依据语义距离最小准则能够以相当大的置信度作出“教育部”的选择。这里需要用到 bu 和 jiao 的音节感知知识,两音节的感知义项都比较多,前者含汉字“不,部,步”的主要义项,后者含汉字“交,叫,较”的主要义项。但难点不在多而在“最小”的估计,这个问题在文【3】中有专题讨论。

这个句子里的 vg 类概念仅双字词就有“教育,预卜,召开,讨论,改革”五个之多,双音词里还另有 vg 类概念。如何从这许多 vg 类概念里找出“主角”E?这就是上面说的第二个困惑。汉语“字义基元化,词义组合化”的根本特点使得汉语的双字词很少具有单一的词性,这对语法层面的句子分析,确实是一个巨大的不利因素。对语义层面的句子分析当然也会带来一定的影响,从形式上说,似乎出现了消除词性模糊的要求。而这一模糊的消除,又不能

仅求助于语义距离的计算。因此,第二个困惑似乎成了汉语理解处理的巨大障碍。

其实不然,关键在于对语句总体结构知识即句类知识的运用。

在文【2】中曾将语句的总体结构概括为语句的物理表示式,并具有 1 + 1,2 + 1,2 + 2 及 1 + 2 四种句式。前两种相应于基本句类,只有一个中心动词。第三种相应于广义作用效应句,表达对象和表现都是两个,因此中心动词不是一个而是两个。汉语的兼语句属于这一情况。第四种是上列三种情况的‘补’,这里的‘1’表示一个主体表达对象,这里‘2’是‘多’的意思,表示该主体对象有多项表现,但通常是两项,而在每一项表现里又都可以涉及另一对象或内容。汉语的连动句属于这一情况。

上述语句总体结构的复杂性表明,即使不存在词性模糊和词的多义性,语义块切分的最终拍板也可能要等到句类分析之后。但是,在‘复杂的总体’当中,存在三项简明的、可把握的特征,这就是 C 语义块的语句扩展性,表现类词汇的语义兼备性和兼词性概念的直接组合性。这三项特征,是语义块切分处理的三把钥匙。

“C 语义块的语句扩展性”这个概念,包含了两层意思。第一层是:某些及物动词不仅必须搭配对象,还必须搭配内容,否则其意义就不完整。第二层是:这个内容可扩展为另一个语句,从而导致第二个 E 语义块的出现。关于及物动词的这一重要特征,在文【7】和〈问答 34〉里有详细说明。2 + 2 的语句结构就是这一特征的语句级表现。这个信息表明,E 语义块的切分,并不是在任何情况下都是‘漫漫黑夜’,至少在‘2 + 2’情况是相当光明的。如果我们找到了一个‘必须同时搭配对象和内容’的动词,如本例的‘召开’,就等于找到了一把对号的钥匙。这类动词,层次网络符号给了明确标志:结构符号用 # 表示,词义库中的类别符号用 v 表示,语义结构方程用 4—4 表示。

“表现类词汇的语义兼备性”这个概念的含义取决于对‘表现类’一词的定义。在文【2】中曾指出:“一个语句的内容无非是两个方面,第一是表达对象,第二是对象的表现”。从字面的意义来说,这个提法本身是不严谨的,但该文对这个提法作了较严谨的说明。这里需要回顾的一点是:对象及其表现,无论在词汇、语义块或语句级都不是截然分开的,两者可以合一。“表现类”就是两者合一的意思。明确了这一点,“表现类词汇的语义兼备性”的意思就不言自明了。汉语里这类词汇特别发达。层次网络符号对它们作了细致的划分:结构符号分别用→|表示,词义库中的类别符号分别用 vb,vc 和 bv 表示,语义结构方程分别用 6—5,7—5 和 8— 0 表示。这些信息如同 l6 和 la 一样,也是语义块切分的辅助标志,两者的差别在于:表现类词汇如果不另加组合标志,它通常是语义块的核心部分,并在语义块的尾部。而 l6 和 la 一定是语义块的说明部分,并在语义块的首部。当此类词汇连续出现时,一律划归一个语义块,而把进一步辨识的任务留给句类分析去处理。它们的连用一定是相应于简单的 1 + 2 结构。汉语对 1 + 2 结构语句的表达常表现出惊人的简洁,实有赖于此类词汇的特别发达。如‘回乡探亲”下海经商”上山砍柴”抗美援朝”进城访友”担水浇地”去医院看病”出国攻博士学位’等等。

“兼词性概念的直接组合性”的说法具有较强的实用色彩,是针对汉语里‘动词连用’的

困扰而提出的。汉语的所谓‘动词连用’现象实际上都是 vg,vr 和 vu 型概念的连用,纯 v 型概念的直接连用也是不符合汉语习惯的。这与同行五元组优先搭配时的‘同性相斥’现象出于同一道理。

按照‘兼词性概念直接组合’的原则,如果 vg 和 vr 类概念连续出现,则在语义块切分处理时,将‘不管三七二十一’,一律并为一个语义块。如本例中的‘讨论深化教育改革’,因为它们都是 vg 类概念。这里顺便说一句,上文所举的表现词汇的连用实例就是这一原则的体现。

当几个 vg 类词汇连用时,谁充当主角 v?这个问题可以给出明确的答案,一般是第一个 vg 类词汇充当主角 v,随后的 vg 类概念是它的内容。这是所谓 VO 类语言的固有特征。与句类标准格式的机制相同。

到此为止,我们已从语句总体结构的角度说明:汉语词性模糊所带来的所谓‘动词多次出现或连用’的困扰,从语义层面来看并不严重。若能进一步运用到文【2】中指出的 la,l6,j1vu 和 uv 类概念所提供的附加信息,我们甚至可以说,这一困扰在许多情况实际上并不存在。就本节的例 2 来说,如果在‘召开’这个位置上,不是一个作用型、而是一个普通的 vg 概念,照样可以确定它的 E 要素身份,因为它的前面是一个逻辑副词 luva 的音节 jiang。

关于‘动词多次出现或连用的困扰’,应该提出一个反问:为什么动词就不能多次和连续出现,而其他词类却有此‘特权’?这是‘中心动词说’的束缚。这个说法是汉语理解难于西语的主要理论依据。例如下面例句的汉英对照分析,

他跑着回来告诉我们这个消息

He came running back to tell us the news

英语显然比汉语容易,因为英语很容易找到中心动词‘跑着回来’。可是,‘跑着回来’与‘告诉’究竟谁是中心?不能厚此薄彼,两者是同样重要、同等地位的表现。强行约定一个中心,无助于理解,反而有多此一举之弊。所以,我们认为:中心动词的表述模式有局限性,它仅适用于西语语法层面,不适用于汉语,更不适用于语义层面。所以我们在在文【2】中提出了句类分析的理论模式。当然,这个理论模式还处于‘幼儿’时期,语义距离的计算只是理解的基本功,关键的检验在句类分析的中级和高级联想处理。而这一步工作才刚刚开始。

上面讨论了语义块切分处理的一般策略,没有考虑分段层选句式(1)的各种具体特征。而这类具体特征往往可以提供简明实用的信息。句式(1)本身的特征可概括为下列五点:

甲　语义块指示符基本齐备。

乙　没有 v 类概念。

丙　只有一个 v 类概念。

丁　没有任何指示符。

戊　语义块指示符不齐备。

这个排列顺序大体上反映了句式(1)的复杂递增性。前三类情况比较简单,后两类情况比较复杂,两者的复杂程度难分轩轾,上面的策略分析是面向这两类复杂情况的。

这五个类别本身也包含重要的信息:甲种情况相应于非标准格式语句,乙种情况相应于基本状态句,丙种情况相应于基本句类,丁种情况相应于标准格式语句。

除上述一般处理策略外,针对不同情况,还需要运用不同的先验知识。

在乙类情况下,切分处理不必管 SB 与 SC 的划分,仅将状态句的信息通知后续的句类分析即可。这种情况西语一般不会出现,是汉语的特色。在诗歌里,甚至可以仅有 SB 或 SC,如马致远的"枯藤,老树,昏鸦。古道,西风,瘦马",只有 SB。而李清照的"寻寻觅觅,冷冷清清,凄凄惨惨戚戚"只有 SC。

在丁类情况下,就需要运用句类的标准格式知识,即 EABC 相对位置的知识。

11.3 语义块的组合处理

有分必有合,分中有合,合中有分,从哲学的意义上说,把上述语义块切分处理叫做组合处理亦无不可。

但既然已把它命名为"切分",则本节所指的处理内容——对语义块进行核心部分与说明部分的划分,自然是换一个名称"组合"比较合适。

在语义网络 1 的设计中,为切分处理设置了十个一级节点,而组合处理仅有两个,这自然意味着组合处理相对说来要简单一些。

语义块的组合结构主要是"偏正并或",已如前述。

偏正结构的汉字标志有"的地之"。虽然"的"的义项很多,但不难辨识。不带调音节"de"的感知模糊度也不大。但在语音识别处理中,它很容易被吃掉,并干扰相邻音节的识别。这并非区区小事,值得汉语人机对话系统予以特殊关注。

并的汉字标志有"和同与及并"。"和同"既是组合的标志,又是切分的标志。它们还有其他义项。这里的多义选一比较复杂,处理过程应分两步,第一步判别它是不是逻辑指示符。如果是,才进入第二步,判别它是切分标志还是组合标志。第一步可依靠语义距离计算的常规手段。而这一手段对第二步无效,这一步的判断主要是利用句类知识。例如:

张先生正在同李小姐谈话。

张先生同李小姐到上海去了。

第一句话里的"同"是切分标志。"谈话"表明这是一个信息转移句,李小姐是"谈话"的对象。这里顺便说一下,对 vg 类的信息转移概念,通常直接在后面依次安排对象和内容,这就是信息转移句的标准格式。但"谈话"是 gv 类概念,由这类概念构成的信息转移句可省去内容,并将对象安排到 E 语义块之前,但必须加上对象指示符。

第二句话里的"同"是组合标志。"到上海去"表明这是一个自身转移句,上海是转移对象,张先生李小姐都是自身转移者。

要解决上一节"v 类概念多次或连续出现的困扰"和这里的"切分标志与组合标志的模糊",都离不开对句类知识的利用。这就是说,复杂的语义块切分组合处理实际上已跨入了

句类分析的范畴。在讨论分段层选处理的文【9】中也谈到了类似情况。这是自然语言理解处理各基本模块之间必然存在的交叉现象,或叫做相互支持。这里的支持确实是相互的,并不是只有低级的分段层选和语义块切分组合求助于高级的句类分析,而后者无求于前者。在文【9】中曾谈到:层选过程可顺便完成语义块类别和句类的辨识,因为这些信息都寓于 1 概念的层次符号当中,分开处理显然有悖于效率原则。这一点显然也适用于语义块切分处理。在切分处理过程中往往也就同时完成了语义块类别甚至语句类别的辨识。

11.4 主辅语义块定义的相对性及其他

本节谈三个问题。一是主辅语义块定义的相对性,二是对象语义块的简单和复杂构成,三是 E 语义块表达的分离性和兼并性。这三个题目都需要专文论述,这里将采取漫谈的形式,简略地谈一下要点。

11.4.1 主辅语义块定义的相对性

本文给出的主辅语义块定义是以陈述句为依托的。对于疑问句显然不适用。四种主语义块的核心部分和说明部分,七种辅语义块的内容都可以成为提问的主项。因此,陈述句语义块的主辅之分,对疑问句已失去意义。但是,疑问句的理解处理仍离不开按陈述句制定的句类分析模式。

在疑问概念的定义里,我们引入“静态问”、“动态问”和“内容问”的概念。这两个概念即源于句类分析的“对象和表现”概念。“静态问”相应于表达对象,“动态问”和“内容问”相应于对象的表现。提问时静态与动态的分野十分清晰,因此,汉语专门为它们设置了两个词:什么和怎么。西语对于新概念的表达通常是另造新词,但是对这两个极常用的概念却“反其道而行之”,采用了组合方式。这一点很有趣。在疑问句的分析中,区分“静态问”、“动态问”和“内容问”是最基本的分析环节,而在这一点上汉语是否比西语容易一些?

陈述句里的辅语义块在疑问句里可以是询问的主项,也就是从辅上升为主。这是主辅语义块定义相对性的第一层含义。它的第二层含义是:虽然在陈述句里语义块的主辅之分有其内在的主次差异(主次是基本概念的二级节点之一),但实际的主次在一定条件下是可以转化的。当陈述的对象或表现直接涉及作为辅语义块定义的有关内容时,这些内容自然也就成为主语义块的核心部分。此理自明,不必细说。

11.4.2 语义块的简单及复杂构成

在文【2】和本文的第一节里都将语义块的构成定义为核心和说明两部分,这个定义只适用于简单构成的语义块,尽管这种语义块的说明部分可以是另一个语句。在传统术语里叫定语从句。

复杂构成的语义块是指一个语义块直接由一个子句构成,或者说,这个子句的对象子块

和表现子块是作为一个整体构成一个语义块,其中无所谓核心部分与说明部分之分。如文【2】中的例句:

张先生怕李小姐发脾气

这里的"李小姐发脾气"这个子句构成这个反应句的语义块 X2AC,对这个子句的对象"李小姐"和表现"发脾气",虽然存在谁可省略的问题,但不能说可省略者为说明部分,而不可省略者为核心部分。

对象及其表现的组合方式,是语言艺术的精髓。主谓、动宾、偏正是三种基本的组合方式,对象在前表现在后者叫主谓,表现在前对象在后者叫动宾,但主谓和动宾都可转换成偏正,这种组合结构的转换,是形成子句结构或语义块的需要。例如此例中的"李小姐发脾气"可转换成"李小姐的脾气"或"生气的李小姐"。汉语对这一转换采用了极为简明的方式:即在两者之间加偏正结构指示符。"物价飞涨"这个过程句,可转换成"物价的飞涨"或"飞涨的物价"形式的语义块;"张先生去了上海"这个转移句,可转换成"去了上海的张先生""张先生之去上海"形式的语义块;"诸葛亮病了"这个状态句,可转换成"诸葛亮的病"或"病中的诸葛亮"形式的语义块;对于"中国队大胜韩国队"这个作用句,不仅可转换成"中国队的大胜韩国队"和"大胜韩国队的中国队"形式的语义块,还可转换成"中国队对韩国队的大胜"形式的语义块;同样,对"张先生支持李小姐"这个关系句,不仅可转换成"张先生支持的李小姐"和"支持李小姐的张先生"形式的语义块,还可转换成"张先生对李小姐的支持"形式的语义块。

由上面的论述可知,语义块构成的简单与复杂的提法只具有形式上的意义,它是语义块是否采用偏正结构的标志,如此而已。

这里还应该谈一下"C 语义块语句扩展性"的概念,既然 AB 语义块都可以是一个子句,那么,强调 C 语义块的语句扩展性有何意义?这是由于,C 语义块中的第二对象和第二表现与句中的第一对象和第一表现存在特定的联系:"第一"是因方或因,"第二"是果方或果,而且两个对象和表现都不可省略。

11.4.3 E 语义块的分离性和兼并性

ABC 语义块通常是"封闭"的,不容许其他语义块插足,但 E 语义块是"开放"的,容许其他语义块插入其间。我们将这一现象叫做 E 语义块的分离性,在文【2】中有详细说明。这里补充说明 E 语义块的另一特性——兼并性。

上面说的子句到语义块的转换就是兼并。兼并就是指主谓、谓宾甚至主谓宾的合并。三者的分合取决于表达的需要。但汉语特别热衷于并。这使我们产生了 E 语义块兼并性的说法,它特指谓宾的合并。这对于句类分析有一定的帮助。汉语有不少这类兼并性的词汇,这类词汇在现代语句词典里都加上"可插入"的标志。一个句子里如果用了这样的词汇,往往就不必另找谓语和宾语了,已经是"毕其功于一役"了。在句类分析时要清醒地运用这一信息。但另一方面,如果对这类词汇施行了插入,对分段层选处理就极为不利。如果分段

处理得到的一串单音词,既不像名字,又不像其他的新词,你就要考虑兼并类词汇的插入运用这一特殊方式了。

这就是引入这一概念的实用考虑。

1995 年春

后　记

本文成文较早。当时语言逻辑概念的 l5 和 l7 还处于备用状态,组合结构符号“对象”采用的是“→”而不是“&”,两类动宾组合词分别采用 vb,vc 表示,而不是后来的 vB 和 vC,主谓组合词采用 bv 而不是后来的 Bv 和 Cv,这些都未作改动。保留这些历史痕迹是为了 HNC 联合攻关组记住一条基本原则和教训,即 HNC 庞大的符号体系是以基元符号体系为依托,并通过类内组装和跨类组装两种方式来实现的。这需要周密的总体规划,而总体规划必须发挥团队的集体智慧。团队的每一成员都应有“以天下为己任”的雄心。

HNC 基元符号体系目前仅完成下列六类:(1)概念类别基元;(2)“词性”基元(即五元组);(3)语义网络节点基元;(4)语义块基元(E,A,B,C,Ms…);(5)句类基元(即基本句类);(6)组合结构基元。基元符号体系及其组装的思路还有待于扩展到语境生成和句群及篇章的处理。因此,充分发扬团队精神对于 HNC 的发展是事关成败的头等大事。过去的许多弯路正是由于团队力量未能充分发挥造成的恶果。

1998 年 8 月 25 日

作用、效应句的句类知识

引　　言

从本篇起到最后一篇,是本论文系列的第三部分,内容是各基本句类的具体知识。各篇所要阐述的问题和行文结构大体相同,但各有侧重。相同的部分有:

(1) 各基本句类的子类及其格式;

(2) 子类的特性说明;

(3) 例句分析;

(4) 句类辨识的主要标志,即决定句类特征的概念层次网络符号,但不包括语言逻辑符号的切分子类[11]。

这里顺便指出,语言逻辑符号切分子类和 E 语义块要素所给出的句类信息应保持一致。这个一致性要求,是消除有关词汇多义模糊或作为发现并纠正错误的依据之一。

七个基本句类大体上是每类一篇,但作了局部调整。一般作用句、效应句和混合句类中的作用效应句并为本篇。作用句中的承受句和反应句独立成篇。共八篇。

作用句按其二级节点配置应有五个子类,但句类格式只有三种,即一般作用句、作用承受句和作用反应句。约束是一般作用的特殊形式,其句类格式与一般作用句雷同。作用的免除一方面是约束的对偶,另一方面是反应的特殊形式,因此也没有独立的句类格式。

本篇按一般作用句、效应句和作用效应句分为三节。

14.1　一 般 作 用 句

本节分三小节。第 2 小节详细说明作用句特征语义块 X 的构成特性,并通过它说明 E 语义块的一般结构特征,其内容是对文【2】有关论述的虚实互补。

14.1.1　一般作用句的子类及其句类格式

一般作用句是作用句的基本子类,约束句又是一般作用句的唯一子类。但本论文系列并不严格区分作用句和一般作用句。下文将省略‘一般’二字。

作用句有三个语义块 A,X,B,共有 6 种排列组合形式,因此,其非标准格式应有 5 种。如下所示:

(1) A + X + B　　张三打了李四。

(2) A + B + X　　张三把李四打了。

(3) B + A + X　　李四被张三打了。李四挨了张三的打。

(4) B + X + A

(5) X + A + B

(6) X + B + A

在前三种格式的右方,给了例句。它反映了一个语言现象:就是汉语的非标准格式作用句优先于(2)、(3)两种格式,而西语优先于第 4 种格式。

14.1.2　X 语义块的构成特征

为了通过 X 具体阐述 E 语义块的结构特征,先给出上列标准格式例句"扩展"形式的三种示例:

(1) 张三又打了李四　　张三不得不打了李四

(2) 张三打了李四一顿　　张三打了李四两耳光

(3) 张三又打了李四一顿　　张三不得不打了李四两耳光

第一种扩展是 X 的前扩展,第二种是后扩展,第三种是前后都扩展。格式的表达可改写成下面的扩展形式:

$$A + (QX + X + HX) + B$$

这个扩展格式里引入了符号 QX 和 HX,分别表示 X 的两类说明部分。QX 代表 X 的前扩展,其内容是:由类别符号 uv,uu,vu,j11,1a,l6 所表达的概念,用于描述 X 的属性、情态、逻辑态及时态特征,前三者相应于属性特征,后三者相应于情态、逻辑态和时态特征。HX 代表 X 的后扩展,其内容是:由层次网络符号 j00,j12,j2,j01, j3,j4,j6 和 hv 所表达的概念,其中的基本概念用于描述 X 的静态特征,依次是 X 的顺序、时间长短、空间(作用点)、数量和程度的信息。例句中的"又"和"不得不"分别是逻辑态特征和情态特征,"一顿"是数量特征,"两耳光"则是数量和效应对象相结合特征。

类别符号 q 和 h 原定义为概念的前缀和后缀,这里改用大写字母,引申为 X 语义块前后搭配的意思,参看本论文系列的附表 2。严格地说,QX 是可前可后、以前为主的意思。HX 是可后可前、以后为主的意思。

X 语义块及其前后搭配可以分离,这是 E 语义块不同于 ABC 语义块的一项结构特征。但汉语的 hv 不能分离,这是语法学将它命名为助词的有力依据。

X 的前搭配是 X 语义块辨识的重要标志[11],但 X 的后搭配(除了 h v)不能作为 X 语义块的识别标志。这是前者的动态性和后者的静态性的必然推论,因为上列静态性内容显然也为 AB 语义块所具有,而动态性内容则为 X 语义块所独有。

对前后搭配中的 l6 和 hv,这里应该指出三点:第一,它们不仅是 X 语义块辨识的重要标志,也是汉语新词辨识的重要标志[10]。第二,它们是 X 语义块的核心部分,而不是像前后搭配的其他部分那样,只是 X 语义块的说明部分。这个提法或定义意味着将时态的表达纳

入 X 语义块的核心部分。仅就汉语来说,纳入与否无关紧要。但比较汉语和西语对时态表达的不同方式,就不难理解,这样定义有利于不同语种句类分析的统一性。第三,在文【1】中,我们把 Q 和 H 作为一种特殊的类别符号来说明,现在我们看到,它们不仅是类别符号,也是结构符号。具体地说,它们是逻辑并的一个特殊子类,不是一般的并,而是有次序、有主次的并。

在文【2】【11】中曾指出,语义块内部的组合结偏重于“偏正并或”三种,并指出语义块的核心和说明两部分“多数情况以偏正或反偏正结构组合而成”。这里我们看到了核心与说明以“并”的形式相互组合的“少数情况”,这就是 X 语义块的 HX 部分。

在上述两文中,我们还谈到了 AB 语义块的封闭性和 E 语义块的可分离性。这些论述里隐含着各语义块的内容彼此“泾渭分明”的假定。但现在我们看到,这个假定并不适用于语义块的所有部分,HX 可以违反它,例句中的“两耳光”就是明证。“耳光”这个词的含义十分丰富,其映射符号的一级近似是 rzz008,作为 r,它在句子里也充当效应对象的角色,但在上面的例句里,通过 HX 的组合方式把它转换成 HX 的一部分了。

本节引入了 E 语义块前后搭配的概念,并给出了 QX 和 HX 的详细清单,目的在于为中级联想处理提供具体的信息(联想脉络)。这里应强调指出:不同的 E 对 Q 搭配基本上“一视同仁”,这就是说,QE 与句类无关或关联性很弱,所以它们是 E 语义块辨识的重要信息【11】。H 搭配则与句类密切相关,HX 侧重于时间长短、数量和程度三者。如果是 X4,或 X 中显含效应,则应增加对空间(部位)信息的侧重,从下面的例句可以看到这一点。

14.1.3 例句分析

这里给出三段语料。

“督邮未及开言,早被张飞揪住头发,扯出馆驿,直到县前马桩上缚住;攀下柳条,去督邮两腿上着力鞭打,一连打折柳条十数枝”。(《三国演义》)

“白孝武接过刺刷,照哥哥孝文赤裸的胸脯抽击了一下,血流顺着胸脯一条条拉下来”。(《白鹿原》)

“随和可亲的鹿子霖率先抽了兆鹏一记耳光。……兆鹏被连续几个耳光击倒之后,黑娃觉得自己屁股上挨了重不可负的一击就狗吃屎似的趴下了”。(同上)

第一段语料:共七句。除第一句外,都是作用句,但格式各不相同。

“早被张飞揪住头发”,作用句:A——张飞,X——揪住,YB——(督邮的)头发。这是一个 B + A + X 格式的作用句,在 A 的前面加了逻辑指示符“被”,作用对象“督邮”被省略了,保留了效应对象“头发”,实际上出现了 B 块分离的复杂现象。

“扯出馆驿”,作用转移句:A 省略,TB 省略,E——扯出,TB1——馆驿。由“扯出”的映射符号 v008 # vg22b5 可知,这是一个作用自身转移句,句类格式是

A + XT2b + TB1 + TC

这类混合句的 A 也是自身转移者。这个信息上面的映射符号是没有包含的,但它指出了该句必须有 TB1,这里是“馆驿”。隐含的信息是 TB2,在下句给出。省略的 A 和 TC 承上文可知分别是是张飞和督邮。

“直到县前马桩上缚住”,转移作用句:TA——省略,T2——到,TB2——县前马桩,X4——缚住,A——省略,B——省略。由“到”的映射符号 v22b6 和“缚住”的近似(这里近似的意思是不管“住”的意义)映射符号 v04,可知这是一个转移作用复合句。“缚住”的作用对象在此句中属于隐知识,其空间信息与 TB2 合一,也是隐含的。

“攀下柳条”,作用句:A——省略,X——攀下,B——柳条。

“去督邮两腿上着力猛打”,是 A + B + X 格式作用句:A——省略,X——着力猛打,B——督邮两腿,这个 B 语义块里既有作用对象“督邮”,又有效应对象“两腿”。

“一连打折柳条十数枝”,作用句:A——省略,X—— 一连打折,B——柳条十数枝。

在这六句话里,作用者始终是“张飞”,但在五句话里省略了。四句话的作用对象是“督邮”,三句话里省略了。两句话的对象是“柳条”,都没有省。这里的省与不省遵循简单的“连续性”原理。以柳条为对象的两句话是隔开的,故不能省。前三句话里的“督邮”都可省,是因为有它们前面那句话(督邮未及开口)里的信息可以继承。但第五句话里的“督邮”不能省,因为其间换了对象“柳条”。在以“督邮”为对象的两句纯作用句里,都有效应对象,前一句是“头发”,后一句是“两腿”。作用概念依次用了“揪,扯,缚,攀,打”,并分别配置了效应概念“住,出,住,下,折”。

通过上面的分析,可以看到:概念层次网络理论所引入的作用效应型组合结构的概念,作用对象和效应对象的概念,句类及句类分析的概念等,在整个分析过程起了关键性作用。当然,难点依然存在。首先是新词辨识的困难。“揪住,扯出,缚住,攀下,打折,馆驿,督邮,张飞”都是新词,虽然其中的绝大部分可以通过新词辨识处理[10]予以确认,并不难辨认出“督邮”和“张飞”是人。但仍有硬骨头,例如“攀下”。《近代汉语词典》里有“攀下”一词,释义“牵连”,相当于《现代汉语词典》“攀”字的第三项解释。不过,此新词也并非绝对不可辨识,《现代汉语词典》收录了“攀折”一词,释义为“拉下来折断(花木)”,因此“攀”字在字义库里必然有“拉下来”的义项,当然其独立性的等级最低。而新词辨识程序通常仅考虑独立性强的字义。这就是说,找到这个义项必然要经过一番周折,是在语句合理性分析作出了否定性判断,进入回归处理时[3]才有可能找出这个正确的义项。

其次是省略模糊,分别在两个混合句里,特别是 B 块部分省略并分离的情况比较复杂,已如上述。

第三是隐知识的揭示,在“扯出馆驿”里“扯出”一词所隐含的督邮的挣扎,在最后一句里所隐含的督邮被鞭打的严重程度。这些都属于句类分析高级联想处理的范畴,是我们正在努力的方向。

第二段语料:三句。第一句是接收转移句,第二句是作用句,第三句是过程句。作用句的 A——白孝武,X——抽击了一下,这里的“一下”,就是上文所说的 HX。B——哥哥白孝

文赤裸的胸脯。此句是 A + B + X 的非标准格式。B 语义块的前面有逻辑指示符‘照着’,其构成有作用对象‘哥哥白孝文’和效应对象‘赤裸的胸脯’。这个作用句不仅在格式和每个语义块的构成方面同上一段语料里的‘去督邮两腿上猛打’完全一样,而且句间的知识利用方式也相同,作用者 A 都通过继承而省略了,工具(柳条和刺刷)都由上一句说明。差异仅在于习用逻辑词汇的历时性变化,罗贯中的虚词运用是‘之乎者也哉’时代的格式,如此而已。

第三段语料:三句。都是作用句,第一句是标准格式,第二句是 B + A + X 的非标准格式,第三句是作用效应句,但作用表达采取了承受句的格式并加了前搭配 QX。

14.2 效 应 句

作用和效应是作用效应链的两极,也是语言表达的两个基本参照点。分别以作用和效应为参照点的表达,在形式上表现为通常所说的主动式和被动式。但是,这两个不同参照点对语句的影响不只是形式方面,更重要的是内容方面,即语句的类别和句类格式转移。

作用句的两个表达对象 A 和 B 是缺一不可的,虽然 A 在具体句子里常被省略,但它通常在上文里出现过,或不言而喻,如文【6】中所说属于 1 号联想脉络的作用句。也就是说它必须显含而不能隐含。效应句不同,它是站在效应和对象的参照点上、而不是站在作用和作用者的参照点上进行表述,因此,可以不理睬作用者。这就是作用句和效应句的本质区别。不过,不同语种对这一区别的重视程度似乎有所不同,这一语言现象十分有趣,但作者并未作深入研究,这里只是提出这一现象而已。

由于效应句不理睬作用者,因此效应句的基本格式是两要素句,如下所示:

(1—1) YB + Y (1—2) Y + YB

(2—1) YC + Y (2—2) Y + YC

第一种格式以 YB 为表述对象,第二种格式以 YC 为表述对象。由于两要素句的表述对象和表现都只有一个,所谓标准格式的提法已失去意义,因为这时谁先谁后都不需要对广义对象语义块另行给出标志。但是,不同的排序意味着表述的侧重点不同,汉语的习惯是以前面的语义块为侧重点。其次,独立的效应句常采用 Y 在后的格式。Y 在前的格式很少独立使用,常在混合句类中以复合效应语义块的角色出现。

效应句的第三种格式是:

(3—1) YB + Y + YC

(3—1)的常见变化格式是:

(3—2) YBC + Y

(3—1)和(3—2)相应于 YB 与 YC 的分合,两者分合的前提是:YC 属于 YB。如果这一条件不能满足,那是只可分而不可合的。

从形式上看,效应句的格式(3—1)与作用句完全一样,都是所谓主谓宾结构。格式(1—2)和(2—2)与省略了 A 的作用句一样。要理解和把握这两个句类的差异,并运用这一差异

所提供的信息,就必须对对象和内容的概念以及两者都是句类函数的概念有深切的理解。这个问题已在文【2】中作了系统的阐述,但这里将对此作更具体的说明。

让我们从作用句标准格式里的对象语义块 B 谈起。这个 B,简单地说,是作用对象。但深入地说,它应该包含三个方面的内容:作用对象、效应对象和效应内容,其一般表示式应写成:

$$B = XB + YB + YC$$

这三者如何界定或定义?先来看四个例子:"改进产品质量";"提高学术水平";"打断了李四的腿";"加快中国经济改革的进程"。这四个例子里的"质量,水平,腿,经济改革的进程"是不能省略的,但可以省略"产品,学术,李四,中国"。这个知识对于理解,对于解模糊及纠错处理显然十分重要。那么,如何表达这一知识?

首先,需要引入作用对象和效应对象的概念。定义:作用对象优先于具体的或整体性的事物,效应对象优先于抽象的或局部性的事物。上面例子里的"产品,李四,中国"是具体的,按定义它们是作用对象;"质量,学术,水平,改革,进程"是抽象的,按定义它们是效应对象;"腿"是具体的,然而在这里是局部的,按定义是效应对象;"经济"是具体与抽象的两可概念,在这里也是局部的,因为前面有中国管着,按定义也应该是效应对象。套用定义的结果,两句(第一和第三句)平安无事,两句出了问题。第二句的"学术,水平"和第四句的"经济,改革,进程"也都成了效应对象。把它们统起来叫做效应对象不可以吗?不可以!因为,这样我们就无法解释"水平"在"提高学术水平"里的不可省略。另外,如果将第一句改成"改进乡镇企业产品的质量",那么,"产品"和"质量"也都变成效应对象了。这些矛盾表明:在 B 语义块的总体构成里,仅仅引入作用对象和效应对象还不够,还缺了一点什么,这个欠缺的部分就是效应内容,或者说,把它定义为效应内容,这个定义与内容的原始定义是一致的。再从另一个角度来说,作用于某一对象,可以说是针对这个对象的整体而言的,而具体的效应通常是体现在对象的某一局部,例如张三打了李四,总是打李四躯体的某一或某些部位,而打的结果总会有某种反应,譬如打肿了,打破了皮,流了血等等,这些就是效应内容。所以,在上一段的开头给出这样的公式:$B = XB + YB + YC$。

按照这个公式,上面例句中多个效应对象的矛盾就能够消除,而例句中的可舍及不可舍现象也能给出明确的解释或规定。第一点无须说明,读者可自行验证。下面来说明第二点。

按照语法学的术语,作用对象、效应对象、效应内容三者都是宾语,当然,效应内容也可能构成补语[2]。两者或三者联合起来,可构成复合宾语。当两者联合时,汉语的顺序是:作用在前,效应在后,对象在前,内容在后。当三者联合时,顺序是:作用对象—效应对象—效应内容。这是汉语复合 B 语义块的组合规则。也就是文【2】中所说的良性语义块构成,是汉语特有的一项宝贵财富。具体的语句可以三者俱全,也可以选取其一或其二。其一或其二都有三种选择。一共有七种选择。于是,问题就归结成如何表达这七种选择。

表达的办法就是语义结构方程[7]。

语义结构方程的应用价值比较明显。例如上面例句里的"改进,提高,加快,打断",前三

者的语义结构方程是 4—4—,“打断”的语义结构方程是 4—3—。4—4—必须有效应内容[7],这就决定了相应例句中“质量"水平”和“改革的进程”之不可省略;4—3—规定有效应对象,这就决定了相应例句中“腿”之不可省略。

语义结构方程诚然可以表述七种选择,但是,如果表达方式有多种选择,就不宜精确描述了,语义结构方程采用 i—0—的模糊表示。例如,“打扫”这个词就属于这个情况,其作用对象是场所,其效应对象是废弃物,其效应内容是使环境整洁。它的宾语可以是三者任选其一,例如“打扫房间”,“打扫垃圾”,“打扫卫生”。还可以有两种二选方式“打扫房间的垃圾”和“打扫房间的卫生”。这里的“卫生”是冗余信息,但它只是冗余,并不违例,而语言允许冗余。

在对作用句的 B 块构成作了上面详细说明以后,对效应句的上列格式就比较容易解释和理解了。

从形式上看,可以认为,效应句是作用句的简化,是作用句的一个特殊子类:它省略 A,突出 X 中包含或隐含的 Y,突出 B 中包含或隐含的 YB 和 YC。从表述的角度来说,它是表述参照点的转移:从作用和作用者转移到效应和效应对象。

从效应句的三种格式来说,可以认为格式(1)仅仅是作用句的简化,而格式(2)和(3)才充分体现了效应句的特色,它把效应内容 YC 从 B 中分离出来变成独立的语义块。这些格式充分体现了我们在上面所指出的作用句与效应句的本质区别:效应句是“站在效应和对象的参照点上、而不是站在作用和作用者的参照点上进行表述”。

这里应该说明三点:第一,不能把效应语句格式中的 YB 与 B = XB + YB + YC 公式中的 YB 混为一谈。格式中的 YB 可以相应于公式的 YB,也可以相应于 XB 或 XB + YB。第二,在谈到作用句的作用对象时,通常是泛指 B 而不是特指 XB。第三,XB 与 YB 的定义和划分,只有当它们在 B 语义块中同时出现时才有意义。整体和局部,甚至具体和抽象都是相对的,有比较才有确切的含义。当 B 中只有一个对象时,通常不必严格区分它到底是 XB 还是 YB。所以,在上一节的例句分析中,对“一连打折柳条十数枝”里的“柳条”就笼统的称为 B。但是,如果要深入揭示此句的隐知识,就不但要区分作用对象和效应对象,而且涉及主辅块变换的复杂问题了。

下面就来讨论效应行概念的句类知识。在讨论之前,必须再次强调:效应行的概念本身具有作用与效应的二重性,效应概念可视为作用概念基元的一个子集。这些概念不仅是构成效应型词汇的主要来源,也是构成作用型词汇的主要来源。通过组合,即通过作用型和效应型两种组合结构,二重性消除了。但反映效应行概念的字义则往往依然保留着该概念固有的二重性。因此,单以这些汉字为 E 语义块要素的语句既可以是作用句,也可以是效应句。两可的模糊必然存在。但不同效应概念二重性的强弱有很大不同,以下各节将对此有所说明。

——产生和消除(生灭)

把产生和消除列为具体效应之首不是偶然的,它反映了效应的基本特征,如果把它定位

在 30b,亦无不可。在主体基元概念的二级节点中,它的活跃性仅次于 11,绝大部分基元概念都可以充当它的内容,这一项链式关联性知识记录在概念关联性知识库的 C 栏目中。它的交式关联节点主要有 14 与 11, 35 与 33。

对 31,目前只设置了一组对偶性中层概念 311 和 312。这两个高层概念涵盖了众多的语言词汇。按照层次网络符号的设计思想,这些词汇的个性差异应由底层层次符号来表达。但底层层次符号的设计是一个复杂的系统工程,目前还处于预研阶段。语言词汇的个性主要通过语义结构方程给出近似表达。这里以 311 为例,对这种表达方式予以具体说明。

311 的基本内涵,用汉语词汇来表示就是“发生,产生,生产”。汉语用一个“生”字表达出三者的高层共性 311。但三者的底层特性有明显差异。发生优先于抽象概念及 rw 或 rvw,但不与一般的具体概念 w 和 p 搭配,生产与发生恰恰相反,它优先于具体概念 w,但不与抽象概念搭配。产生介于两者之间。这就是它们在概念关联性方面的区别,用语义结构方程表示如下:

发生	5—7—3—5	0,1	gv
产生	5—1—8—8	0,1	vg
生产	4—1—2—2	1,0	vg

结构方程的表示规则是:第一层数字表示组合结构,第二层表示对象语义块的构成特征,第三层表示语义块(通常是 B,但也可以是 A 或 C)的概念类别优先性,第四层通常表示句式优先性。除了这些各层独立定义的含义之外,还有层间交互的综合信息。例如这里的“产生”,前两层是“5—1”,表示效应型组合概念搭配作用对象,在形式上这是违例表示,因为效应型概念只应该搭配效应对象。但结构方程正是利用这种违例表示来表述“产生”这一概念的二重性。“产生”的这种二重性就决定了它的句式是不确定的,所以它的后两层数字是“8—8”,表示它的作用者和对象都优先于一般事物。

与“产生”相反,“发生”和“生产”都有确定的句式,分别用第四层的 5 和 2 表示,两者都是三个主语义块句式,但“发生”是一个对象两个表现的“1 + 2”句式,而“生产”是两个对象一个表现的“2 + 1”句式。

上列三个双字词的词义仅取决于其中一个字“生”的 1 号义项 v311(这个信息在字义库中),与搭配的汉字“发,产”无关,这个信息由中排的两个数字“0,1”及“1,0”给出,0 号义项是空集的约定。后排的字母表示概念的类别性,在词义库中仅用 1 或 0 表示,这里是它的说明。总之,通过语义结构方程,对上列三个概念的共性和个性作了足够精确的表达。其中的“生产”最为特别,它把效应型概念“生”变成了纯作用型概念。

在二级基元概念节点的设计中,曾有许多问题令人困扰,强交式关联性的表达就是其中之一。例如效应的生与灭,立与破,推动与抑制,关系的支持与反对,排斥和干扰等,都同作用强交式关联,那么,是否把这一重要特征纳入中层表示?纳入显然有利于计算机的思考,但具体实行存在很多矛盾。最后是采用了两种途径加以表示。一是通过概念关联性知识库给出概念层面的表达,二是通过语义结构方程给出词汇层面的表达。“生产”的表达就是一

例,在这里,虚组合的概念起了重要作用,它还大大减轻了字义表达的负担。

这里也顺便说明一下概念二级节点的命名。由于二级节点是高层概念,自然语言往往没有相应的词汇,这时只好选用它涵盖的某些中低层概念作代表,用‘产生与消除’命名 31 就是一例。

——利与害

利与害是效应的基本属性,人类活动和生命活动的基本规律就是趋利避害。当然,利与害是相互依存并可以相互转化的,虽然这种相互转化是所有对偶性概念的共同特征,但 32 的表现尤为突出。有利于此,则往往有害于彼,反之亦然。所以老子说‘福兮祸所伏,祸兮福所倚。老子的这一名言应列为效应节点 32 的第一项高级联想知识。

上面我们看到,效应基元概念 31 具有较强的二重性,32 也是这样。构成作用型 32 的具体词汇,其语义结构方程通常是 4—1 或 4—3—。构成效应型 32 的具体词汇,其语义结构方程通常是 3—2 ,由此可知,利害效应句优先于格式(1—1),即 YB + Y 的形式。

32 本来可以只设置一组对偶性概念 321 和 322。但考虑到‘免除伤害’这个复合概念,即‘保护’的概念特别常用,为它设置了底层概念 3219,它一定是作用型概念。由 321 和 322 构成的动态概念多数属于作用型,所以,利害句以作用句为主。

——显与隐

33 只有一组对偶性中层节点 331 和 332。33 的二重性最弱,或者说其效应色彩最为鲜明。在句类格式的优先性方面,33 与 32 恰好相反,它以效应句为主,而且主要取第三号格式。这就是说,以 33 为映射符号的单字词或由它构成的新词,如果充当了 E 语义块的核心,可默认该语句为效应句,并且必须取第三号格式。这一默认规则也适用于 39,3a 和 3b。在效应句的句类知识中,这是很特殊的一项。

上述默认规则可能引起‘显隐’概念无作用句的误解,如果在默认规则的‘以 33 为映射符号’的前面加上‘仅仅’一词,并对新词的组合方式加以限制,例如限制在后缀式,则上述误解可以消除,但这样又缩小了默认规则的适用范围。由于作者尚未对有关显隐概念的全部词汇作总体研究,宁可暂时模糊一点。但应该指出,如果出现了违反上述默认规则的情况,可通过增加字义库义项的办法予以补救。至于双字词显隐概念的作用型与效应型之分,结构方程或直接表示都可以给出无模糊的标记。例如‘揭露’是作用型概念,而‘暴露’是效应型概念,两者的这一区别,其映射符号 v9331 和 v331 也是表达得一清二楚的。

在概念关联性方面,33 与 23、71 及 72 强关联。对态度,希望,情绪,精神状态,能动性和素质的表述,首先是它们的显或隐。汉语的‘示’字,为所有上述节点所共享,道理就在这里。7y 类反应句[15]与显隐句不会发生混淆,因为 7y 有它特定的内容。但显隐句可能与信息转移句混淆,具体地说,就是由 311 构成的句子可能与省略了 TB 的一般信息转移句[17]发生混淆。因为信息既是天然的 T3C,又可充当 YC。然而,这是典型的两可情况,并不影响句类分析。

——扩展与减缩,立与破,推动与抑制

这三个效应概念放在一起讨论,因为它们是效应概念中的一个集群。集群的提法意味着它们具有某些特殊的共性,主要是以下两方面。第一,它们的二重性最强。第二,它们在语句中所要求的效应内容 YC 具有特定的约束性,34 优先于基本概念的量与范围 j4,故 34 可称量的效应;35 优先于基本概念的质与类 j5,故 35 可称质的效应;36 优先于基本概念的度,由于度不足而需要推动,由于度太过而需要抑制。36 还有一项特殊的约束,就是它的对象优先于过程。这些约束是这个集群所拥有的极为宝贵的句类知识。使效应句的这三个子类具有鲜明的个性。

对 34,仿效 32 的方式,也设置了一组复合对偶概念 345 和 346。它们相应于汉语的"提高"和"降低"。这一组对偶实际上是 34 与 35 的并。

对 351,设置了四个底层概念,它们是:

3518　　建设

3519　　改进

351a　　恢复

351b　　巩固

而 352 尚未发现有此需要,这一现象表明:351 与 352 对偶而不对称,"立"的内涵比"破"丰富。所以,中国古语说"创业难,守业更难"。

本集群的概念基元在构成效应型概念时,往往采取 7 号结构方程。这项知识对于汉语的新词辨识有一定的参考价值。例如汉语的"动"字,是 361 的比较精确的表示。由它组成的"动怒,动气,动情,动人,动容,动听,动心"都是以 7 号方程构成的效应型 361 概念,而由它组成的"打动,带动,调动,发动,鼓动,撼动,轰动,惊动,开动,启动,牵动,煽动,耸动,挑动,掀动,引动"都是以 4 号方程构成的作用型 361 概念。成语"惊天动地,惊心动魄,不动声色,大动干戈,无动于衷"里的"动",其效应色彩多于作用色彩,因为这些成语也可表达为"天惊地动,心惊魄动,声色不动,干戈大动,衷心无动",这正是效应句的特征。

34 和 35 没有对立统一概念的常用语言词汇,但 36 有,就是"调节"。

——连断、通阻

这是两组对偶,层次符号分别是:371 与 372,375 与 376。

这也是一个集群。与上一个集群类似,除了效应概念固有的二重性之外,还各有特定的优先联想脉络。连断优先于关系,通阻优先于转移。换句话说,对关系的作用首先表现为连断,对转移的作用首先表现为通阻。因此,370 又具有作用效应—关系的二重性,374 又具有作用效应—转移的二重性。

371 与 411 强关联,372 与 412 强关联。汉语对 371 的表达主要是"接,连"二字,对 372 主要是"断,隔"二字。双字词"焊接,锻接,嫁接,铆接"打断,截断,切断,熔断,斩断,折断"表现了 370 的上述二重性。

汉语对 374 的表达主要是"通,阻"二字,双字词"通车,通风,通航,通话,通气,通天"阻

碍,阻挡,阻击,阻截,阻拦,阻塞,阻援,阻雨,阻止"表现了 374 的上述二重性。

上面谈到 360 的 YC 优先于过程,这里谈到 374 的 YC 优先于转移。概念层面的这种分工是非常自然的。从联想脉络来看,由于转移与过程密切相关,362 又与 0 4 强关联,所以,(376,362,04)之间应存在较强的交式关联性,而它们与(10,20)之间应存在较强的链式关联性,事实正是如此。从上列阻字的词汇就可以清晰地看到这一点。把握概念层面的这种关联性或联想脉络是自然语言理解唯一可行的起点。

当然,这仅仅是一个起点。就上述的局部联想脉络来说,它尚未涉及过程或转移的具体特征,而深入的理解离不开这些具体特征。以 376 为例,高层表达用阻字,但对底层的 376,它还引入了"隔,挡,遮"等字,于是有"隔声,隔热,挡风,挡雨,遮阳,遮光"的表达,376 当然不能反映这些概念的个性特征,但作为一级近似,仍然准确地表达了它们的共性:阻止转移物到达它的终点 TB2。

——选存弃

在效应概念的二级节点中,"存弃"与"合分""得失"这三组对偶概念构成一个集群,作为集群的共同特性是:第一,它们的二重性最弱,换句话说,单纯由这些概念构成的语句通常是效应句。第二,它们的联想脉络都十分复杂。第三,它们与人类追求活动 b0 和关系 40 强交式关联。这个集群特性恰好与 34 到 36 的集群形成鲜明的对照。前两点不言自明,第三点表现为后者与专业活动强关联。这两类高级智能活动的实际表现虽然很难分开,但语境的分野是清晰的[3]。追求活动是战略性的,侧重精神和横向联系,侧重抽象和高层概念;专业活动是战术性的,侧重物质和纵向联系,侧重具体和底层概念。

"存弃"是生物进化的基本规律,所谓优胜劣败,新陈代谢,就是存与弃。生命过程每时每刻都发生存与弃。记忆与忘却 68 是存与弃的意识反应,决策 842 是存与弃的思维表现。追求活动的四个二级节点,改革 b1 与继承 b2 是存弃问题,竞争 b3 与协同 b4 是合分问题。甚至可以说,《红楼梦》的主旨是探索"存弃"的真谛,而《三国演义》的主旨是阐述历史发展的"合分"势态。两位作者都在第一回中把这一点说得很清楚,只不过曹雪芹的用词不像施耐庵那么通俗,因而引起了较多的误解。

存与弃有显含的对立统一概念 380"选"。但存与弃并不对称,既有共同的,又有各自的联想脉络。

汉语的"存"字,有"存亡,存在,保存"三项基本含义。这是汉语"字义基元化"的又一生动字例。存亡的概念已安置在 141 和 142,存在是基本逻辑概念 jlvg 115。但汉语将"生存""存在"和"保存"这三个概念用一个存字串联起来,是很不寻常的。它表现了过程的生存与效应"存"的内在联系;它把哲学上十分深奥的"存在"概念与过程的存亡和效应的存弃概念联系起来,这显然是务实而聪明的思考方式。生存,存在和保存是密切相关的,汉语用"存"字清晰地展示了这一联想脉络。

"存弃"这一组对偶概念还与效应的另一组对偶概念得失 3a 密切关联,将"取存弃"构成一组效应型三重对偶概念,也曾纳入效应二级概念节点的设计考虑。因为,生命过程的新陈

代谢是“取存弃”的循环运作，当然，这个运作过程还必须有转换 24b 的参预。转移的传输过程更是经常伴随着“取存弃”的运作，不过，在名称上，转移以入代取，以出代弃而已。但是，这一组三重对偶概念的联想脉络过于专业化，所以，放弃了另设的考虑，而分别用“存弃”和“取舍”（即 3a 的获得和付出）来代替。

汉语对“弃”字的运用与“存”字有异曲同工之妙。弃沿着它的联想脉络可形成各种各样的组合概念：优胜劣败的弃又名淘汰，自暴自弃的弃就是淘汰的意思；新陈代谢或新旧更迭的弃叫废弃；关系二级基元概念“用与弃”的弃叫舍弃。作为“选与弃”或“取与舍”这两组对偶性概念一极的“弃”，叫放弃或抛弃，也就是弃暗投明，弃旧图新的弃。这个“弃”是 382 的本义。而废弃，淘汰以及其它的“弃”，则是 382 与其它基元或基本概念的复合。

上面说明了存与弃各自独立的联想脉络，这是两者对偶而不对称的表现之一。其另一表现是 380 选与 382 弃对偶，但不与 381 存对偶。

由单纯 38 概念构成的语句通常是效应句，并主要采用格式（3）。在概念层面不能对 YB 和 YC 给出约束，这方面的信息只能由具体词汇的语义结构方程提供。但概念“选拔”或“选举”c380 是一个例外，它不仅构成作用句，而且往往是作用效应句。

——合分、聚散

这是两组对偶。层次符号分别是：391 和 392，395 和 396。

39 是效应概念中唯一的具有自身句类格式的子类。这一特殊性来于它的特殊的交织性。

合与分是关系的基元概念，并列为关系二级概念之首。同时它又是效应的基元概念。这就是说，合与分这一组对偶概念需要从关系和效应两个不同的角度进行表述，这是这一类概念的共同特征。但关系的合分与效应的合分还各有自身的个性：关系的合分双方应具有一定的独立性，即双方分离时，仍可独立存在。这可视为关系合分的必要条件。但效应的合分不需要这一必要条件。例如“砍刨切削”等概念就是效应的分，而不是关系的分。砍了头，人就死了。人与头，关系上是不可分的，但效应上是可分的。这就是说，效应的合分比关系的合分内涵更广。当然，合分不限于双方，但这不影响上面的论点。

效应合分表述的标准句类格式有：

YB9 + Y9 + YB

YB + Y9 + YB9

格式中的 YB 和 YB9 分别代表合分对象的整体和局部。YB9 相应于一般效应句句类格式中的 YC，这样定义保证了语义块定义的一致性，便于语言逻辑概念切分子类[11]的统一设计，又符合自然性法则，因为局部是整体的当然内容。

从形式上看，似乎这个句类格式里的 YB 和 YB9 具有互换性。但这是假像。两种格式的选择取决于 Y9 的特性，如果 Y9 优先于对象，则取第一种格式，如果 Y9 优先于内容，则取第二种格式。这一优先性信息可由语义结构方程或字义库明确无误地得到。例如下面的例句就显然不能互换。

张三加入了李四公司　　张三退出了李四公司

美国分为 40 个州　　前苏联分成 15 个独立国家

这里的“加入”和“退出”优先于对象，而“分”优先于内容。这就决定了它们各自的句类格式。

像所有的效应概念一样，合分也有作用型和效应型两类。以上所述是效应型的合分，作用型的合分概念就构成作用句。例如：

沙俄吞并了一大批弱小的邻国　　欧洲殖民帝国曾瓜分非州大陆

这里的吞并和瓜分是作用型的合分，要求按作用句的句类格式构成语句。例句中的沙俄和欧洲殖民帝国是作用者，邻国和非州大陆是作用对象。这两个作用合分句可变换成各种非标准格式，但上面的效应合分句则不具有这种变换性。这一表述方式上的区别，并不限于合分句，是一般作用句和效应句的区别特征。

通过对合分概念的效应型和作用型划分，我们窥见了合分概念内涵的一角，而且就这一角来说，也不过是略见端倪。再深入一步，就需要引入加减乘除的概念，前者是加减式合分，后者是乘除式合分。乘除概念的引入又将带来一系列新概念，但这就超出了语言表述的范畴，所以，语言必须适可而止。具体的做法就是仅在概念知识库中[6]建立起 39 与 34 的交式关联表示，与 j3 的交链式关联表示。

合分概念的联想脉络，汉语通过“合”“分”两字的组合词汇给出了相当完备的指示信息。表示与 41 关联的有结合，分离；表示与 435，b4 和 443 关联的有联合，合作，分工，分担和配合；表示与 81 关联的有综合，分析，分辨，分类和区分。这些概念里含有一个共同的对偶性核心，就是从效应角度定义的 391 和 392，从关系角度定义的 411 和 412。

当然，39 的联想脉络不可能通过“合分”两字的组合词汇而包揽无遗，39 与 54 的强关联性就没有得到体现。结构与结构物的核心概念也是合分，对结构的理解，就要从它的合分构成着手。

39 的另一组对偶概念聚散，是合分的补充。聚与合、分与散并无必然联系，聚而不合，散而不分是常见的现象。但聚散终究是合分的重要（虽然它既非必要，也不充分）条件，所以安排在 394。从概念关联性来说，聚散与转移强关联，其句类格式的一般形式是：

YB94 + Y94 + (TB2)

格式中的 TB2 代表转移的终点[17]，由于 TB2 的存在，聚散句也可视为是混合句类[21]，但括号表示它可有可无，因此，这是典型的两可情况。还应该说明一点，转移句中的 TBk 仅优先于具体的空间概念（含 pc 类概念），但不包括广义空间，而聚散句包括后者。

——获得与付出

这一组对偶概念的命名用“得与失”或“取与舍”皆可，但“获得与付出”要更确切一些。层次符号是 3a1 和 3a2。

在概念关联性方面，3a 与关系的拥有和失去 46，转移的入出 200 强交式关联，虽然三者之间有某种因果表现，但在概念关联性知识库中仍可纳入“狭义”交式栏目[6]。

由于在 38 中已对此集群特性作了说明，这里对 3a 已不必多加解释，而只指出一点，就

是由 390 和 3a 构成的作用型词汇在基元概念的二级节点中是最多的,这从一个侧面支持了前面的说法,即本集群概念是人类追求活动的核心。

效应二级概念的最后一个积累与消耗 3b 是从 38 分离出来的,也可以安排在 384。只是由于它与上述集群特性差距较大而另行设置。这表现在两方面,一是它的二重性较强,二是它与专业活动的联系更为紧密。

本节已行文过长,最后引用一段语料,它以效应句为主,读者可用它自行印证本节的有关论述。

十一届三中全会以来,在邓小平同志建设有中国特色社会主义理论的指导下,我们党和人民锐意改革,努力奋斗,整个国家焕发出了勃勃生机,中华大地发生了历史性的伟大变化。社会生产力获得新的解放,安定团结的政治局面不断巩固。十一亿人民的温饱问题基本解决,正在向小康迈进。我国经济建设上了一个大台阶,人民生活上了一个大台阶,综合国力上了一个大台阶。在世界风云急剧变幻的情况下,中国的社会主义制度经受住严峻的考验,显示了强大的生命力。

14.3 作用效应句

作用效应句是混合句类中最重要的一种类型,从句类格式来看,它只是作用句的自然扩展,所以从【21】中提出来放在本篇讨论。

作用效应句的标准格式是:

$$A + X + B + YC, \quad YC = E + EC$$

从上一节对作用句 B 语义块的详细说明可知,在形式上,如果说效应句是作用句的简化,那么,作用效应句就是作用句的扩展,两者都是将 B 语义块中的 YC 变为独立的语义块 YC,但这两个变换或两个 YC 有本质的区别:效应句可保留‘2 + 1’句式,或转化成‘1 + 2’句式,或退化成‘1 + 1’句式,而作用效应句则一定扩展成‘2 + 2’句式。这就是说,效应句的 YC 是文【2】中所阐述的 C 语义块融合性表现,而作用效应句的 YC 是 C 语义块的扩展性表现。

作用句的这种扩展性当然与 X 要素的特性有关,但与 B 语义块的构成不同,它不是唯一地决定于 X 要素的特性,也同语言的总体表述方式和效应表现的具体特征有关。按照作用效应链的观点,X 要素在不同程度上都应该具有这种扩展性,但不同概念和词汇的这一特性的强弱之差很大。显然,具有强扩展性的概念和词汇是自然语言理解关注的重点对象,对这一类概念和词汇必须给出明确的标记,因为它是句类分析极为宝贵的信息。

具体的标记方法有两个:一是采用语义结构方程的 4—4—k—4 表示方式,这用于作用型组合概念,在汉语就是用于双字词。二是采用双作用结构符 # #,在汉语用于字义表示。

汉语的‘迫使,勒令,鞭策,要求……’等词汇,‘使,令,叫’等字的义项之一就属于强扩展性 X 概念,由它们必须构成作用效应句,后者常用于省略语义块 A 的作用效应句。传统语言学的所谓兼语句就是典型的作用效应句。这个句类按西式语法的规范,显得有点‘反常’,

但从句类分析来看,它反而是个性鲜明、易于处理和理解的句类之一。这一句类知识的运用,可取得模糊消解及纠错处理的良好效果。

由一般 X 概念所构成的作用效应句,汉语通常对 YC 语义块给出标志符 l0300,或同时对语义块 A 和 YC 给出联合标志符 l2100 和 l2300。充当前一种标志符的典型汉字是‘得”,充当后一种标志符的典型搭配是‘一……就”。当然,由于“得,一”二字的义项甚多,用法十分灵活,即使是无模糊文本,对这类作用效应句的判定和理解,也不像强扩展性双字词所构成的作用效应句那么便利。

结 束 语

本文可视为文【2】的重要补充,它进一步阐述了作用与效应,对象与内容,E、B 语义块构成这三组重要的概念。

关于文【7】中所阐述的概念关联性知识库的具体内涵,这里也作了示范性说明。

1995 年 6 月

作用承受句和作用反应句的句类知识

引　　言

本篇是【14】的姊妹篇，讨论作用句的两个子类，承受句和反应句各占一节。这两个子类概念层面的句类知识比较丰富，所以着重这两方面的阐述。

15.1　作用承受句

承受句的信息标志比较简明，就是基元概念层次符号 01。下列六个二级节点都提供明确无误的承受句信息：01，6m01，901，c01，5201，5301。这里“二级节点”的提法是以对主体基元概念的挂靠层为准，不计本体层。本体与挂靠的层次总是分开计算，这是由两者的定义所决定的。

15.1.1　承受句的子类及其句类格式

承受句可分为四个子类。第一是主动的承担，第二是被动的承受（受到，遭受），第三是一般承受，最后是与反应交织的感受。这四类承受句的标准格式可统一写成下面的形式：

X1B(k) + X1k + XC(k)

这个式子的第一项表示承受者，第二项是承受句的 E 语义块，第三项是承受的内容。这个表示式试图表明承受句是 k 的函数（k 的层次符号在下文说明），但不符合文【2】给出的规范。从句类分析来看，承受概念的子类首先应分为主动性和非主动性两类，因此，上面的函数表示形式应改写成下面的主被动两种标准格式：

X1A + X11 + XBC

X1B + X12 + XAC

主动形式中用 X1A 表达主动承担者，因为主动承担者就是作用者。被动承受者用 X1B 表示。主动承担的内容用 XBC 表示，以区别于被动承受的内容 XAC。XBC 的核心概念是 rc01 和 gv900，而 XAC 的核心概念是 vg00。这一重要知识在概念知识库中通过 C 函数予以表达。这就是说，主动承受句所表述的是人类的高级智能活动，即由复合基元概念的 9，a，b，c 行所描述的活动。

主动和被动承受是承受的两个极端，多数情况的承受介于两者之间，称之为一般承受句，标准句类格式是：

X1B + X10 + XBC

表示主动承受的层次符号是 011,表示被动承受的层次符号是 012,表示一般承受的层次符号是 01 或 010。至于与反应交织的感受则用(01,02)的形式来表示。这就是承受句一般表示式中 k 的定义。这种表示方式显然是吸收了汉语有关双字词的表达思想,在层次网络符号的设计过程中,受益于此良多。

显然,承受句对 JK2 的 C 块素和主动承受句对 X1A 都有很强的约束,这是承受句类所提供的宝贵信息,对解模糊及纠错处理十分有益。

承受句的常用非标准格式有:

XBC + X1A + X11

X1A + XBC + X11

这里只给出了主动式承受句的非常用标准格式,一般承受句也如此。但被动式承受句通常仅采用标准格式,理由在下面说明。

15.1.2 承受句的特点

本节将从承受句语义块要素之间的强约束性和承受这个概念与其他作用概念的交织性这两个角度,讨论承受句的特点。

特点 1:关于非标准格式主动承受句指示符的省略。

在讨论句类格式时,我们曾指出[2]:当语句采用非标准格式时,一般需要对一些语义块给出标志,只有在有关语义块的个性特征十分突出时才容许违例。如"红旗漫卷西风"之类的倒装句。承受句的语义块 JK1 和 JK2 一般满足违例的条件,从理论上说,承受句的非标准格式可以不另加语义块标志。当然,语言的实际表现绝不会与理论的预想完全一致,但这并不影响这一概念层面的知识对承受句句类分析的指导作用。例如下面的例句:

这项工作(由)张先生负责　　XBC + X1B + X10

张先生对这项工作负责　　X1B + XBC + X10

这是一般承受句的两种非标准格式。第一种非标准格式的对象指示符 l0101"由"可以省略。在省略情况下,语义块感知处理[11]就有赖于对承受句上述句类知识的运用,如果仅仅运用 lv 准则,将产生把"这项工作张先生"切成一个语义块的错误。

特点 2:关于主动承受句与作用句的转换。

从概念的交式关联性来看,主动承担与作用的交织性很强,X1A 对 XBC 的主动承担可视为 X1A 对 XBCB 的作用,即 X1A 相当于作用者 A,XBCB 相当于作用对象 XB。这意味着作用句和主动承受句这两种句类格式可以相互转换,语言的实际表现正是如此。文【2】中的例句"中科院声学所和自动化所联合承担了汉语人机对话系统的研制任务"就表现了这一转换。

作用句向主动承受句的转换规则是:

作用句的 JK1(A)—→承受句的 JK1(X1A)

作用句的 JK2(B)—→承受句的 JK2(XBC)

作用句的 E(X)—→承受句的 EH

转换后,承受句的 E = EQ + EH,EQ 一定由概念节点 ϕ01 表示。但应强调指出,EH 经常与 EQ 分离,分离出去的 EH 被 JK2 吸收,出现在承受句的句尾。如上面的例句所示。这个转换是有条件的,就是 X 必须属于智能活动。这些都是主动承受句极有价值的句类知识。

主动承受句向作用句的反转换规则不过是上列转换方向的逆。这时,要取消承受句的 EQ。对句类转换实行反转换是句类分析的重要内容,因为,反转换以后得到的句类格式更便于句类检验,即语义块要素之间关联性的检验。

特点 3:被动承受句与被动式作用句的相互替代。

被动式作用句是作用句的非标准格式之一,句类格式是:

B + A + X(汉语)　　B + X + A(西语)

它的 A 必须带有逻辑指示符。这种格式的作用句在西语里最为常用,但汉语比较少用。原因之一就是汉语用一般被动承受句来替代被动式作用句。这种替换从理论上说,应该是无条件的。因为一般被动承受句就是被动式作用句,两者都以作用的承受者为参照点进行表述。传统语言学对此早有明确的认识,这里不过是从句类知识的角度加以说明而已。下面举几个作用句与一般被动承受句相互转换的例句。

张三打了李四。	李四挨了张三的打。
必须惩治贪官污吏。	贪官污吏必须受到惩治。
沙暴经常袭击鄂尔多斯草原。	鄂尔多斯草原经常遭到沙暴的袭击。
法西斯德国杀害了几百万无辜的犹太人。	几百万无辜的犹太人惨遭法西斯德国的杀害。
海峡两岸的中国人都尊重孙中山先生。	孙中山先生受到海峡两岸中国人的尊重。

例句中的最后一句是主动式反应句,下文还会提到。

特点 4:感受与反应的交织。

作用与效应,作用与反应,都是对作用效应链的表述方式。但概念层次网络理论约定:效应这个概念更广泛一些。所以,概念的组合结构之一命名为效应。效应的具体内涵由主体基元概念 ϕ3 的二级概念节点来范定,而反应的具体内涵则由主体基元概念 ϕ0 的二级概念节点 02 和复合基元概念 ϕ7 的二级概念节点 71 来范定。

承受是作用与反应之间必然存在的过渡,所以它同作用和反应都形成交织,上述特点 2 就是承受与作用的交织表现,从上面的例句我们看到,汉语用‘受到,遭到,遭受’等词汇表达这一交织性。特点 4 是承受与反应的交织表现,汉语用‘感到,觉得,感受’等词汇来表达这一交织性。

15.2 反 应 句

对反应的表述首先要涉及到反应者和引起反应的引发者,反应者和引发者往往形成相互作用,所以,在反应句里,反应者和引发者都可充当施事和受事的双重角色。

引发者通过一定的表现引起反应者的反应,反应者除直接反应外,还会有后续的反应表现。所以,反应句的语句要素通常不仅是反应者、反应和引发者三项,还要加上引发反应和后续反应两项。对这五项要素采用下面的符号予以标记:

反应	X2
反应者	X2B 或 X2A
引发者	XA 或 XB
引发表现	XC
反应者的后续表现	X2C

反应者的 A 和 B 两种标记用于表现反应者主动权的差异。主动权在反应者记为 X2A,否则记为 X2B。引发者的符号与反应者相反。

在构成反应句时,通常并不是每项要素都形成一个语义块,而是将某些要素合并。合并方式有两种:一是将引发者及其表现合并,二是将反应及其后续表现合并。当 XA 与 XC 合并时,简记为 XAC;当 X2 与 X2C 合并时,简记为 X2C。

如果反应句仅涉及上述两种表现之一,有下列三种基本格式:

1 X2A + X21 + XBC

2 X2B + X22 + XAC

3 X2B + X2C

如果两种表现同时出现,则构成下面的复合反应句:

4 XAC + X2B + X2 + X2C

这是作用效应句的特殊形式之一,在文【2】中曾引用过这样的句例。

对反应句句类知识的把握要比前面已介绍过的作用句、效应句、作用效应句和作用承受句要困难得多。难点在于反应句的下列两项特性:一是反应概念的内涵子类与句类格式不是简单的一一对应;二是反应句的每一种句类格式都有两种句式,上列基本格式 1 和 2 在形式上是‘2 + 1’或‘1 + 2’句式,但它们也可扩展为‘2 + 2’句式。这就是说,对于某些反应概念子类,既存在句类格式的两可,又存在句式的两可。当然,双重两可可转换成单重的多个选择,这就是把上列三种基本格式扩大为六种,但在问题陈述时,不如用双重两可的说法简明。格式和句式的两可性各句类都不同程度的存在,但以反应句最为严重。本节将侧重对两可性处理策略的讨论。

上面关于反应句句类格式的论述似乎假定了引发者来自外部,但实际上并未作此假定。上列句类格式同样适用于反应的引发者是反应者自身的情况。格式 3 不考虑引发因素,当

然适用。对格式 2 来说,这只是 XA = X2B 的特殊情况。

引发因素的内外之分虽然不影响语句标准格式的划分,但对格式的选择却有重大影响,这包括非标准格式的运用。这就是说,引发因素的内外之分是语句合理性和完备性判断的重要信息。因此,反应概念,主要是心理概念的子类划分,必须把引发因素的内外之分作为首要的判据之一。在下面可以看到具体运用这一判据的实例。

反应概念作为效应概念的一部分,当然也具有效应概念的二重性:第一种格式里的 X21 是作用型概念,第二种格式里的 X22 是效应型概念。二重性有强弱之分,对强二重性概念将默认为作用型概念,对弱二重性概念将默认为效应型概念,例外和两可的情况必然存在,可在词汇层面给出相应标记。这是处理二重性概念的一般原则,在上一篇论文【14】里有详细说明,这一原则当然也适用于反应概念。

反应句的特征信息集中在两个来源,一是以 02 和以 02 为挂靠层的复合基元概念,二是 71。02 主要是指生命体的反应,汉语的‘反应’一词有脉络性模糊,它包括化学反应、热核反应之类的反应,但这些反应包含在效应的概念里。

生命体的反应明显的分为三个层次:生理及本能层次,心理层次和理智层次。这三个层次的反应概念分别用层次符号 6m02,71,902 及 c02 表示。

理智反应的主动权在反应者,所以,以 902 和 c02 为 X2 核心的反应句一定是第一种格式。心理反应按主动权的强弱分为三个子类:711 半主动,712 强主动,713 弱主动。生理反应属于弱主动,而本能反应属于强主动。这样,反应句的类型,可粗分为下面三类:

类型 1　　7110,7111,712,902,c02

类型 2　　713

类型 3　　7112,713,6m02,m = 1,2

类型 1 将命名为作用型反应,类型 2 和 3 将命名为效应型反应。即类型 1 将构成主动反应句,类型 3 将构成被动反应句,类型 2 将构成一般反应句。但这种概念层面的知识一定存在违例的个性表现,即例外和两可情况。对这种违例或两可情况,应分别在字义和词义表示中采取下面的具体处理方案。

字义表示:对作用型反应概念的效应型义项加挂靠层 30,对效应型反应概念的作用型义项加挂靠层 00,对两可性义项加挂靠层 02,不加挂靠层则取默认属性。

词义表示:对作用型反应概念用 4 号语义结构方程,对效应型反应概念用 5 号方程。反应概念的语义结构方程绝大多数是虚运用。即用 4 号和 5 号方程分别充当作用型和效应型反应概念的区别标志,这是对例外的处理。除了例外,还有两可情况,作用与效应的两可也就是前面所说的句类格式的两可,将虚用 a 号语义结构方程予以标记。而对句式的两可,则虚用 b 号语义结构方程予以标记。对于双重两可,则虚用 c 号语义结构方程予以标记。

作用型反应句作为一般作用句的特殊子类,除了上述 A 语义块的角色(兼充施事与受事)特征和 E 语义块的概念来源不同之外,还有两点更有实用价值的差异。一是 B 语义块的构成不同,作用句 B 语义块的一般构成是 XB + YB + YC,它有时需要细分作用对象和效应

对象[14],而对作用型反应句的 B 语义块则不需要这种细分,只需要粗分为对象及其表现两部分。二是一般作用句扩展为‘2 + 2’句式时,一定是作用效应句,而作用型反应句则可以是一般的‘2 + 2’句式。下面要谈到的 7121 子类就属于这一情况。

下面就来分别说明这两大类反应句的句类知识。

——作用型反应句的句类知识

作用型反应句即由格式 1 所表达的句子。按其内涵可分为两个子类:即 711,902 及 c02 子类,712 子类。如前文所指出的,格式 1 可取‘2 + 2’、‘2 + 1’、‘1 + 2’三种句式。‘2 + 2’句式就是 XBC 分成 XB 和 XC 两个语义块。两个内涵子类可分别命名为态度子类和意愿子类。优先‘2 + 2’句式的只是意愿子类中的 7121 子类,例如‘希望,祝愿’等概念。意愿子类的另一子类 7122 和态度子类则优先‘2 + 1’或‘2 + 1’句式。这是作用型反应句在语义层面的一项宝贵知识。

当然,概念层面的优先规则永远有例外情况,对例外就需要在词汇层面予以个别说明,办法仍然是通过语义结构方程。例如诅咒这个概念,其层次符号是 j821/vg7121(恶意的希望),但它不强求‘2 + 2’句式,具有句式两可的特性,其 XBC 优先于类别概念 p 和 pe,这些个性知识由语义结构方程 b—k—1 给出。

对 7121 概念构成的‘2 + 2’句式,这里还应该说明一个特殊情况,就是对自身的希望,这时 XBCB = X2A,因而 XBCB 可以省略,这一知识目前尚未给出表示的手段。

作用型反应句具有作用句的各种非标准格式,汉语最常用的是上一节所说明的被动承受句形式:

X1B + X10 + XBC

上面的例句‘海峡两岸的中国人都尊重孙中山先生’就是由 7110 概念构成的反应句,它可用被动承受句予以表达,变为‘孙中山先生受到两岸中国人的尊重’。当然,作用型反应句也可以采用以下两种常规的非标准格式:

X2A + XBC + X2

XBC + X2A + X

这时,上面的例句分别变为‘海峡两岸的中国人对孙中山先生都十分尊重"孙中山先生为两岸中国人所尊重’。第一句引入了逻辑指示符‘对’,第二句引入了逻辑指示符‘为……所’。

71 各子类的概念都非常丰富,其中 7110 和 713 尤为丰富。对这些概念个性的表达将主要采用组合近似的方式,例如:

重视 (vg7110,j721)
轻视 (vg7110,j722)
歧视 (vg7110,v422)
敌视 (vg7110,v432)
唾弃 (vg7111,v382)
侮辱 vg7115 # v3229

奉承　(j832/vg7110,l6,rc441)

妒忌　(j842/vg7114,l6,r930a1)

这些映射符号安置在两字之一的义项中,对另一字取空义项。整个双字词的语义表示则通过语义结构方程,以便提供更多的信息。式中的 l6 是 l16 的省写,这是逻辑概念组合的统一约定。这里应该说明一点,就是 l6 后面的内容是指对方,而不是指自己,奉承里的权势 rc441 是指对方的权势,妒忌里的成功 r930a1 是指对方的成功,这是定义。这个信息也来于映射符号 711y,因为当 y≠2,3 时,它都是指对待他人的态度。这种隐式表示方式显然不利于程序的运作,具体的改进方案,则寄望于来者。

从上面的例子可以看出,汉语'视'字的义项之一是 v7110 相当精确的描写。重视,轻视,歧视,敌视以及藐视,傲视,仇视,小视等词汇里的'视'都是取 v7110 的义项,组合结构都是采用 1 号语义结构方程。但实际的词义表示将视语用的需要而采用不同的结构方程。而且如上文所说,对于 7110 概念,绝大多数采用虚组合方式。这就是说,在查寻上列词汇的语义时,实际上并不检索'视'字。但 v7110 的义项仍应列为'视'字的义项之首,虽然它的高层层次符号多达四级,但组合性极强,且是构造新词的活跃语素。

这里顺便介绍一下 711 的中层设计。

对 7110 选取下列两组对偶:

71101　尊重

71102　鄙视

71105　关注

71106　漠视

对 7111 选取下列两组对偶:

71111　宽恕

71112　苛求

71115　爱护

71116　虐待

对 7112 选取下列两组对偶:

71121　谦虚

71122　傲慢

71125　满足

71126　计较

上面对态度子类中的 7110 概念作了较详细的说明。态度子类中的另一类概念由 902 和 c02 表述。两者的区别在于:前者心理因素起主导作用,后者理智因素起主导作用。对 902 目前只安排了两组中层对偶概念:

9021　同意

9022　反对

9025　相信

9026　怀疑

这一类概念对句式无优先取向，“2 + 1”和“2 + 2”两可。由于作用型反应概念中的大多数都有优先取向，902 的“无”反而是值得注意的特色。

作用型反应概念，无论是态度子类或意愿子类都同关系 40，特别是与二级概念节点 43 强关联。而且两者互为因果，态度和意愿影响到关系，关系又影响到态度和意愿。这就是说，两者互为因果函数。对这种关联性，在概念关联性知识库中以函数 M 表示。

711 与 720，722 强关联。前者是后者的 Rt 函数，后者是前者的 Pr 函数。

意愿子类还与追求活动 b0 强交式关联，后者是前者在理性水平上的升级。

——效应型反应句的句类知识

由效应型反应句的定义可知，它有三个内涵子类：由 7112 所定义的态度子类，由 713 所定义的情感子类，由 6m02 所定义的生理反应子类。其句式可以是“1 + 1”、“2 + 1”、“1 + 2”或“2 + 2”句式。格式 3 可取前三种，但不会取第四种。格式 2 可取后三种，但不会取第一种。

由前面给出的反应概念的句类知识表格可知，态度子类 7112 和生理反应子类都优先于句类格式 3，情感子类则属于两可。

那么，对情感子类 713 能否像前面对意愿子类 712 那样，通过第四级层次符号的约定消除它的格式两可性？

情感子类的概念内涵远比意愿子类复杂。汉语过去有七情之说，七情是“喜怒哀惧爱恶欲”，这个概括当然并不完备，但可供参考。下面对七情作一简单分析。“欲”应纳入 712，不来讨论。喜哀惧的引发因素可内可外，词汇表达当然不反应引发因素的内外区别，但语句表达却与引发因素的内外之分有密切关联。内引发的喜哀语句表达不存在句类格式中的 XA，或者说，XA 与 X2B 合而为一，引发表现来于反应者自身。这就是说，喜哀惧的表达有多种格式可供选择。三者之中的惧又有所不同，不论引发因素的内外，它都必须有第二对象，但这个对象可能不是引发者。一个人犯了错误，害怕受到责备或惩罚就属于这个情况。喜哀的表达，第二对象可有可无，喜的表达尤为多种多样，“1 + 1”、“2 + 1”、“1 + 2”和“2 + 2”的句式都可能选用，也就是说，它的句类格式和句式都不受约束。因此，它是句类分析难于处理的概念之一。

相对而言，爱恶怒的特性就不像喜和惧那么复杂，它们必来于外在的引发，并且必有第二对象。其中怒的表达很少采用标准格式，这是一项很特殊的概念层面知识。

引发因素仅在自身的情感概念不在七情之内，它有愧、悔、咎、憾等，其语句表达可以有第二对象，但第二对象可以是参照对象 Re。这也是一项很特殊的概念个性。

基于上述分析和本节开头所说的反应概念子类划分的首要原则，情感概念应分为三个子类：引发因素可内可外子类，外引发子类和内引发子类。这三个子类将分别安置在 713y 的三个子区，y 值的分配如下：

内外引发子类　　y = 0　3

外引发子类　　y = 4　7

内引发子类　　y = 8　b

扩大七情之说，可把喜忧惧归于第一子类，爱恶怒恨归于第二子类，憾、愧、悔、咎、窘等归于第三子类。y 值具体设置见文【6】6.2.3 节的 φ7 概念节点表。

由该节点表可知，情感概念中的“喜”7131 和“忧”7132，“爱”7135 和“恶”7136 具有对偶性。它们在数字形式上似乎符合对偶性的定义，但实际上未能表达这一重要特征，因为心理反应的高层表示定义为四层。这项知识只能在概念知识库中予以表达。

上面定义的各 713y 都是情感概念的子集，每一子集都有丰富的内涵，对它们的个性表达属于中层表示的范畴。对各子集的汉字命名仅仅是一个汉语标志，上面的命名并未沿用七情之说，改用了现代汉语的词汇。

情感概念的联想知识主要是它的 Pr 函数，其中的某些函数项优先类别符号 r，例如，7131 的 Pr 函数有 r930a1，r321，rb301；7132 的 Pr 函数有 r322 和 rb302；7138 的 Pr 函数有 r930a2。

通过对情感概念的子类划分，部分消除了句类格式选择的两可性，外引发子类优先于格式 2，内引发子类优先于格式 3，但内外引发子类仍保留句类格式的两可性。

结　束　语

乔姆斯基曾说过：自然语言不是一种“well-defined”的东西，而是一种“ill-defined”的东西。我们在文【2】中提出的句类分析模式试图表明，这个说法在总体上是错误的，因为自然语言语句物理表示式的存在及其可穷尽性就是“well-defined”的根本保证。当然，自然语言也有其“ill-defined”的一面，如上一篇【14】所述效应概念与作用概念的两可性，这里所说的反应句的格式及句式两可性，复合语义块的非良性构成等。但通过本文的分析可以看到，即使是“ill-defined”这一面，也并非是无句类知识可以利用，理解处理仍可以大有作为。我们正在研制的“语义块感知处理”和“初级句类分析处理”软件模块，将对这个问题进行实质性的探索。

反应句是基本句类中最复杂的子类，但通过本文的分析可以看到，它在概念层面仍然拥有丰富的确定性知识支持句类分析。这大大坚定了 HNC 理论关于“运用语句物理表示式即可取得相当于大脑语句感知过程的预期和判断能力”的信念，从而写了上面的话。

1995 年 6 月

转移句的句类知识

引　　言

转移句的最大特点是:各语义块个性鲜明,每一子类都有自己独立的句类格式。在七个基本句类中,转移句的解模糊和纠错潜力最大。主体基元概念语义网络的低层设计,目前惟有转移跨出了对偶、对比和包含性中层概念的范围,进行了较为完备的底层设计,显得比较完整,主要是来于先行启动这一潜力的考虑。

17.1　转移句的子类及其标准格式

转移句以外的句类,对象和内容在直观上的分野都比较清晰,惟有转移句与众不同。文【2】对此有专门阐述。转移内容 TC 定义为被转移的物或信息,转移对象定义了四个,主对象是转移内容的接收者,记为 TB,两个对偶性辅对象——"起点和终点",还有路径或中转点分别记为 TB1、TB2 和 TB3,转移的作用者(发动者)记为 TA。

转移句各子类的标准格式如下:

一般转移句	TA + T + TB + TC
物转移句	TA + T2 + TB + T2C
信息转移句	TA + T3 + TB + T3C
一般接收句	TB + T1 + T1C
针对性接收句	TA + T19 + TBC
传输句	TC + T0a + $\sum$TBm
自身转移句	TA + T2b + $\sum$TBm
交换句	TBC1 + T49 + TBC2
替代句	TB1 + T4a + TBC2

转移句共有九个子类,前三类由转移概念的二级节点决定,后六类由三级节点决定,由此可见,转移概念语义网络的设计,不仅是二级节点服务于子类的划分[2],三级节点也同样如此。

一般转移句由四个语义块构成,共有 24 种排列方式,但汉语主要选用了下面的四种非标准格式:

(1) TA + T + TC + TB　张三嫁祸于李四

(2) TA + TC + T + TB　张三把这个消息告诉了李四

(3) TA + TB + T + TC　张三向李四透露了这个消息

(4) TC + TA + T + TB　这个消息由张三告诉了李四

第一种非标准格式需要对语义块 TB 加标志符 l02,如例句中的"于";第二种需要对语义块 TC 加标志符 l0320,如例句中的"把",第三种需要对语义块 TB 加标志符 l0220,如例句中的"向";第四种需要对 A 语义块加标志符 l01,如例句中的"由"。标志符的安排规则是:当广义对象语义块两两连用时,对后者给出标志。这符合自然法则。上面的例句正是如此。当 TA,TB,TC 三者连用时,应对第二和第三个语义块给出标志。但这种排序方式过于违反自然性,汉语很少采用这种格式。

对第四种非标准格式,汉语可以省去(4)的 TA 标志,直接表达成"这个消息张三告诉了李四"。也可省去(1)的 TB,如"张三嫁祸李四"。这种可省略性,是由于 TC、TA、TB 的个性差异比较大的缘故,与"红旗漫卷西风"的省略语义块标志符类似。

一般转移句的 TB 和 TC 在传统语言学里称为间接宾语和直接宾语。从句类来看,所谓双宾语句一定是一般转移句。它有两个子类,物转移句信息转移句。分别以层次符号 22 和 23 为标志。后者的 C 语义块具有语句扩展性,可形成文【2】中所阐述的"2 + 2"句式。信息转移句有一个以层次符号 239 为标志的子类,其 T3C 的扩展将形成"2 + 2"句式中的兼语句。

接收句是三要素句,以接收者 TB 为第一对象,以层次符号 21 为标志。它有两个子类:一般接收句和针对性接收句,前者以层次符号 210(包括 21a 和 21b)为标志,后者以 219 为标志。两个子类的 T1C 有本质的区别,下一节有详细说明。

传输句和自身转移句也都是三要素句。在句类格式中,特征语义块的数字串分别写成 0a 和 2b。传输句 E 块 T0a 的写法意味着它有两个二级子类:T2a 和 T3a,即物传输句和信息传输句。自身转移句的特征层次符号也是如此, 但以 T2b 为主,也有 T3b,T0b。

传输句以 TC 为第一表述对象,这意味着它也具有语句扩展性。例如"人类第一次登上月球的消息迅即传遍了全世界"。这里 TC 的主体"人类第一次登上月球"是一个完整的句子,但在具体表达时与"消息"以偏正结构组合起来,"蜕化"成语义块。

自身转移句以 TA 为第一表述对象,它的三个语义块都具有鲜明的个性,这是极为难得的"机遇",在下一节将具体讨论对这一"机遇"的利用。

交换一定是相互的。在其标准格式中,语句要素 TBC 成双出现。但应该指出,交换句的 TBCC 不一定成双出现。这有两种情况,一是 TBCmC 之一隐含到 T49 中,买卖和借贷的概念就是这样,它们隐含着 TBCmC 之一的货币或某种契约。二是交换双方共享 TC,讨论就是这样。这就是说,交换句还有两个二级子类,分别以 E 块 T499 和 T49a 为标志。

交换句必有交换的双方,这一点与关系句必有双方一样,因此,交换句的格式变化在形式上与关系句类似。

替代这个概念是转移与效应的边缘概念。因此,替代句本身虽是一个独立的句类,但上面给出的标准格式实际上很少独立使用,往往是充当复合句类的第一个语句。

17.2 转移句各子类的特征

上一节对各子类的格式变化或点到为止,或避而未谈,因为这个问题与各子类的特征密切有关,放到这一节来讨论比较合适。

—— 一般转移句的 TB 省略

省略了 TB 的一般转移句是三要素句,标准格式是:

TA + T + TC

省略的前提是不必指明 TB,转移型生理活动 6222 属于这一情况,以 201 或 202 为 T 的特征要素的转移句也常有这个情况。

省略包括两种特定的意义,一是 TA 与 TB 合二而一,如本能活动的"食"。这时,T = v62221,TC 优先于 pw65221(熟食)、jw6(生食)或 jw528、jw518。后者与 v62221 的底层概念 v62221a(喝)、v62221b(抽、吸)优先搭配。二是 T 中包含了 TB 的信息,新闻业属于这一情况,这时,T = (vc232,va34)。

文【1】中曾谈到人类活动的本能、智能和社会三个层次,这一层次性表现也是句类知识的重要内容之一。作为一项普适知识,不必针对具体句类作具体说明。但当它影响到语句的格式和句式时,就需要予以关注,本能活动的"食"就属于这一情况。

省略转移句的变化格式与一般作用句相同,汉语常采用以下两种非标准格式:

TA + TC + T

TC + TA + T

——信息转移型作用效应句

在转移句里,信息转移的三级节点 9239 经常用来构成使 TC 获得语句扩展性的作用效应型概念。因此,可以考虑把这一特性赋予 9239 的某一四级节点,但利弊大体均衡,所以目前没有这样做。仍采用组合结构符号 # 和语义结构方程 4—1 的方式[7]予以表达。汉语的"督促,敦促,催促,命令,号令,勒令,密令,责令,指导,指挥,指点,指使,唆使,指示,指引,引导,引诱,告诫,警告,劝告,劝戒,劝勉,规劝,勉励,激励,策励,鼓励,鼓动,鞭策……",属于这一类概念。从这一概念集可以看出,其中有一个子类与效应基元概念 36 强关联,并着眼于未来。这个子类形成作用效应句中一个独特的联想脉络和语境:如果 TA 是人,则其层次和等级属性 g55 和 g56 一定高于 TB;TB 不仅是信息的接收者,同时是作用的承受者;四个语句要素的总体关系与作用效应句完全一样,只不过在形式上把(A/X)#(B/YC)转换成(TA/T)#(TB/TC),TB 后面的 TC 一定是 TB 的表现。(注:这里说的就是混合句类 T3XY * 31,在写作此文时尚未形成如此清晰的概念。)

—— 一般接收句的 TC 构成

一般接收句中的 T1 与一般转移句中的 T 对偶,即汉语中"发和收"的对偶。

一般接收句类似于作用句中的承受句,承受句的 X1C 包含作用句的两项要素:A 和 X。

一般接收句的 T1C 包含另一转移句的三项要素：TA，T 和 TC。

T1C = TA + T + TC

例如“张三收到了李四寄来的生日礼物”里的“李四寄来的生日礼物”，就包含上述三要素。但这个复合语义块的核心是 TC，不能省略，而 T 和 TA 都可以省略，特别是 T。（注：这里的 T1C 是句蜕块，当时尚未使用这个名词，但思路是明确的。）

——针对性接收句的 TBC 构成

T19BC 有两种构成方式是：（注：在语句的物理表示式中，广义对象语义块前面的句类符号约定可略去特征要素的数字符号，所以，T19BC 是 T19 句类 JK2 的完整符号表示。下面的表示式仍采用省略形式。但有些情况不能省略，例如 T2C 和 T3C。因为两者有本质区别，T2C 不具有语句扩展性，而 T3C 具有。）

（1）TBC = TBCB

（2）TBC = TBCB + TBCC

这就是说，针对性接收有两个子类，第一类只包含对象 B，第二类则包含对象及其表现。这里的对象及其表现，都与转移无直接关系。这就是针对性接收句与一般接收句的本质区别。（注：这里两个子类的说法是不正确的，这只是 T19BC 语义块构成的特性，而这一特性与语境有关。）

第二个二级子类的层次符号标志是 9219 或 c219。它是高级智能活动。这是用主体层（挂靠层）前面的“附着”层（本体层）作为子类区分标志的又一例子。

219 的核心概念是汉语的“寻找”，它关心的是对象的未知位置。未知的原因之一是它离开了原来的位置。从这个意义上说，针对性接收句第一子类的 TBCB 同转移仍有一定的关系。但第二子类关心的是对象的隐蔽性表现，这是 9219 的定义。隐蔽性包含位置的未知，但一般是指其他的未知表现而位置确知。因此，第二子类的内涵要比第一子类丰富得多。

从针对性接收第二子类的定义可知，其核心概念是汉语的“探”字。“探”的目的之一是变未知为确知，即变隐为显。目的之二是为了获得。这两者都是效应型概念，9219 常用来构成混合句类。这与 9239 常用来构成作用效应型句类非常类似。（注：应为 T3XY * 31 混合句类。）

针对性接收第二个二级子类按上述两个目的又可分为两个子类。但这两个二级子类的区分也不是依靠 9219 的底层扩展，而是靠附着层的改变。9219 主要是变隐为显，c219 主要是为了获得。汉语对后者的表达主要是“查，察”二字。这两个汉字有某种专业性分工，“察”字常用于 a5（法律），但界限十分模糊，不宜在概念层次，而只能在词汇层次加以说明。

——传输句的句类知识

转移的两个静态要素 TA 和 TB，在一般转移句和接收句中分别充当了标准格式的第一表述对象 JK1。我们曾在作用句中看到了各个子类是各静态要素“轮流坐庄”的语言现象，转移句同样具有这一现象。现在该轮到半静态要素 TC 来“坐庄”了。这就是传输句。

传输句有一点与接收句类似,就是 TA 和 T 变成了 TC 的属性知识。它把表述的重点转到了 TC 和辅对象 TBm。例如下面的句子:

三峡水电站发出的强大电力将输送到能源缺乏的华东和华南地区

这里的‘三峡水电站将发出强大的电力’是一个省略了 TB 的一般转移句(发出有两个义项,一是 v202,二是(v311,v202),此句应取第二义项,因而严格说来它是一个效应转移句)。传输句先把这个省略了 TB 的转移句蜕化成 TC 语义块,而后扩大 TB 的信息含量(此话的含义下文有进一步说明)。这就是传输句的本质。

上面的句子可改成下面的形式:

三峡水电站将把它的强大电力输送到能源缺乏的华东和华南地区

三峡水电站将向能源缺乏的华东和华南地区输送强大的电力

强大的电力将由三峡水电站输送到能源缺乏的华东和华南地区

这是一般转移句的三种非标准格式,而其标准格式‘三峡水电站输送能源给缺乏的华东和华南地区强大的电力’,反而显得十分别扭。这正是我们把传输句列为转移句的一个子类的原因之一。

以前我们看到过作用句与承受句的转换[15],作用句与效应句的转换[14],这里又看到了一般转移句与传输句的转换。这种转换只是表述参照点的变动,按句类划分的各语义块所充当的角色并不发生变化。这里我们看到了这一论点的新证据。

上文有‘传输句扩大 TB 信息含量’的提法,现在对此加以说明。对转移句的表述对象,我们定义了三个一般对象 TB 和一个特殊对象 TA。从空间关系来说,TA 常与 TB1 在一起,TB 常与 TB2 在一起。在语句表达时,这两两‘在一起’的对象往往可以合并,或互相替代。常见的是用 TB1 代替 TA,用 TB2 代替 TB。在一般转移句里,TB1 和 TB2 不充当语义块的要素,但可以充当 TA 和 TB 语义块的说明部分,或作为辅语义块出现。在传输句里恰恰相反,TA 不充当语义块的要素,而充当 TC 的说明部分,TB1 和 TB2 则可充当 TB 语义块的要素,这就是‘传输句扩大 TB 信息含量’的第一层意思。

传输句标准格式里的 TBm 不仅包括 TB1 和 TB2,还包括 TB3。这是扩大 TB 信息含量的第二层意思。这里需要对 TB3 加以说明。按语义块表示的定义[2],B1 和 B2 可代表对偶的双方,而 TB1 和 TB2 正是转移起点和终点的对偶。在起点和终点之间必有一系列的过渡点,它们形成路径,并用 TB3 表示。这里有两个概念:路径和过渡点。但语义块可合而为一,因为它们是包含性概念。其层次符号是 204 - 和 204 - 0,而起点和终点的层次符号分别是 205 和 206。

凡是含有不只一个一般对象 B 的句类,例如关系句和转移句的交换替代子类,其标准格式具有稳定的模糊性[18]。传输句的 TBm 可包括 TB1,TB2 和 TB3 三项,这就是说,TBm 可由三个语义块构成。例如‘成千万吨的的煤每年从山西通过大秦铁路和秦皇岛港海运到华东地区’这个传输句里就有五个语义块。除了 TC—成千万吨的煤,T22a—海运之外,还有 TB1—山西,TB3—大秦铁路和秦皇岛港,TB2—华东地区。这里,每个 TBm 前面都给出了语

义块标志符。TB3 里包括了路径大秦铁路和中转站秦皇岛港。

语义块 TBm 所用的指示符属于辅要素途径的 lj1342k。例句中的‘通过、从、到’的逻辑语义映射符分别相应于 k = 4,5,6。(注:这里‘从、到’映射符号的说法是错误的,它们的语言逻辑符号是多义的。主要是 l15;‘从’更是 l19 的主要反映射词。这里应取义项 l02205,而‘到’应取 hv。)由辅要素指示符所指示的通常是辅语义块,但转移句中的传输句和自身转移句例外,因为,TBm 是它们必需的要素。其中的 TB1 或 TB2 更是不可或缺,如果单用 TB2,汉语可不加指示符。例如‘货运上海"货发上海’的‘上海’都是 TB2。关于途径和工具在转移句中的特殊性和主辅语义块的转换在【2】和【11】中都有所说明,这里就不来重复了。

传输句和自身转移句必须有 TBm,因此,其语句物理表示式用了特殊的语义块表示符号 ∑TBm,这就是把它独立于一般转移句而自成一个子类的主要原因,这一知识在句类分析时必须加以利用。

——自身转移句的句类知识

在所有的转移句子类中,自身转移句的个性最为分明。它取消了一般转移句中的两大要素 TA 和 TB,TA 与 TC 合并变为 TA,TB 由 ∑TBm 取代。以 T2b 为 E 块的自身转移句第一表述对象 TA 的个性最强,它一定是人、动物或 pw22b。

这里说明一下 TBm 的概念优先性。TBm 的定义是空间位置,自然与 j2(包括 wj2、pj2、pwj2)有最密切的联系,但任何物和人以及它们的一部分都都可以充当空间位置的替代表示。因此,“孙悟空钻进了铁扇公主的肚子’这样的自身转移句并不违反 TBm 的定义,因为,“铁扇公主的肚子’也属于广义空间 wj01。这就是说,可表达 TBm 的概念仍然十分宽泛。但这并不等于说对 TBm 的个性无可利用,关键在于语境信息的取得。例如,如果涉及专业性活动 ay,则 pe 类概念的优先级上升 如果涉及战争 a4 或交往及娱乐活动 97,则 wj2 - 000 类概念的优先级上升。这些依赖于语境的句类知识属于论文【6】中所阐述的概念关联性知识,在 B 类关联栏目中有详细清单,但目前未给出语境的标记。

关于自身转移句的句类知识,还应该特别指出下列两点:

第一是自身转移句的非标准格式往往不另加语义块标志,汉语的口语尤为常见。例如“我们去过上海’可以说成‘上海我们去过’或‘我们上海去过”。这种省略必须以语义块个性鲜明为前提。

第二是自身转移句常充当复合句类的构成部分。这一特性与替代句相似。在自身转移句后面,往往跟随着复合句的主表现语义块。例如,“他回家探亲去了"他去上海开会了’等等。

——交换句的句类知识

交换句的第一个特点是它的‘本能、智能、社会’层次性表现十分突出,第二个特点是它的语句要素 TBC1 和 TBC2 必然具有对仗性。这两项知识对于解模糊及纠错处理十分有效。

含有两个一般对象 B 的语句,如关系句和这里交换句和替代句,其格式变化最为复杂,这个问题将在‘关系句的句类知识’[18]中一起讨论。

结 束 语

在引言中曾谈到,转移句各语义块个性鲜明,这是它的第一个特点。

转移句的第二个特点是,格式和句式的模糊度最小。

本文的叙述大体上是围绕着这两个基本特点而展开的。

让计算机把握和运用句类格式和句式知识,是使它形成文【1】所说的预期和判断能力的基础,在这个基础上利用概念层面和词汇层面的概念关联知识,就能形成有效的解模糊及纠错能力。

转移句是检验这一设想的最好句类。

1995 年 5 月

混合句的句类知识

引　　言

本文不仅是句类知识论文系列的最后一篇,也是〈论文〉系列的最后一篇,因此,本文将首先对 HNC 理论的核心内容——语句表示式作一个总结性的说明。

混合句的句类知识,主要是对混合句类构成的物理阐释及其构成方式的说明。这是本文的主要内容。

由作用效应链的 6 种基本句类可生成 $6 \times 5 = 30$ 种混合句类,我们还可以进一步推论,如果基本句类共有 M 种子类,则应有 $M \times (M - 1)$ 种混合句子类。在这一推论中,仅考虑了两两混合,也未涉及句类格式的变化。由此可见,语句类型的数量将是十分庞大的。但应该指出,HNC 理论可以穷尽基本句类的子类数量,在这个基础上,就不难对语句类型进行穷尽式的研究,这应该是指日可待的。

21.1　关于语句表示式的一般讨论

到本文为止,我们已研究了全部基本句类——仅表述作用效应链一个环节的句类,给出了它们的相应语句表示式,现在可以对这些语句表示式作一个综合说明。

1. HNC 以主语义块连加的形式构成语句表示式,如 XJ = A + X + B, TJ = TA + T + TB + TC 等。我们把这些表示式称为句类格式。这些格式带有句类和语义块的命名。这些命名符号代表了相应表示式(包括句类表示式和语义块表示式)的物理意义,所以,这些以主语义块物理表示式相加构成的表示式,命名为语句的物理表示式。如果抽去这些物理意义,仅从数学上加以表述,可以给出下列纯数学形式的语句表示式。

$$J2 = JK + E$$

$$J3 = JK1 + E + JK2$$

$$J4 = JK1 + E + JK2 + JK3$$

这三个表示式分别代表两主块、三主块和四主块句,其一般形式可写成

$$J_{n+1} = JK1 + E + \sum_{k=2}^{n} JK_k$$

在这个表示式里,将 E 块固定安排在第二号位置上,这个安排代表了语句的语义块自然顺序,汉语和印欧语系等 SVO 型语言都遵守这一自然顺序。

表示式里的 JK 命名为广义对象语义块。

2. HNC 所定义的语句表示式只包含主语义块,不包含辅语义块。从语句表示式中排除辅块,是构造语句表示式的一项重要举措,其意义不在于减少表示式的复杂性,而在于带来思考空间的净化。

3. 上列由语义块自然顺序构成的语句表示式叫做标准格式。如果改变语义块的自然顺序,则需要对相邻的 JK 给以标记,这些标记就是语言逻辑概念的 l0 和 l2。我们把违反自然顺序的语句表示式叫做非标准格式。

4. 非标准格式又分为两种情况。一种是在所有相邻 JK 之间都加上标记,这种情况被命名为规范格式。另一种是在某些或全部 JK 之间不加标记,这种情况被命名为违例格式。

5. 上面的说明表明,JK 是否另加标记与格式之是否标准有关,也就是说,与 JK 的位置有关。但辅块是否加标记,则与其位置无关,也就是说与格式之标准与否无关。辅块的这一特性也可理解为 HNC 把它们定义为辅块的依据之一,上面所说‘带来思考空间的净化’就是指的这一点。

6. JK 的类别必须划分出 A,B,C 三类,这一点在文【2】中曾详加阐述。自然顺序的含义不仅是 E 块必须在第二位置,而且规定了 A,B,C 的顺序:A 必须在 B 之前,B 必须在 C 之前。

7. 必须进一步指出,A,B,C 只是语义块的类别基元,实际的语义块可以由它们复合构成,基本的复合形式是 BC 或 AC,复合语义块可能以 C 为主体,也可能以 A 或 B 为主体。上述 A,B,C 的自然顺序对复合语义块则以其主体的类别属性为准。

8. 以 C 为主体的复合语义块也具有 C 块的对象、表现二重性,既可扩展为语句,也可由语句蜕化而来,这两种汉语常见的语言现象分别简称为块扩和句蜕。

9. JK 的上述特性都与句类密切相关,例如,物转移句的 JK3 = T2C 不具备块扩特性,而信息转移句的 JK3 = T3C 则优先块扩;作用句的 JK2 通常以对象为主体,取语义块表示式为 JK2 = B,但反应句的 JK2 通常以 C 为主体,取语义块表示式为 JK2 = XBC。这里的‘通常’表示这一知识属于概念层面,是特定句类的优先表现,不能形成绝对的规则,但可以作为默认规则来使用,这就是说,当 HNC 词汇知识库未针对特定词语作出例外说明时,就可以默认它们为规则。

10. 语句的数学表示式用‘J = ’起写,语句的物理表示式则用‘EJ = ’来起写。“J = ”的右方以 JK_k 和 E 表示语义块;“EJ = ”的右方则以语义块的具体物理命名来表示,从文【14】到本文所给出的句类格式都是语句的物理表示式。

11. 实际的自然语言语句不仅有规范格式和违例格式,还有省略格式,即省略语句表示式中应有的语义块。汉语不仅经常省略 JK1,还可以省略 E。汉语还有更惊人的表现,就是不仅省略 E,还进一步省略 $\sum JK_k$,只剩下一个孤零零的 JK1,如马致远的著名诗句‘小桥流水人家,古道西风瘦马。’传统语言学定义的祈使句常省略全部 JK,剩下一个孤零零的 E,如‘站住,停止射击’。口语的‘谢谢,再见’也属于这个情况。至于李清照的著名诗句‘寻寻觅觅,

冷冷清清,凄凄惨惨戚戚'则是这两种彻底省略的连用了。

21.2 混合句类的物理阐释

混合句的基本特征是对作用效应链的运作进行表述,而不是像基本句类那样,孤立地表述作用效应链的一个环节,因此,它至少涉及作用效应链的两个环节。

在一个语句里,对作用效应链的运作进行表述有两种基本方式,第一种方式只用一个 E 块,第二种方式采用两个 E 块。前者命名为内混合句类或混合句类,后者命名为外混合句类或复合句类。

本文只讨论内混合句类,不涉及外混合句类。

在外混合句类中,有一种汉语偏爱的特殊类型,即文【2】定义的作用效应句

$$XYJ = A + X + B + YC,\quad YC = E + EC$$

将被当做基本句类的子类之一来处理,这样做的目的是为了把某些常用的多重混合简化为双重混合。

上面给出了混合句类的定义,下面需要进一步说明的是混合句类与基本句类的模糊性以及对这一模糊的处理策略。

混合句类与基本句类之间存在模糊。这一模糊来于两个方面,一与 E 语义块构成方式有关,二与构成 E 块的概念节点有关。文【14】曾经指出:E 块构成通常具有

$$E = QE + EQ + EH + HE$$

的复合形式,式中的 QE 和(EQ + EH)都具备形成混合句类的潜在条件,但这样的混合句类只能由软件自动辨识,知识库很难提供帮助。因此,对于由 E 块复合构成形成的混合句类,拟全部简化为基本句类来处理,这种简化,可产生净化思考空间的效果。

QE 的 jlvu12 类概念,EQ 的 vv 类概念都会产生复合句类,也就是说,它们对句类的生成是有贡献的,但软件可以暂时无视它们的贡献,一律当作基本句类来处理。例如下面的例句:

我们如期完成了任务。

我们一定如期完成任务。

我们一定努力如期完成任务。

这三个句子可以都当作效应句,尽管第二句加上了 QE"一定",第三个句子加上了 EQ"努力"。

主体基元概念所定义的全部高层节点都是划分基本句类子类的依据。但是,应该强调指出,由 E 概念的 HNC 符号并不能完全确定它对应的句类,下面将对此详加阐述。

1. 某些主体基元概念本身就具有二重性,文【14】中详细阐述的效应网络一级概念节点的二重性就是典型代表。至于复合基元概念,绝大多数都具有跨类特征。从整个概念体系来说,ϕ 网络的全部节点不过是概念海洋中的"基元一族"。实际的语词概念多数由它们复

合而成。因此,从构成 E 块的具体语词到句类的判定需要另行制定一套符号,为计算机做句类辨识提供明确的辨识信息。这个辨识信息将命名为句类代码,句类代码是句类辨识和句类分析的基础,因而是最重要的知识项,是 HNC 知识表示的纲、统帅和灵魂。

2. 正是基于这一认识,我们对 80 年代以来风起云涌的语料库热保持冷静,因为,我们深知,句类代码这一统帅语句分析的最重要知识不可能通过语料库的统计加工得到,只能由人来教给计算机。计算机不具备动物的感官系统,更不具备人类大脑的知识加工和存储系统,即使将来巨型机的存储容量超过了大脑的神经元数量(1,000 亿),也配置了相应的‘视觉’和‘听觉’系统,它能否达到人类的基本智能——例如像常人那样运用和理解自然语言,仍然是未知数。突破这一重大奥秘的关键何在?HNC 理论认为:使计算机先掌握句类代码知识,是关键的第一步。

3. 句类代码的判定和标定将是一项十分繁重的任务。首先,要确定完备的基本句类子类,并判定相应的基本句类代码,这项目标已不难实现,第二,由基本句类代码生成混合句类代码,这将在下一节作简要讨论。第三,对每一个可能构成 E 块核心的语词,标定它的句类代码,这将是 HNC 知识库建设中最关键、最费思考、机器不可代替(但可以提供语料帮助)的艰巨工作。

4. 句类的判定存在理论上的不确定性,在主体基元网络一级节点的设计过程中,我们有意按‘纯净性’安排节点的顺序,前面的节点比较纯净,后面的节点与其他网络节点具有更多的‘纠葛’。例如,节点 φ3a 与 φ20,φ46 强关联,其反映射语词所构成的句类往往不仅涉及作用和效应,也涉及转移或关系。在 HNC 知识库建设过程中,必然会经常遇到句类代码两可的矛盾,这个矛盾只能在实践中逐步解决。在解决这个矛盾过程中,逐步提高理论水平,把句类的不确定性减低到最小限度,将是一项意义重大的发展。

5. HNC 符号体系的设计,目前,仅完成了高层和中层的结构及其细则,底层结构则仅有框架而无细则,我们设想,每个网络根节点的底层将更多体现句类的混合。这将是未来 HNC 符号体系底层设计的重要原则之一。

21.3 混合句类表示式

同基本句类一样,混合句类表示式当然也有物理表示式和数学表示式之分。

混合句类的数学表示式可沿用基本句类的表示式,差异仅在 E 块的表示式不同,以 E1E2 代替 E,如下式所示:

$$J_{n+1} = JK1 + E1E2 + \sum_{k=2}^{n} JK_k$$

式中的 E1 和 E2 是参与混合的两基本句类,E1E2 将作为混合句类的句类标记。

混合句类的主语义块数量,当然不会像基本句类那样有 4 的限制,但一个句子的主块数量不可能很多,它受到所谓‘Miller 魔数 7 ± 2’的约束。

混合句类主块的数量通常不难确定,关键是从参与混合的两句类 E1,E2 中各取哪些语义块。理论上也许需要从两基本句类轮次选取,这会造成表示上的极大不便。我们将约定,混合句类的语义块顺序是:先从 E1 中选取若干个 JK,而且必须从 JK1 起依次连续选用,不能跳跃。然后选取 E2 中的 JK,起点可以任意,同样也要连续选用。

这样约定之后,混合句类的句类代码就可以采用下面的构成方式:

E1E2 * kmn

式中 k 表示 JK 总数,m 表示从 E1 中选用的 JK 数量,n 表示从 E2 中选用 JK 的起点。在表示式中,若 m = k,意味着混合句类的全部 JK 取自基本句类 E1,若 m = 0,意味着全部 JK 取自基本句类 E2。

按句类物理表示式的定义,混合句类各语义块物理表示式应采用 E1E2 为函数名,但这样作显然流于烦琐。我们将约定混合句类语义块表示式仍沿用原基本句类的语义块表示式。但应指出,这是一项简化,混合句类中相应语义块的含义实际上是有变化的,未来的深层理解要运用这一知识。

1995 年 12 月初稿
1997 年 8 月改定

附表 1　数字及小写英文字母的意义说明

小写字母意义说明

0　13	高层层次符号,用 16 进制表示(下同)
0　7	对偶性中层层次符号的简化表示,0　3 和 4　7 分别为两组对偶
8　11	底层层次符号
cnk,dnk	对比性中层层次符号表示
emk	对偶性中层层次符号表示
f	“语法”概念类别符号
g	五元组的因静态概念表示,取自汉语拼音 gai nian
h	后缀符号,取自汉语拼音 hou
i,k	层次符号的一般变量表示
j	基本概念类别标记,取自汉语拼音 ji ben
l	语言逻辑概念类别标记,取自汉语拼音 luo ji
m,n	层次符号的一般变量表示
p	具体概念的人,取自英语 people
pe	社会组织,这里的 e 是字母,而对偶性表示里的 e 是数字
p -	一般团体
q	前缀符号,取自汉语拼音 qian
r	五元组的果静态概念表示,取自英语 result
s	综合概念类别标记,兼有基元、基本、语言逻辑三大类概念的综合特征,取自英语前缀 syn
t	底层层次符号的变量表示,可带数字下标
u	五元组的属性概念表示,取自汉语拼音 shu xing,在双拼表示里 sh 一般用 u 替代
v	五元组的动态概念表示,取自英语 verb
w	具体概念的物,取自汉语拼音 wu
x	物性概念表示,兼有抽象具体的双重特征,取自汉语拼音 xing
y	高层层次符号的变量表示,可带数字下标
z	五元组的值表示,取自汉语拼音 zhi

附表 2　大写英文字母的意义说明及句类表示示例

大写字母意义说明

A	特殊对象语义块,亦称作用者块素,或简称 A 块
B	对象语义块,亦称对象块素,或简称 B 块
C	内容语义块,或特殊表现语义块,亦称内容块素或简称 C 块
D	一般判断句 E 要素
E	特征语义块,亦称 E 要素或 E 块,取自英语词根 eigen
FK	语义块块素符号,但不用于 E 块表示
H	语义块块素指示,用于语义块的形式分解,不单独使用,前后之后的意思,取自汉语拼音 hou
J	句子,取自汉语拼音 juzi
K	语义块形式符号,取自汉语拼音 kuai
P	过程句 E 要素,取自英语 process
Ph	短语
Q	同 H,与 H 配合使用,前后之前的意思,取自汉语拼音 qian
R	关系句 E 块标志,取自英语 relation
S	状态句 E 块标志,取自英语 state
T	转移句 E 块标志,取自英语 transfer
X	作用句 E 块标志
Y	效应句 E 块标志
jD	基本判断句 E 块标志

语义块分解的两种方式

物理分解示例

B = XB + YB + YC	良性分解
RB = RB1RB2	可分可合分解
YC = YCB + YCC；YCC + YCB	对象内容分解,如果存在确定顺序,称良性分解
YCB = YCBB + YCBC；	
YCC = YCCB + YCCC；	此分解过程可无限延拓

形式分解

K = KQ + KH	
KQ = KQQ + KQH	
KH = KHQ + KHH	此分解过程也可无限延拓
FK = FKQ + FKH	
FKQ = FKQQ + FKQH	
FKH = FKHQ + FKHH	

基本句类

XJ	**作用句**
PJ	**过程句**
TJ	**转移句**
YJ	**效应句**
RJ	**关系句**
SJ	**状态句**
DJ	**判断句**
jDJ	**基本判断句**

基本句类一级子类示例(总计 57 种)

X1J	**承受句**
X2J	**反应句**
X3J	**免除句**
X4J	**约束句**
T1J	**接收句**
T2J	**物转移句**
T3J	**信息转移句**
T49J	**交换句**

混合句类示例(一级子类,总计 3192 种)

YXJ	**效应作用句**
RXJ	**关系作用句**
XRJ	**作用关系句**
YSJ	**效应状态句**

复合句类示例(两级复合也是 3192 种)

T * XJ	**转移作用复合句**
T1 * X2J	**接收反应复合句**

语句物理表示式示例

XJ = A + X + B	**一般作用句**
T3J = TA + T3 + TB + T3C	**信息转移句**
XYJ = A + X + B + YC， YC = E + EC	**作用效应句,作为基本句类处理**
XY02J = A + XY02 + B + YB2	**作用句与双对象效应句的混合句,一般采用规范格式**
XRm110J = A + XR + RB2 + RC	**作用句与单向扩展关系句的混合句**
T2b * XJ = TA + T2b + TB2 + X + B	**自身转移句与作用句的复合句**

混合句类和复合句类的物理表示式由计算机根据现场数据由基本句类物理表示式自动生成。

第 三 部 分

HNC 理解处理的 52 个论题

一论中西语言的基本差异

——汉语昭昭语义块,西文短语细标明

汉语有明确的语义块指示标记,却没有完善的短语标记,西语恰恰相反。

因此,汉语的基本特色是:语义块昭然,短语模糊;而西语的特色是:短语昭然,而语义块模糊。汉语有"把、被、向、对……"之类的语义块指示符,西语是不存在的;西语有比较完备的短语指示符,如"the,a,for,with,……"等,汉语没那么完备。

于是,汉语会经常出现"给他‖一个苹果"之类的语义块分隔模糊,而西语基本不存在;另一方面,西语会经常出现"the cow with crumple horn that Farm Giles likes"之类的语义块构成模糊,而汉语基本不存在。

这是中西语言的基本差异之一。因此,汉语计算机处理应从语义块感知入手,而西语应从短语感知入手。

这一点,可谓天经地义!

但应该指出:西语在短语感知之后,仍必须上升到语义块感知和句类分析,采取先下后上之策略。而汉语在语义块感知和句类辨识之后,最后仍须回到短语感知,采取先上后下之策略。

如果置汉语与西语的这一根本差异与不顾,盲目模仿西语的处理模式,在短语类型上呕心沥血,随着句法分析之路亦步亦趋,岂不悲哉!

基于 HNC 理论的句类分析之路在绝大多数情况下已经畅通。但仍有一道不容忽视的难关,这就是在两 JK 之间不存在指示符的"佯谬"现象。这是汉语过于忽视语法造成的佯谬,是必须面对的佯谬。

这一现象将命名为 BC 佯谬,对它的处理将命名为 BC 辨识。

BC 辨识的难度与句类有关,也与格式有关。

句法分析对此当然一筹莫展,但句类分析应有天成妙策,有待妙手取之。

在 1.0 版(注:指"HNC 汉语拼音智能输入系统"的 1.0 版,后同),对 BC 佯谬不要求全歼,但也绝不可全部放过,策划组应为此制定具体目标与对策。

论句类假设的策略

1-1.1 引　　言

HNC 理解处理与传统句法分析的根本区别在于:一个从语义块感知入手,形成句类假设,转入句类分析;一个从短语辨识入手,形成句法树。

汉语是 HNC 的理解处理策略最好的试验场。

语义块感知和句类假设以语句的数学和物理表示式为依托,这些表示式的基本类型已完全确定,复杂类型可以组装,物理表示式是对概念联想脉络的具体描述,因而,句类分析可走向模拟大脑语言理解之路。

句法树以句型为依托,句型只是句子的形式表述,惟有列举,不可演绎,总量不定,无挂接联想脉络之求,不可能走向理解之路。

对文字文本,汉语的句类假设之路比较平坦,因为如主文所述,汉语有明确的语义块指示标记。但是,对于非文字文本,如拼音串或语音识别输出的音串,由于面对巨大的语音模糊,句类假设需要一个好的策略。

本文将围绕这一策略问题进行讨论。

为了读者阅读本序列论题的方便,先给出下列术语说明表。这些术语都有其常规意义,但我们赋予了特殊约定,需要注释。

音串——两处理标记(逗号、句号、问号、分号)之间的拼音串,第一个音串的左边是段落起始标记。

切口——两相邻音节不能组成双音词的交界点。

虚切口——多字词的左右边界。

音段——两相邻切口之间的拼音串,一个音串通常由若干音段组成。第一个音段的左切口就是音串左端。多字词构成独立音段。

奇段——音段音节数为奇数。

偶段——音段音节数为偶数。

单音段——只有一个音节的音段。

双音段——只有两个音节的音段。

混合音段——含有多字词的音段。

双音词——区别于双字词,是双字词的模糊集,在双字词后用 * 表示。

单音词——区别于单字词,是单字词的模糊集。

上层——即音段上层,指音段中奇偶音节相继构成的双音词序列。

下层——即音段下层,指音段中偶奇音节相继构成的双音词序列。

夹层——表征多字词的结构体,包括左右虚切口和多字词内部的双字词。

1-1.2 从 lv 准则谈起

句类假设的全称是句类格式假设,也可简称假设,6 年前最早起的名字是 lv 准则。

lv 准则里的 l 只是 l0,不包括 l1,因为后者不进入句类表示式。将 l0 与 v 绑在一起就意味着运用句类格式知识,这是 lv 准则的要点。

lv 准则包含下列 6 条基本规则:

规则 1:l0 必须在 E 或 EH 之前,不能在 E 或 EH 之后。

规则 2:如果 E 之前只有一个 JK,则一定是标准格式。

规则 3:对广义作用句,如果 E 在句尾,必定是规范或违例格式。

规则 4:对广义作用句,如果 E 在句首,必定是 !310 格式。

规则 5:规则 1 可以推广到 l1。

规则 6:如果在 l0 之后出现多个 v,则优先选取殿后者为 E。

上列 6 条规则既是格式判定的依据,也是消除多个动词干扰的依据。

1-1.3 多动词干扰的两种基本类型及相应的抗干扰方案

多动词干扰的两种基本类型是:分离型干扰和连见型干扰。前者指两动词之间插入了其他的语义块,后者指动词连续出现。

lv 准则的前 5 条规则适用于排除分离干扰,规则 6 适用于排除连见干扰。但仅有这些规则还不能形成完整的抗干扰方案。对分离干扰,还需要应用句类转换和语义块的理论以及块扩与句蜕的理论。对连见干扰,还需要使用"抓两头"的概念。而最后的王牌是句类检验。这是多动词抗干扰方案的要点。下面分别作具体说明。

——动词分离的抗干扰方案

先考虑两动词分离的简单情况,两动词依出现先后分别记为 v1 和 v2。

规则 1:如果 v1 为 E_t,则 v2 为 E。

规则 2:如果 v1 为 EQ,则 v2 为 EH。

规则 3:如果 v1 为 E,则 v2 实现块扩或表现句蜕。

规则 4:如果 v2 为 E,则 v1 表现句蜕。

规则 5:如果 v1 与 v2 共用某一语义块,则 v1 和 v2 分别为 E1 和 E2,并共同构成复合句。

这 5 条规则是演绎的结果,不是综合或统计的结果,综合或统计是得不到这个结果的。演绎的立足点是 HNC 的语句理论,包括语句格式理论和语义块构成理论等。如果以句法树为立足点,那就不可能进行这一演绎。

这些演绎规则覆盖了 v1, v2 可扮演角色的各种可能性,这一点可以保证,但关键在于规则前提的确认。

前 4 条规则前提的确认,属于句类假设的综合处理,下面分别作具体说明。

规则 1 是句类转换理论的直接应用。当实际的语句出现转换时,理解处理要实行反转换,恢复原语句的本来面目,以利于句类检验,即为句类检验提供更确切的联想脉络。E_t是句类转换的信息标记,在 HNC 知识库的句类代码栏目中有明确表示。

规则 2 是 E 块构成理论的应用,查清 E 块主体构成的分离,同样是为句类检验提供更确切的联想脉络。E 块分离信息在 HNC 知识库的 @K 栏目中有明确表示。

规则 3 和规则 4 是块扩理论和句蜕理论的应用。块扩的指示信息现在放在知识库的 @S 栏目中,最近在考虑换到概念类别栏目。句蜕的指示信息蕴涵在语义块的物理表示式中或 @S 栏目的 JK 构成表示式中。

规则 5 就是复合句类的定义。其前提的确认需要通过语句和语义块的自足性检验。

在规则 1 与规则 3,或规则 2 与规则 3 之间,可能出现两可疑难。疑难的消解度取决于上述自足性检验的水平。

上列 5 条规则的软件实现,就是前一阶段晋耀红的中心工作。

HNC 联合攻关小组的读者从上面的分析应能比较深入地理解,为什么我们向许嘉璐先生汇报提纲的内容,可以作为 HNC 技术取得突破性进展的标志。

上述动词两分离构架是动词分离的基本构架,多分离构架总是两分离构架的再分离,即分离嵌套现象,如块扩里的句蜕,句蜕里的块扩,再块扩,再句蜕等。西语的书面语常出现这种分离嵌套现象,汉语则较为少见。这里顺便说一句,【2】中提出的两种表现说,文中以最多篇幅阐述的 C 语义块概念都是基于动词的两分离现象而萌发的。

——动词连见的抗干扰策略

对动词连见的抗干扰策略,演绎的方法不像动词分离那样万无一失,但‘抓两头’的概念是演绎的并且是有效的。

“抓两头”的简略陈述是:在首首优先,在尾尾优先。说得详细一点就是,如果动词连见现象不是出现在音串的最后,就优先假定第一个动词为 E,如果连见出现在音串的最后,就优先假定最后的动词为 E。

“抓两头”的概念来于反证法,如果不是这样,则剩余动词将难以安排。抓‘首’,后面其余的动词可作为块扩或句蜕 JK 的 E,抓‘尾’,前面其余的动词可作为 QE(如果前面的动词具有 u 特性)或前面句蜕 JK 的 E。总之,后续安排比较容易。

对动词连见现象值得统计,但当前的条件还不具备,大家可在平时阅读时注意并记录,请刘老师安排一人总揽其事。

汉语动词的多元性表现在排除多动词干扰时也能发挥一定的作用,特别是纯 v 类概念及(v,ug) 类概念的利用。也就是形成一些优先性规则。请萧老师总揽此事。

关于多动词抗干扰方案就写这些,不过是原有意见的总结,但不能以软件要求的规范形式来写,只能表示一个老人的遗憾了。

1-1.4 语音文本的特殊措施

上述 lv 准则和多动词抗干扰方案是以一般语言信息串为对象,并不限于文字文本。但是,文字文本的模糊度很小,使用上述规则时节外生枝事件比较少见。语音文本则不同,它的模糊度大得多,节外生枝事件比较严重,需要采取一系列的特殊措施。

但是,并不需要对所有节外生枝事件都采取特殊措施,下面说明语音文本的各种节外生枝,同时说明相应的对策。

1. 双音词模糊集中的动词,这个情况与文字文本中的兼类动词(如 v,g 之类)并无本质区别,属于句类假设中的排队问题。当然,增加了后续解模糊处理的难度。

2. 文字文本的 l0 或 l1 基本上是无模糊的,但语音文本一定是模糊的,需要对这些特殊音节采用特殊方式来处理。

3. 对汉语的文字或语音文本都应该以音段为处理单元,而不能以词为处理单元,因为汉语的词一般不能定型于理解处理之前,必须在它之后,需要首先定型的只是语义块的标记,即 l0 l5 的语言逻辑概念。文字文本以单字词,双字词及多字词为主,三字段及多字段较少出现。语音文本则相反,以奇音段、偶音段和混合音段为主,当然也会出现双字词及多字词,但不能都作为可靠亮点,因为它们可能是伪词。关于伪词处理,将在论题 30 中讨论。

4. 汉语的绝大部分音节都含有动词,这是最大的节外生枝,因此在语义块感知和句类假设阶段,对单音动词必须忍痛割爱,先不予理睬,而后通过 K 调度予以补救。

1-1.5 结 束 语

此文从动笔到今天,已有一个月过去了,我想按雷良颖教授要求的规格试写此文,但本性难移,只得凑合。另外,句类假设还涉及奇段和偶段的不同处理策略,假设之后的排队,都已在或将在其他论题中阐述,此文就此结束。

1998 年 6 月 4 日

论“述语”之辨识

——鹤立鸡群与鸡肥鹤瘦

“汉语理解处理难于西语”。

在中国计算语言学界,这几乎是众口一词。

HNC 则答曰:“否”。

“难于”的主要依据是:

1. 西语词间有间隔,汉语词间无间隔,多了一条拦路虎——分词“瓶颈”。这一难关迄今屡攻不克。

2. 西语有发达的形态,词性分明,词性可作为判定语法成分的基本依据。汉语无形态之变,词性模糊,词的语法成分很难判定。尤为严重的是,汉语常出现动词满天飞的现象,述语很难判定。而西语的中心动词永远是形态分明的。

但是,论者不知:

1. 如果采用 HNC 的“中间切入,先上后下”处理策略的话,分词“瓶颈”可转化为“瓶底”。这个问题将在论题 29、30、31 中详细阐述。

2. 汉语的述语常若“鹤立鸡群”,既有 QE“鸣锣”于前,更有 hE 或 HE“击鼓”于后,述语本体也不甘寂寞,常以 EQ + EH 的复合结构凸现其“九五之尊”。这也许是几千年发达封建王朝造成的文化痕迹吧。

对雄立鸡群之鹤,辨识何难之有?

这就是 HNC 的答复。

当然,并非所有情况述语都如同鹤立鸡群,也有鹤不雄立,鸡肥鹤瘦的语句。基本判断句 jD 就属于这种情况。

鹤鸡之说虽然只是一个比方,但在多数情况反应了汉语述语的结构特征,可以作为述语识别的有效判据。对这一判据的运用,我在“HNC 技术漫谈”中已有详尽的阐述。

但是,应该指出,鸡鹤的比方只适用于语义块结构比较简单的语句。对下列四类语句的述语辨识,需要作进一步的考察,这四类语句是:

1. 复合句类
2. 块扩语句
3. 句蜕块语句

4. 并合型述语

复合句类的基本特征是两个述语共用一个或两个语义块。汉语常见的两种复合句类叫连动句和兼语句。连动句的两述语共用 JK1,兼语句述语 1 的 JK2 充当述语 2 的 JK1。就述语 1 为三主块句的复合来说,总共有四种复合形式,连动和兼语只是其中的两种。另外两种复合形式是:两述语共用 JK2,述语 1 的 JK1 充当述语 2 的 JK2。“鸡你不吃我吃”属于前者,“这个报告将提交大会讨论”属于后者。(注:E1 的 JK2 也充当 E2 的 JK1)

兼语句中反应作用效应链典型运作过程的语句特别命名为作用效应句,HNC 并把它当作作用句的基本子类之一来处理。其他基本句类或混合句类也可以通过 X0 转换变成作用效应句的形式。

单向扩展关系句在形式上很像兼语句,但实质不同。因为当 RC 扩展为 !31 形式的语句时,其 JK1 不仅是 RB2,还有 RB1,换句话说,这时 RB1 和 RB2 共同构成 RC 的 JK1。

块扩语句和句蜕语句都是汉语的特殊语言现象,其语法根源是由于汉语无关系代词充当从句接口,于是,从句只得以句蜕块的形式出现。

复合句、句蜕语句和块扩语句都提出了双述语辨识问题。

庄咏璆和刘志文的相应策论已对此作了专题讨论,我在论题 17、27、28 中也将有所补充。

并合型述语也呈现为 v 连见现象。对并合型述语应采用并合准则,它与“HNC 技术漫谈”中所说“首选”准则在形式上是相互冲突的。对两准则的选用应采用数据驱动方式,如果相连 v 交式关联,优先选用并合准则,否则优先选用首选准则。

论 E 假设策略

E 假设即句类假设。胡适之先生曾提出过“大胆假设,小心求证”的考证策略,后来郭沫若先生为了响应批判胡先生的运动,针锋相对地提出“小心假设,大胆论证”的反策略。两位先生各有侧重,但显然不必针锋相对,假设和求证都需要“大胆与小心”的结合,即全面与重点、整体与局部、广度与深度、直觉与定量分析的结合。

“全面”要求不放过任何一个可能的 E,即任何一个可能的动词,这对于双字词并不难做到。但是,又必须做到重点突出,先排除一批待选者,不让他们进入句类检验。随后,要对待选者进行排队,确定句类检验的顺序。先排除,后排队,这是 E 块感知的基本原则。

下面先讨论排除准则。

准则之一　紧靠“的”前面的动词一定可以排除,紧靠“的”后面的动词,除“是”和句尾(包括以“,”标记的小句)动词外,也都可以先行排除。

句尾动词常常是分离 E 块的 EH,这时它前面往往带“的”。例如:

声学所负责 HNC 系统的研制。

我们一定要深入进行国有企业的改革。

上面的两个例句代表了 E 块核心分离的两种典型情况,例句 1 表现了承受句的句类格式知识,例句 2 表现了 vv 概念“进行”的特性。

基于“的”排除动词的 E 块资格是就语句的全局而言的,就语句的局部来说,它可能仍然是局部的 E,即句蜕块的 E。这很类似于英语动词的形态变化或前面加“to”并不改变其动词本质。例如:

日本经济的持续衰退不利于东盟经济的复苏。

这里的“衰退、复苏”形式上变成了名词,实质上仍然起动词作用。这个句子是作用因果句 XP21 * 20 ,E 块前后的 JK 是两个句蜕块,都是由过程句 P3 蜕化而来。对这两个语义块的理解必须从蜕化前的原语句“日本经济持续衰退”和“东盟经济复苏”出发。因此,必须透过动词“衰退、复苏”词性变化的形式抓住其不变的本质,这个本质里蕴涵着“经济”与“衰退、复苏”的概念联想脉络。“的”前动词与“的”后动词都改变了词性,但从语义块构成分析的角度来看,前者是真改,后者是假改。这一点十分重要,我们欢迎语言学界对此作系统深入的研究。

这里还必须回答一个问题,“的”后面的动词(串)充当 EH 的情况只出现在句尾么?我认为,对汉语必须作这个假定,或者说作这项限定。当然,我们需要通过大规模真实语料对

此进行验证。

准则之二　l9 类概念之后的动词可以排除。单字词‘这、那、哪、某、任’和由它们构成的双字词‘这个、这些’等等属于此类概念。当初在设计语言逻辑网络概念时,正是由于指示代词这一宝贵特性而将它们独立出来构成 l9。

准则之三　“h□g”类概念之前的动词可以排除。单字词‘性’,双字词‘问题,方式’等属于此类概念,其含义是:它本身充当块素 FKH,与 FKQ 一起,构成静态概念 FK,即将前面的概念 g 化,包括 v 类概念的 g 化。

准则之四　后见排除,对连续出现的两个动态概念取第一个,排除第二个。

前面三条准则的使用,可以仅涉及语言表层,本准则的使用则必须进入语言深层,即运用词语的 HNC 符号。因为两个连见的动词可能属于并合型 EQ + EH 结构。如果两概念同行,或句类代码相同,则应确认为并合型 E 块,不满足上述条件才按‘后见排除’准则处理。

准则之五　“是”字准则。紧靠单字词‘是’的动词,不论前后,一律先行排除,将单字词“是”作为基本判断句 jD 的绝对激活因子。

本准则的使用看起来比较简单,实际上非常复杂,首先要注意到,应用前提只限于单字词,不包括含‘是’的双字词或双音词,这一点本来不必说明,单字词已明确规定了这一点。但由于对这一类不言自明的约束条件的疏忽而造成的频繁失误实在令人痛心,所以这里采用了不厌其烦的下策,但下策终究是下策,这一问题的根本解决,只能依靠参战者语言和 HNC 素质的增长。

本准则使用的难点当然不在于把握上述前提条件,而在于 jD 句类的特征。首先,所有的基本判断句都不存在直接的 E－JK 约束,只存在两 JK 之间的语义约束。jD 又属于其中最难把握的子类,因为其 JK 之间的约束知识主要是常识性知识。这就是说,常规的句类检验对于‘是’字句毫无意义。这里值得顺便说一句,“shi”属于汉语所有音节中最活跃的音节,“是”作为 E 块,又一定不穿下装,也少穿上装,因此‘shi’的直接感知必然伴随着草木皆兵的风险,汉语对 shi 的单音节使用达到令人恐怖的程度,例如常用于(X0, E)转换的‘使’,作短语后标记的‘时’,作基本命名的‘市、事、师’以及数字‘十’等等,因此,对汉语语音的处理,“shi”是第一号音节难关。

对于仅涉及对象及其属性说明的简单的‘是’字句,当‘是’出现在奇音段时,即使不指定“是”,句类的判定并不难。问题在于汉语常把一般句类转化为‘是’字句,这时就需要实施句类反转换,以便对原句类进行句类检验。这就使得‘是’字准则的应用出现十分复杂的情况。看下面的例句。

1. 张三是这部电影的导演。这部电影的导演是张三。
   张三是导演过这部电影。这部电影是张三导演的。
   这是张三导演的一部电影。是张三导演的这部电影。
2. 这一战略选择是符合我国当时的实际情况的。
3. 投资规模是决定经济增长速度的重要因素。

4. 科技水平较低,科技人才严重缺乏,劳动力素质较低,社会发育程度极低,商品经济意识不强,是造成西部地区经济发展水平低的重要原因。

5. 90 年代深化农村改革的重点是:继续稳定以家庭联产承包为主的责任制,不断完善统分结合的农业双层经营体制,积极发展农业社会化服务体系,逐步壮大集体经济实力。

6. 中国共产党第十五次全国代表大会是一次极为重要的大会,是在世纪之交,承前启后,继往开来,保证全党继承邓小平同志遗志,坚定不移地沿着十一届三中全会以来正确路线胜利前进的大会。

7. 前一任务是为后一任务扫清障碍,创造必要的前提。

8. 建国后特别是近二十年来我国已经形成可观的综合国力。

9. 我们党确立起已被实践证明是正确的建设有中国特色社会主义主义的基本理论和基本路线。

第 1 组例句是作用句"张三导演了(过)这部电影"的各种(jD, X)转换。对于基本判断句 jD 的反转换处理,需要作为一个专门问题进行研究,并形成相应的特殊处理系统。

第 2 个例句是比较判断句的 jD 转换(jD,jD0)。去掉"是、的",就完成了反转换,恢复成原来的句类。这个比较判断句出现了两动词连见的现象,但不能采用"后见排除"准则,这一点在下面还要谈到。

第 3 个例句是作用关系句的 jD 转换(jD,XR211)。反转换操作是去掉"是、的重要因素"。

4、5、6 三组例句代表了汉语"是"字句句群的常见类型。以后详谈。

7、8 两种例句代表了汉语中"是"字不充当 E 要素的两种典型情况:

1."是"后面紧跟"为、为了"。

2."是"前面有"特别"。

对这两种特殊情况,应建立相应的特殊处理系统。

第 9 个例句里"是"字句充当句蜕块。在"是"之前,已找到了 E 块"确立起",它是作用句,前面已出现作用者 A,A + X 不可能构成句蜕块,它后面必须有 B 块,不能采用省略形式,因此后面的"是"字句只能形成句蜕块。如果不运用这一句类知识,本句的句类分析将陷入困境。

jD 句类是最常用的句类之一,以上所说,不过是 jD 句类最基本的句类知识,对它的系统阐述需要多篇专文。但是,掌握了上列要点,就能处理大部分 jD 句类。

下面来讨论排队准则。这可以概括为下列优先准则:

1. 带上装、下装者优先。

2. 纯 v 优先。

3. 亮点优先。

汉语 E 块的上下装十分丰富,这为 E 块的判定带来了极大的方便。对汉语 E 块上下装

的分布情况作一个统计一定十分有趣和有益。我希望 HNC 联合攻关小组中出现有志于此的志愿者。

汉语的 v 概念多数属于 vg 或 v,g 型概念,纯 v 概念不多,遇到这样的概念一定要把它作 E 候选,不可放过。能否作这样的假设“汉语的纯 v,如果后不跟‘的’,前不跟 l9,则它即使不是整句的 E,也必然是块扩或句蜕的 E”?我倾向于先运用这一准则,如果将来发现有问题,再来进行修改。

对上面的例句 2 如果进行反转换,即去掉‘是、的”,就会出现动词‘选择符合’连见的现象,选择是 vg 型概念,符合是纯 v 概念,这时应优先作出比较判断句的假设,而不是按照排除准则,仅作作用句的假设。

亮点有两重含义,一是指无模糊,二是指不产生游离音节。亮点可以两者具备,也可以只有其一。指定字和无模糊的双字词或多字词通常只是满足第一个条件的不完全亮点,不能组成词的指定字才是完全亮点。

符合上述优先准则的假设称为良性假设,不符合的称为非良性假设。

上述三条准则的单独使用都比较方便,难点在于各准则的联合使用。我将在论题 2-2 中对此进行讨论。

E 假设还涉及概念类别信息的运用,这个问题请参看论题 33。

关于 v1 的处理策略

2-2.1 引　　言

我在联合研讨会第一次会议上发言的五点意见,张全博士已有记录稿,这里我着重对第二点意见的第一条作一些补充。

第二点意见总题目是“加强”,也可叫“当务之急”,是现有系统经常出现失误的薄弱环节。

在“加强”题目下,列举了 7 个薄弱环节,实际上当然不只这些,但这 7 个环节最为重要,而且不难解决。

薄弱环节之首是对第一个 v 的处理(以下简称 v1 或 v1 问题,意味着可能有 v2)。

这个问题是 E 感知的一部分,E 感知是语义块感知的关键,为什么这里单提出 v1 问题?这是因为,E 感知的各项信息的利用,软件基本上都注意到了,惟有这个问题似乎忽视了,造成了一种令人十分惋惜的状况,好像抱着了西瓜,却丢了芝麻,然而这芝麻是万万不能丢的。所以我把它列为亟待加强之首。

2-2.2 一 般 分 析

v1 问题应区分三种情况:

1. v1　前面无 JK,或 v1 在音串之首。
2. v1　在音串的最后。
3. v1　前面有 JK。

由语句表示式和格式代码知识可知,第三种情况属于一般正常情况,第一和第二两种情况属于特殊情况,当前的软件似乎这两个特殊情况很不适应,本文为此而写,所以并不讨论情况 3。

下面的讨论还有一个前提,就是认为 v1 已通过论题 2-1 所阐述的排除处理。没有这个前提,就不能得到下面给出的简明分析表。

详细的讨论只围绕着情况 1,情况 2 仅给出概要分析,细节留给读者去思考。一般分析结果用下面的表格来表示。这张表格体现了我对这个问题的思考过程。读者应能看到,这一思考过程除了一般的综合分析之外,主要是运用演绎的方法,而演绎的基础无非是句类块

扩句蜕等概念,没有这些概念就无从进行这里的演绎。细读此文的读者若能对此有所体会,则作者不胜欣喜。

情况 1(v1 在音串之首):

无 v2	!310		例句 1-1-1	作用句
	句蜕块		例句 1-1-2	
	S04		例句 1-1-3	
v2 在音串最后	v1 为 E	v2 块扩	例句 1-2-1	广义作用句
	v1 为 E	v2 句蜕	例句 1-2-2	广义作用句
	v2 为 E	v1 句蜕	例句 1-2-3	广义效应句
	复合句类		例句 1-2-4	
v2 之后存在 JK	v1 为 E	v2 块扩	例句 1-3-1	广义作用句
	v1 为 E	v2 句蜕	例句 1-3-2	广义作用句
	v2 为 E	v1 句蜕	例句 1-3-3	
	复合句类		例句 1-3-4	

情况 2(v1 在音串之尾):

标准格式		例句 2-1	广义效应句
规范格式		例句 2-2	广义作用句
违例格式		例句 2-3	转移句

这张分析结果表格共四列。第一列是情况的粗分类,对情况 1 又分了三个子类。第二列是对每一子类的格式分析,列举了各种可能。第三列给出相应的例句编目号,放在本文的下一节。第四列给出句类优先性知识。

无论是情况 1 还是情况 2,当 v1 出现时都意味着句类假设的全部条件(指句类代码及句类格式的信息)已经具备,处理进程可进入检验(情况 2)或等待检验(情况 1)的状态。情况 1 的第一子类永远意味着 !31 格式,它可能转化为句蜕,但这个“ !31 ”的本质不变。

应该说,E 块分别在音串的首中尾是语句的三种基本形态。但是,关于语义块布局的阐述,语句表示式绝大多数以 E 块居中(而且一定紧跟在 JK1 之后)为标准格式的事实,都容易造成 E 块必须居中的错觉。就汉语来说,这三种基本形态都常用,我的看法,这虽然不是汉语的基本特点,但可视为重要特点之一。

这里还应该再次强调一下 E 排除处理的重要性,排除处理应充分利用概念类别信息,这个问题在论题 1-1 中言有未尽,在论题 33 中 将作进一步阐述。

2-2.3 例句汇编

例句都以分段形式给出,采用了下面说明的符号,我希望它成为一种规范。顺便说一声,这样的分段形式五年来我键击过几千句,受益良多,并不觉得是浪费了时间。

在下文中，

…　表示音段的分界；|　表示虚切口；＊　表示该双音词存在模糊集；

#　表示该新词存在伪词干扰；　　表示新词，应通过段接构成。

例句 1-1-1

高举…邓小平 | 理论…的…伟大旗帜＊，

全面…贯彻…党…的…十五…大…精神＊，

（协约国）草拟…了…一份 # …条约＊，

瓜分…了…德国，

造成…了…一个…欧洲…政治＊经济＊疯人院，

稳定…和…加强…农业…的…基础…地位＊，

切实调动…和…保护农民…的…积极性，

落实党…在…农村…的…基本政策，

稳定…农产品…总量　，

大力＊…推进…农业…科技＊革命，

坚持…稳中 # 求进 # …的…指导＊方针＊，

当…了…一家…海运…公司…的…顾问＊，

例句 1-1-2

生活＊…在…沿海＊…城市＊…的…汉族…居民，

消灭…波兰…的…白色…方案，

这个…当时…被欢呼为一项 # …妙举　…的…条约＊，

例句 1-1-3

（待补）

例句 1-2-1

确保…农业…增产、农民…增…收、农村…稳定，

那…就…表明＊日本…的…金融…体系已经＊…崩溃，

要严格＊…防止＊集体＊资产…流失＊，

发现绝大多数为…华人拥有，

实行＊…村…务…和…财务＊公开，

实现＊经济＊…和…社会协调…发展，

组织＊一批…干部…到…基层…挂…职…任职＊，

例句 1-2-2

确保…农业…的…持续稳定…增长，

加强…农业基础…设施建设，

加强生态…环境＊建设，

加强…对…干部…的…民主监督＊，

增加…对…农业…的…投入，

例句 1-2-3

准备…白色…方案…的…工作正在加紧 * 进行，

投资 * …增…速…有所…增加，

造成…这种 # 结果…难以理解，

例句 1-3-1

帮助基层…解决实际 * …问题，

帮助…新来 # …的…成员…安排…食宿　，

健全…村…级民主…选举制度，

全然…未能…察觉…当今…的…亚洲…巨变将…会…怎样…

　　…改变 * 世界 * …的…面貌，

竭力 * 避免…官府 * 注意 * …他们…的…财富，

盼望日后…能…荣归故里 * ，

盼望…有…一天…能…发财…归来　，

总…想…赚…够…了…钱就 # 返回 * 家乡 * ，

例句 1-3-2

（待补）

例句 1-3-3

来自 * …亚洲、欧洲…和…美国…的…直接 * 投资 * …均…呈…下降…趋势 * ，

全国…累计…批准…外商 * 投资 * …企业…31 万零 5 百 70…家，

造成…这种 # 现象 * …的…原因 * …主要是…“过客心态”，

旅居…他乡仅仅…是…出于 # …纯粹…的…商业…动机 * ，

保持…低姿态…以…不…受贪官污吏　…的…搜刮。

2-2.4　对一般分析结果的讨论

1. 例句 1-1-1 分析

全部 12 个例句都是作用句，其中 11 个一般作用句，1 个承受状态句。从这少量的语料能得出什么结论？

首先，一般作用句在“1-1-1”情况居大多数绝不是偶然事件，而是势在必然。因为，专业性或社会性活动的作用型语句，其作用者往往是不言而喻，古汉语惜字如金，对不言而喻的作用者往往隐而不显，这是汉语的传统语言风格。这一语言风格近半个世纪在应用文中，特别是现代八股中更有所加强，其原因不难理解。

其次，一般地说，广义作用句（除判断句）都可采用 !310 格式，其中，约束句尤为突出，为此专门设置了 X401 子类。

第三，广义效应句很少采用 !310 格式，正是由于这个缘故，对效应句专门设置了 Y011

和 Y012 子类。

在全部 57 个语句物理表示式中,只有这三个是 E 块居首,设置这三个特殊语句表示式的目的是为了突出一般共性中的个性表现。

E 在首并随之出现 1-1-1 情况,可作为一般或广义作用句的最终判据。这就是说,当处理过程面临一般作用句与其他子类或广义作用句与广义效应句的两可局面时,软件可据此作出最终判断。

展望未来,在对汉语大规模真实语料进行句类分析以后,可以得到 !310 格式的各种定量句类知识,这些统计知识自有其学术价值,但 HNC 技术更感兴趣的是对上述论断的验证和深化。

2. 例句 1-2 和 1-3 的综合分析

汉语多动词难点有多种表现,形式上可分成连见和分离两种基本形态,由这两种基本形态可构成各种复合形态。两种基本形态中又有简单与复杂之分,简单形态就是两两连见或两两分离。多动词难点应从简单形态的剖析入手,以往的分析工作不够系统和深入,本论题系列希望前进一步,最后写成论文"汉语多动词难点的分析与处理"。

这里的综合分析是对两两分离形态的分析,包括 v1 和 v2 分居音串首尾和一首一中这两种特殊情况。上一节对这两种特殊情况的分析结果具有一般性,即表现了两两分离的一般规律。"一般寓于特殊之中"的哲学论断在这里得到了生动体现。

动词两两分离的一般规则可概述如下:

(1) v1 构成整个语句的 E 块,v2 实现块扩。

(2) v1 构成整个语句的 E 块,v2 表现句蜕。

(3) v2 构成整个语句的 E 块,v1 表现句蜕。

(4) v1 和 v2 联合构成复合句类。

这里引入了一项基本约定:E 块之前不存在块扩,只存在句蜕。块扩与句蜕都是子句的特殊形态,可理解为彼此的相互变换(见论题 28)。因此,这一约定不是理论上无可奈何的含糊,而是工程上有利无弊的净化。

3. 例句 1-2-1 与 1-2-2 的对比分析

把 1-2-1 中的某些例句加上"的"字可以变成 1-2-2 例句,反之,把 1-2-2 的某些例句去掉"的"字可以变成 1-2-1 例句。这表明块扩与句蜕可相互变换。这里应强调指出,这一变换是无条件的,这就是说,两组例句都可以相互变换,不过,有些例句的具体变换形式比较复杂,不是简单地增删一个"的"字就可以实现的。

4. 例句 1-2-1 与 1-2-3 的对比分析

这实质上是 v1 为 E 与 v2 为 E 的对比分析。

对比分析的基本结论是:v1 为 E 时优先广义作用句,v2 为 E 时优先广义效应句。有人可能担心,凭这么几个例句就能得出这个结论么?要知道,这个结论不是基于综合,而是基于演绎。从语句的物理表示式可知,广义作用句不存在 E 块在尾的标准格式,而广义效应

句恰恰相反。

有人会问,这一句类知识有何实际意义?

答曰:当 v1 或 v2 需要解模糊或纠错时,若这一句类知识正好派上用场,岂非妙不可言!

2-2.5 有待下回分解

本文暂时留下了一个问题,回避了一个问题。留下的是情况 2 的具体分析,回避的是复合句。

复合句与块扩的辨识,从何处着手?

为了便于软件的操作,是否有必要在知识菜单中给出可块扩动词的更简明信息?望读者思考。

1998 年 5 月 25 日

论 E 块主体构成及其分离

汉语‘字义基元化,词义组合化’的特点在 E 块主体构成上也有突出的反映,在形式上,我们用

$$E = EQ + EH = EQ + E = E + EH$$

表示式加以表达。这里给出了三种组合方式,EQ + EH 表示两部分构成贡献相当,EQ + E 表示主要贡献在构成的后部,E + EH 表示主要贡献在构成的前部。E 的句类知识主要由 E 提供。

EQ + EH 的构成通常是并合型结构,西语也经常采用,不表现汉语的特色。汉语特色表现在后两种类型。

EQ + E 表现为 vv + v 或 $E_h + E_l$ 的组合,E_h 表示高层 v 概念,E_l 表示低层 v 概念。vv 类概念是 HNC 引入的 v 类概念之一,它要求补充另一个 v 类概念,才能形成意义完备的 E 块主体。英语也有类似的构成,采用‘中心动词 + to + 动词’的组合形式。有的时候 vv 之后接另一个 vv,则两 vv 构成 EQ,后继的 v 构成 E。如‘努力进行’就属于双 vv 的 EQ 构成。

汉语的字义基元化,是使字义向高层或底层双向发展,而西语词义的扩展则主要是向底层发展,这就是西语的词一般词义较多,而汉语双音词词义较少的根本原因。当然,现代汉语双音词也有底层化倾向,从语言的进化来说,这是一种不良倾向,但由于汉字已停止‘新生’,字义基元化的能力也已大大衰退,这种不良倾向必将进一步发展,这是国际化潮流和废除文言文必然带来的文化代价。

西语只有少量的 E_h 类概念,汉语则丰富得多。双字词有‘进行,给以,予以,感到,展开,开展,受到,使得……’。EQ 与 E 经常出现分离,也就是说在 EQ 与 E 之间插入 E 的宾语。如‘进行产业结构调整”开展政治体制改革的研究’。传统句法分析要追究这里的‘调整,改革,研究’是动词还是名词,HNC 的回答是:这种追究只徒具形式,没有本质意义。理解的本质在于不论是‘政治体制改革’还是‘改革政治体制’的词序,‘政治体制’充当‘改革’的对象这一概念关联性的本质不变(更准确地说,‘政治体制’是‘改革’这一作用型概念的效应对象 YB)。同样,‘产业结构调整’和‘调整产业结构’的词序也不影响‘调整’与‘产业结构’的关系本质。当‘政治体制’与‘改革’,‘产业结构’与‘调整’相结合时,前者的 YB 角色和后者的 X 角色,不应该由于两者出现顺序不同而变化。大脑的感知就是对概念之间这一相互关联性的把握。在这一概念联想激活过程中,词性的作用显然是一个疑点,也许以西语为母语的人会对词性有所依赖,但以汉语为母语的中国人显然不应该依赖于词性。

EQ 与 E 之间当然还可以插入 E 的修饰语,如‘进行全面调整’,‘开展初步研究’,还可

以宾语和修饰语同时存在,如‘进行产业结构的全面调整’;‘开展政治体制的初步研究’。这时 E 的修饰语必须在宾语之后。这里要强调的是,对 EQ 与 E 之间的插入成分必须进行‘修饰语—宾语检验’。表面上看,词性在这一检验过程中显然具有速效之功,但大脑的感知在多大程度上依赖于词性,也是一个未知数。对汉语来说,这里存在词性模糊问题,因而必须以概念关联性为最终判据。

EQ + E 两种类型都可以插入,但形式上有所不同,$E_h + E_l$ 类型的插入宾语不加标记,而 vv + v 的插入宾语则必须另加标记。这里用了规则性的陈述,但是否完全符合汉语的实际情况有待来日以大规模真实语料来进行验证。

E 块主体构成的第三种组合方式 E + EH 为汉语所特有,是一种动—静结合方式,E 为动,EH 为静,如‘做手术,搞对象,有信心……’之类。这种组合方式可统称动静结合,‘动’多数为单字词。

在 E 与 EH 之间也可以插入宾语,但往往是宾语的一部分,如‘做肝脏手术’,这里的肝脏只是效应对象 YB,而作用对象 XB(患者)是分离在外的。

动静型主体构成的研究过去注意不够,在知识表述上也尚未形成规范。

E 块主体构成的多种组合方式是汉语的重大特色,本文不过略及其要。

论 E 块“上装”

——外衣内衣,坎肩随意

本文标题及其副标题有不严肃之嫌,但它确实表达了 E 块修饰语的特色。

E 块修饰语(上装)可划分为多少类,这里不来综述前人的论述,只简单陈述一下 HNC 的观点。

人类上装有内衣、外衣之分,内衣必须贴身穿,外衣必须穿在最外面,两者不能交换“位置”。中式上装还有坎肩,它通常穿在外衣与内衣之间,但有随意性,也可罩在外衣的外面。

E 块各类修饰成分的排列顺序具有内衣、外衣与坎肩的特性。

E 块修饰语的语义分类有:

1. 势态逻辑修饰——E 块外衣,如“应该、必须、可能、也许、必定……”之类。

HNC 将此类概念以 jlvu1(或 jluv,jluu)表示,它构成修饰语的外衣。在词类上属于传统语言学的情态动词。

2. 属性修饰之一——E 块的衬衣

E 块属性修饰语在词类上属于传统语言学的副词,HNC 将其中专用于 E 块修饰的部分抽出来记为 uv 类概念,这类概念的同行性比较充分,是发现 E 块的充分条件。当然,具备同行性的 uu 或 u 类概念也可充当内衣,但不能作为发现 E 的激活因子,因为它们不具备充分性。

3. 时态修饰——E 块的汗衫与裤衩

时态修饰有动静之分。静态修饰指“现在、过去、将来、进行、完成”等,动态修饰则指对其过程特性的说明,如“越来越、日见、马上……”等等。时态的静态修饰应纳入汗衫,HNC 记为 l6,而动态修饰则应纳入坎肩,HNC 记为 l7。

汉语对时态的静态表述似乎作了明确的二维划分,以“现在、过去、将来”为一维,以“进行、完成”为另一维,前者变为“汗衫”,后者变为“内裤”,相应的汉字是“着、了、过”。

4. 属性修饰之二——E 块的另一套裤衩

汉语的“内裤”十分发达。除了时态修饰的“着、了、过”之外,还有“到、去、来、上、下……”等。它们对 E 块的效应和空间特性予以说明,这里的“效应”和“空间”是广义的,实际上是基元概念和基本概念两方面的代表。

HNC 将所有的内裤记为 hv,hv 类概念是汉语里最重要的语素。大多数为单音形式,双

音形式为数不多,如‘一下、起来……’等。

hv 类是发现 E 要素、特别是单音 E 要素的重要线索,常用的 hv 应作为 E 块感知的激活因子。

5. 语言逻辑修饰——典型的坎肩

HNC 将这一类概念记为 luva 或 ljuva,在词类上也属于传统语言学的副词,但它们必须在 E 要素之前,是 E 感知的激活因子。

对语言逻辑概念的坎肩特性需要作进一步研究,其本体第二层还处于备用态,即留作此用。

6. hv 的对称概念——qv,E 块的另一套汗衫

它同样用于对 E 块的效应和空间说明,如‘去,来……”,HNC 记为 qv。以上对 E 块包装作了详细说明,包括外衣、衬衣、坎肩、两套汗衫与裤衩,每类包装都有明确的 HNC 符号。

E 块上装有序的概念是张全博士提出来的,我在这里姑妄言之,希望引向深入的研究。

论 E 块“下装”

今天是 1998 年 6 月 17 日,是本论题系列转变阐述方式的日子。

52 个论题分八组:第一组(论题 1—5)论句类假设,第二组(论题 6—12)论 JK 和 fK 感知,第三组(论题 13—17)对句类理论作补充说明,第四组(论题 18—20)论句类转换,第五组(论题 21—24)论音节感知,第六组(论题 25—34)论句类检验及块内处理,第七组(论题 35—39)论语义距离计算,第八组(论题 41—51)补充阐述 HNC 的一些基本概念。最后的论题 52 是本论题系列的总结。

在八组论题中,第一组和第六组涉及句类分析的两头,基于加强两头的需要,针对当前软件的弱点,按照一个外行心中自发的文档格式写了有关的论题。

从本篇开始,将恢复漫谈形式。谈要点,谈火花,谈万绿丛中一点红,少说貌似全面的废话。

汉语对 E 块精心打扮,上装下装,都有层次,不仅是义的需要,也有音的需要。义寓于音(包括韵律),但音有其自身的需求,于是出现冗余。深浅 = 深,树木 = 树,河流 = 河,都是音的需求。汉语的 ill 表现,很多是来于音的需求。从义看,ill,从音看,well。所以,“强大语义描述体系”的提法不及“强大概念处理系统”(HNC 句类分析)看得透,因为后者包括对语音冗余现象的清除。大脑反正是这么干的,你得向这个方向前进。

下装“了”体现了音义两方面的需要,所以它是最特殊的助词,可作助词之助。音串中间的“了”一定是 E 的标志,因而可形成规则。“的”的否定作用只可否定该动词的“级别”,不是整句的 E,但不能否定它作为动词的“资格”,它仍然是局部的 E。

HNC 扩展助词之定义,统称 hv。它是两类下装之一,是汉语的特色,戏称内裤,不雅而贴切。

hv 的绝大多数是单音词。HNC 择其精华,构成 HNC 知识库的一个栏目,近日刘薛、陶已完成此项工程,层选处理必须充分利用这一信息资源。

HE 是另一种下装,各语种都有。

HE 是对 E 块基本特性的说明,这里的基本特性是指以基本概念表达的属性,各种 E 块都可能具有,与句类无关,其物理表示式为 K(jm)。这是定义,是一个命题。但逆命题“K(jm) 就是 HE”不成立,因为,K(jm) 既可充当 HE,也可充当 fK 和 JK 或它们的块素。此话本不必说,但常出现类似的误会,故防患于未然。

K(jm) 的定义是“以基本概念为要素的短语”。用于 HE 的最常见 K(jm) 是“时数质度”,用于辅块的常见 K(jm) 是时间和空间。

hv 是发现 E 的重要线索,对单音动词更宝贵,是层选处理和段接处理的重要武器。

HE 是为了净化语句物理表示式而引入的概念,不仅有益于思考,也有益于对规范格式的发现。但对于广义效应句,它将带来两可的困扰。但容许程序两可,不就万事大吉了么!52 个论题的最早清单中,此文有副标题‘年方二八,裙裤咸宜”,就是此意。这里把广义效应句比作少女,因为语言的进化必然以广义效应句为先也。

1998 年 6 月 17 日

论 JK 构成及其分离

只要数学的命题是实在的,它就不是可靠的;
只要它是可靠的,它就不涉及实在。
——爱因斯坦《相对论侧记》

如果有人问:语义块的想法和语句物理表示式的想法,哪一个先出现?我要说,"物理表示式"这个词语肯定晚了好几年,但一个想法(概念)在脑子里已经形成的时候不一定需要相应的词汇,"语句的物理表示式"就属于这个情况。在 1990 年到 1991 年的一段时间里,我追求的目标,我日夜思考的中心就是建立现在叫做"语句物理表示式"的东西,从这个过程看,应该说是表示式在前,语义块在后。但实际上,也许说两者几乎是同步产生的更符合实际。

没有语义块的概念,就没有语句表示式;反过来,没有语句表示式的概念,也就没有语义块。建立语句表示式的关键是确定语义块的"序",这个"序"显然扑朔迷离,篇首引用的爱因斯坦怪论,是打破这一扑朔迷离的思想武器。只有先认定"序"的存在才能有所作为。在这个前提下,广义对象语义块(JK)、辅语义块(fK)和特征语义块(E)的概念是必然的推论。这三类语义块的划分是语言逻辑概念网络设计的灵魂。这个网络曾有三个版本,但与语句表示式直接有关的局域网络是始终不变的,道理就在这里。同时,它也再次印证了上述"一个想法(概念)在脑子里已经形成的时候不一定需要相应词汇"的说法,因为 JK 的概念是后来才标明的,但它实际上早已存在。这里我还想补充一句,越是深刻的想法越是这样,因此,我对"语言是思维的外壳"的提法是持保留态度的,刘志文先生忆及此否?

广义对象语义块(简记为 JK)通常都具有内部结构,因此,JK 也需要相应的表示式来表示这个内部结构(简称构成)。为此,需要引进 JK 构成基元的概念,从这一点开始,我就同菲尔墨先生分道扬镳了。本来我们就不在同一条道路上,他在追求对各种短语的语义命名,我在追求语句表示式。然而,我的语义块命名与菲尔墨的格命名是遥相呼应的。可是,当我意识到 JK 的基元与句类的结合才是菲尔墨的格时,我豁然开朗,困扰菲尔墨先生的完备性问题对我已不复存在。而菲尔墨先生的历史性工作可以划上句号了。

读者或许对我写的 HNC 论文都很少引用文献感到吃惊。这里我想说一句,如果在 1990 年到 1991 年期间,我按常规及时写出论文,肯定会引述一些当时还记忆犹新的文献。可是在事过境迁,我离开他们越来越远以后,就无意于回首往事了,因为那仅仅是回首而已。加之我历来不记笔记(包括在中学和大学的学习期间)的坏习惯,更使这一回首成了一种雪上加霜的负担。但是,菲尔墨等先生我始终是念念未忘的。

JK 的三构成基元 A，B，C 的 A，B 来于语法学的主语和宾语，或语义学的施事和受事。在菲尔墨先生‘格’的引导下，语义学家后来又发明了很多的‘事’，即语义角色，我国的鲁川先生是其集大成者。但是，在 HNC 理论明确 JK 的物理表示式应该采用‘句类符号 + JK 基元符号’的表示结构以后，把 JK 的语义角色作为句类的函数来处理，于是，JK 语义角色的完备性问题就同语句物理表示式完备性问题统一起来了。这个问题已随着语句物理表示式的穷尽发现而划上了圆满的句号。

关于这一穷尽发现，有的同行要求拿出证明，否则就‘宁可信其无’。对此不妨引用一段英国数学家哈代的一段名言，作为本文的答复。

> 严格说起来，根本没有所谓的数学证明，……归根结底，我们只是指出一些要点，…李特伍德和我都把证明称之为废话，它是为打动某些人而编造的一堆华丽辞藻，是讲演时用来演示的图片，是激发小学生想象力的工具。
>
> 转引自克莱因《数学：确定性的丧失》
> 中译本 1997，p.323

“句类符号 + JK 基元符号”形成语义块的物理表示式。式中的句类符号和 JK 基元符号理论上都可以是复合形式，即这些符号的连用。句类符号连用表示混合句类，这一连用是有序的，XR 与 RX 的意义不同，XR 表示作用对关系的影响，而 RX 表示关系对作用的影响，其他类推。JK 基元符号的连用则采用约定的天然顺序，即 A 先于 B，B 先于 C 的顺序。

对基本句类，句类符号无连用，JK 基元符号有连用。对混合句类，句类符号连用仅用于特征块，JK 仍然采用原来的基本句类形式，这是一项统一约定的形式省略。但必须明白，其物理实质是不能省略的，同一个 JK 物理表示式，对基本句类和混合句类的物理意义是有所不同的。

对 JK 的基元 C，读者应深刻领会【2】中所阐述的各种特性，联合攻关组成员不妨回忆一下我一年来关于 C 的各种‘奇谈怪论’。这里我想说的是，从 C 的立场来观察语法学关于词性与句法成分对应的说法，倒是真有点奇谈怪论的味道了，因为这表明在语法学的心目中根本没有 C 的概念，这是多么违背语言的实际，甚至可以说是基本常识呵！为什么语法学竟然在如此长的时间里对此视而不见？这只能从西语的形态，从西方人对西方语言的过于自信中去寻找答案。就自然语言理解来说，西语的形态及由此而来的语法学是否起了一叶蔽目的消极作用？我认为计算语言学的发展历史清楚地表明了这一点，句法分析到处都碰得头破血流之后才领悟到要注意语义，而且还坚持必须先句法后语义的观念。俞敏先生曾说过，章黄学派用古老的弓箭射中了现代学者用大炮未能击中的目标。这个话是陆宗达先生对我父亲说的，当时我正好在场。可惜这几位先生都已仙逝，无从了解俞先生说此话时的具体背景及其深意了。上面的话丝毫没有否定语法学，包括汉语语法学巨大成果的意思，这些成果我们要积极继承。但是，也应该看到，这些成果终究只涉及语言理解的技术性或局部性问题，于根本性问题无补。

C 的中文名字是内容,英文名字可用 Content,第一个字母建议大写。可惜中文没有这个方便,所以我经常宁可用 C 而不用内容二字。有些句类的某些 JK 必含内容,而大多数 JK 只是可能含内容。语义块物理表示式的统一约定是:必含内容者才明写 C,否则一律不明写 C。

【2】中说“两类对象,两类表现,表现与对象的融合性,果表现的语句扩展性,这四点,是形成 EABC 四种主语义块概念的理论依据”。这段话实际上已明确无误地包含了现在叫做块扩和句蜕的思想,也包含了最近在论题 2-2 中阐述的动词分离抗干扰方案的规则。

JK 的构成有三方面的意义:

1. 对象内容分解。例如,B = BB + BC, A = AA + AC, RB = RBB + RBC, YC = YCB + YCC, XBC = XBCB + XBCC, XBC = XBCC + XBCB, …
2. 对象的类型分解。例如, B = XB + YB, RB = RB1 + RB2, …
3. 数学上的形式分解。K = KQ + KH, KQ = KQQ + KQH, KH = KHQ + KHH, …

不同句类各语义块物理表示式的构成特性,是句类知识库的基本内容之一,也是词汇知识库的重要内容之一。

虽然爱因斯坦先生有‘只要它是可靠的,它就不涉及实在’的深刻体会,但语句物理表示式的可靠性终究需要 JK 的封闭性的支持。本文最后要说的话是,汉语标准格式的 JK 确实是封闭的,汉语处理应为此感到幸运。而非标准格式下的分离,实践已表明并不难处理。

1998 年 6 月 20 日

论 辅 块

7.1 引 言

在这个论题下应该首先阐述辅块的定义及其意义,辅块的辨识(或感知)及其指示标志,辅块的类型及其特性等。这些问题以前都有所说明,厌恶重复的心理抑制了我写此文的兴致。但此文又不能不写,因为有许多模糊观念有待澄清,因此,下面将以答问的方式阐述一些基本问题,包括上列几点,最后,谈一下自足性问题。

7.2 辅 块 答 问

问 1:菲尔墨的格有必需、任选和禁止之分,您定义的语义块有主辅之分,是否可以认为,语义块的主辅大体相应于格的必需与任选?

答 1:对于学术评论,我喜爱一针见血的方式和语言。当然,这里说的一针见血是指透彻性,而不是简单化的贴标签,对这一喜爱值得谈一件往事。大约是 1946 年的冬天,我随同父亲去看望当时的武汉大学哲学系主任万卓恒先生,万先生住在珞珈山上叫做‘半山楼’的两层楼房里,四周林木森森。先生一生未婚,住室内外的环境和布设,先生的气度,使人产生一种如同面对一位高僧的感觉。那一次万先生谈起黑格尔对孔子、老子和庄子的看法,我的印像是,黑格尔对中国这三位先哲的评价都不高,尤其是孔子,但万先生对黑格尔评价的评价很赢得我父亲的赞许,具体内容已印像不深(当时肯定也没有完全听懂),但是,对万先生关于‘尖刀和抹布’的比方说法却印像极深。接下来,两老(我父亲当时 46 岁,万先生应不超过 50 岁,不过当时的习惯称呼是耀老和卓老,我父亲字耀先)以尖刀和抹布为‘识’的两极标准评论了很多历史人物。这一段少年生活的花絮在我的意识里种下了尖刀和抹布的深刻印像,不知不觉之中对我的思考和写作习惯产生了无所不在的影响。

关于菲尔墨的格理论,我以尖刀的方式在论题 13 中有所说明。我想,你的这个问题可以在那里找到答案。

问 2:先生曾反复强调‘两可’的概念,语义块的主辅之间也存在两可,两可大体上相当于集合论的交集,如果直接定义一些两可型或交集型概念节点,岂不对软件操作大有帮助?

答 2:两可的疑难不是定义问题,不是静态问题,而是智力运用问题,是动态问题,是动态的模糊消解问题。著名心理学家皮亚杰说过,智力是你不知怎么办时动用的东西,我欣赏这个说法的实质,但不喜欢这种西语(皮亚杰是瑞士人)的陈述方式,汉语的成语‘见机行事’

是更传神的表达,因此,我用‘见机行事的创造性处理方式’来表达两可处理的基本特征。我概括的语言五重模糊有其静态和动态两个侧面,为了强调动态侧面,又提出两可这个术语。

概念的抽象具体之分及概念节点的系统设计,语义块的主辅之分及语义块基元的系统设计,基本句类的划分及其一级子类的系统设计都存在两可问题,正是处理这一系列两可疑难的艰辛使我对利用机器可读词典寻求语义原语的效果深表怀疑,因为问题的要害不在于寻求最小原语数量,不是一个简单的义项归并问题,而是要寻求自然语言概念体系的理论模式,这里需要深刻的创造性思考,统计方法只能提供工具性帮助,不可能替代思维的创造性作用。

这里顺便说一下,语义原语的英文是 semantic primitive,HNC 译为语义基元,将 primitive 译为原语不妥,因为,语义块的分类,抽象概念多元性表现的分类,甚至任何分类,都存在寻求 primitive 的问题,HNC 定义的 A, B, C 语义块就是主语义块的三种 primitives,五元组(v, g, u, z, r)是抽象概念多元性表现的五种 primitives,HNC 把它们分别命名为语义块基元和概念类别基元的一种——形态基元,这比较确切。如果命名为语义块原语和概念类别原语,岂非别扭之极?

这里不妨说一句多余的话,两可之‘两’实质上是‘多’,这同‘2 + 2’句式的‘2’是‘多’一样。

在处理上述一系列两可疑难的过程中,曾试图引入疑难度的概念予以定量表述,后来放弃了。因为,这需要统计工具的支持。因此,我虽然不赞成语料库学派的大方向,但我一直十分关注他们的方法学成果,HNC 理论和技术的发展不久的将来必须在 HNC 知识库和 HNC 语料库的基础上开展这方面的研究。两可问题的彻底解决需要 HNC 联合攻关队伍的迅速壮大,需要当前核心队伍的迅速成长,遗憾的是当前的小环境不能保证 HNC 发展所需要的基本条件。至于我个人,这里不妨转载一下丁丑除夕夜我写给家妹的一首小诗:

我也楚狂人,生平不步尘,此身惟我有,无意问营营。

此文读者可能不习惯这种五言唐诗,允许我稍加注释。第一句是抄袭李白的诗‘我本楚狂人’,后两句是抄袭苏东坡的诗,苏诗原文是‘长恨此身非我有,何时忘却营营?’‘营营’者,名利也,这里我反其意而用之。诗是抄袭的,狂劲是浅薄的,但刻画是真实的。

你提出了一个根本性的问题,但深度很不够,所以我作了上面的发挥。

问 3:先生对两可概念作了很好的阐述,但我还希望听到您对主辅块两可问题的具体说明。

答 3:主辅块的两可源于对语义块的主辅之分,如果对语义块不作这一区分,也就不存在主辅块的两可疑难。那么,为什么要提出语义块的主辅之分?为了建立语句的数学和物理表示式,使语言变成一个 well-defined 的东西。这些表示式就是计算语言学界所孜孜以求的语言模型之一——句子层面的语言模型。

关于这个模型问题,可以说存在两种态度,一种是得过且过,在短语结构模型的基础上

修修补补,不去触动它的根本缺陷,希求通过受限的约束避开语言的种种不规范现象,也就是避开对语言本质的探索。另一种是相信乔姆斯基关于自然语言是一个 ill-defined 的东西的说法,脑子里存在大量比喻的和夸张的,乡土的和诗歌的,儿童的和怪诞的例句,并为之困扰而不知自拔,不相信对自然语言的表述可以出现牛顿力学对力学现象或麦克斯韦方程对电磁现象的突破,但是他们不曾想过,如果当年牛顿不是专注于天体的运动,而是专注于羽毛在狂风中的飞舞,麦克斯韦不是专注于电磁场在自由空间中的一般规律,而是专注于方孔的衍射,他们也将一事无成。在建立自然语言模型这一重大探索中,必须紧记在所必为和有所不为的辩证法,并深思康德的下列两段名言:

> 理性必须一手拿着原则,拿着那些唯一能使符合一致的现象成为法则的原则,另一手拿着自己按照那些原则设计的实验,走向自然,去向自然请教,但不是以小学生的身份,老师爱讲什么就听什么,而是以法官的身份,强迫证人回答他所提出的问题。
>
> 自然的最高立法必须是在我们心中,即在我们的知性之中,而且我们必须不是通过经验,在自然里面去寻求自然的普遍法则;而是反过来,根据自然的普遍的合法则性,在存在于我们的感性和知性里面的经验可能性条件中去寻求自然。

我正是遵循着这样的方法和信念,开始进行语义块和语句的物理表示式的探索。在这一探索中,面临的第一个困扰就是一个语句语义块个数的确定性与不确定性两个侧面的同时存在。抓住确定性因素,把它作为联想脉络的主体,将不确定性因素作为联想的支脉,另行处理,这就是 HNC 的策略思想和语义块主辅之分的由来。联想脉络意味着有序,有序就意味着各语义块应有自己的位置,沿着这一思路就不难发现,确定性因素在语句中具有位置确定性,而不确定性因素具有位置不确定性,这进一步加强了语义块主辅之分的科学性与合理性。

但是,主辅语义块的区分只是一个起点。关键性的飞跃是在关于主语义块类型的思考中:对特征语义块 E 的类别基元和广义对象语义块 JK 的类别基元的发现,后者是前者的函数的发现,从而得到主语义块是句类函数的结论。这个结论标志着 HNC 理论对第二个理论模式,即语义块和语句物理表示式的探索已进到"蓦然回首"的境界。

"回首"以后的大量配套工作多数以细节为主,理性"法官"的工作尤为烦琐,这些都与我的性格很不相宜,加上伴随"回首"而生的一种心理上的惴惴不安,使我疏于战术和细节的弱点更加膨胀,语言逻辑概念 14 以后残缺不全的状态就是这样形成的。

当然,在总体布局方面我是反复推敲极为慎重的,严格遵循"待定"不能影响"已定"的原则,并自信对此可以确保。然而,正是这种自信使我对某些"待定"过于漫不经心,主辅语义块的两可问题就属于这一情况。

对于两可问题的处理,理论上和技术上的要求是有所不同的,理论上必须透彻,但技术上可以从权。我在"给萧友芙老师的信"中说:"在工程意义上,辅块也可以这样定义:辅块是

句类代码之外的语义块”,这就是从权。当然,由于省略格式和语义块分离现象的存在使上述从权策略只有在无省略和无分离时才有简明的实际效果,根本的解决之道仍然是语义块本身的检验。

上面的说法是‘非主即辅’的从权法,反过来,也可以采用‘非辅即主’的从权法,这两种方法可以交替使用。这应该成为软件语义块主辅辨识的基本策略。实际的音串,有时是主块容易辨认,有时是辅块容易辨认,更多的情况是辅块容易辨认,这一类的见机行事应该不难实现。因为:

第一,辅块的位置具有‘不定之定’的基本特征,这个‘定’就是它绝不会出现在 E 块之后,这是汉语的一个十分可爱的特点。喜爱收集语言‘鬼怪’的语言爱好者,可能会找到否定这一特点的特例,但是,我们不能因为‘鬼怪’的存在而丧失‘自然的最高立法必须是在我们心中’的信心,上帝保佑,不是所有的人都被‘鬼怪’吓破了胆。

第二,辅块之是否带标记不理会句类格式之变,这是辅块之有别于主块的又一个极为可爱的特点。

第三,不带标记的辅块必然是以 K(jm) 为内容的条件辅块。其中主要是时间条件辅块。

以上所述,是对主辅语义块两可疑难外围的清除,于是,两可疑难便集中在带有 l18, l19, l5 标记的语义块,而硬骨头是 l18。更具体地说,这块硬骨头就是对音节 wei 的感知处理。因此,对‘wei’的处理是仅次于‘xiang’的第二号特殊处理问题。

问 4:先生的这一番论述我需要认真思考,先生能否对语言逻辑概念节点 l18, l19, l5 作一些具体说明?

答 4:严格说来,这不属于本论题系列的范畴,而是正在编辑的基本资料《HNC 概念符号体系手册》的任务。不过,我愿意在这里谈一下‘辅块的类型及其特性’这个题目,并且先介绍一下概念节点设计中的一些重要思路,这对于该手册的编辑也会有所帮助。

HNC 的 HN 不是一般的层次设计,一般的层次设计仅仅考虑子类的划分,如同生物学的分类。但概念的 HN 不能这么简单,它不仅具有层次结构,还具有网络结构和其他结构,这就是我把它命名为‘层次网络’HN 的原因。层次的数学表达比较简单,但网络的数学表达则极为复杂,不可能用表示式来体现,但重要的是,必须在表示式里有所体现。这就产生了高中底三层次表达的思想:高层表示层次性,中层表示其他特性,底层体现网络性。我在【1】的结束语中说:“底层设计是一项复杂的系统工程,我们寄希望于与语言学家及同行们的合作”。这句话依然具有现实性,不过希望已经落实,它历史性地转移到联合攻关组的肩上了。

这里所说的中层结构就是【1】中所概括的对偶性、对比性和包含性。在高层与中层、中层与底层之间,推而广之,在本体层与挂靠层之间,最简单的办法是设置层间标记,但模拟大脑的强烈意识和初期的 4 比特观点使我从一开始就否定了这种简单办法,而采用规范化和数字约定的办法。这一举措的功过还有待评说。

所谓规范化就是规定高层的层数。大家现在所看到的层数 2、3、4 的简明结果实际上经

历了十分复杂的过程,它包括节点的排序和序中隐含的局域网络知识。这些,在【14】到【20】中有所说明,但不够系统。我打算在本论题系列完成以后,对〈论文(Paper)〉系列加以整理。

隐含局域网络也体现在辅块类型的设计,即 l1 二级节点的设计中。l1 原来有三组局域网,l11 到 l13 为第一组,l14 和 l15 为第二组,l16 和 l17 为第三组。后来增加的 l18 和 l19 可视为第四组。第一组是纯粹的辅块,从第二组开始,就不那么纯粹了,而且越来越不纯粹。这种过渡特性存在于作用与效应之间,主动与被动之间,基本句类与混合句类之间,块扩与句蜕之间,总之,存在于一切具有对偶或对比特征的事物之间。对于这类过渡现象,要有"定"与"不定"的两手,要见机行事,"不定"则粗,可"定"则细。辅块类型,细分为九,粗分为四,更粗为二(辅块与两可块),这就是我的建议。

1998 年 5 月 31 日

带括号式指示符的辅块

语言逻辑概念的 l0 到 l3 是语义块的指示符号,放在所指示语义块的前面,在 HNC 的早期文献里,曾叫做语义块切分指示符。与此相对应,把语义块内部组合的指示符号,如“的、之、和、及……”等,叫做组合指示符号。后来觉得‘切分指示符号’的切分二字有歧义,把前者简称指示符,把后者简称组合符。组合符各语种都有,指示符则不然,汉语和日语都比较发达,而西语阙如。不过汉语和日语的指示方式有所不同,汉语放在语义块的前面,它的左边是另一个语义块的尾。日语则放在语义块的后面,它的右边是另一个语义块的头。日语指示符主要是语法意义,汉语指示符还有语义‘格’的信息。

语义块指示符是规范格式的需要,标准格式不使用指示符,违例格式省略了应有的指示符,因此,违例格式和 4 主块句的标准格式存在所谓 BC 模糊。

汉语的第一个 JK 一般不加指示符,这符合理所当然法则。但是,如果第一个 JK 的前面有辅块,这个理所当然就站不住脚了。这时,汉语在辅块的最后另加一个结束指示符,以间接指示第一个主块的开始,这就是括号式指示符号的起因。例如,“在公园里孩子们玩得很开心”,这里的‘在’与‘里’搭配形成一个括号式指示符号。

成对的括号式指示符是汉语的一项创造,如:

在	里,上,中,下,内,外,旁,……
	之(上,中,下,内,外,东,西,南,北)
除,除了	外,之外
对	来说
像	一样,那样,似的
跟,和,同,与	一样
从	中
……	

汉语括号式指示符的形成,是进化自身的天才设计,这没有什么可奇怪的。这里需要申述的是,汉语的许多个性或特色,经过西语有色眼镜的过滤之后,可能十分模糊甚至消失,语义块指示符号就是一例;句类转换、块扩和句蜕是第二例;音形义三极而不是音义两极,从而每个音节都可能是单音词,是第三例。汉语特色并不都是理解处理的困难之源,也蕴涵着宝贵的甚至是关键性的信息。我认为,语义块指示符就是这种关键性的信息。

本文最后需要说明一下本论题序列对有关术语使用的一种偏向。HNC 论述中很少使用主语、谓语、宾语、状语、定语、补语、短语等术语,这并非 HNC 排斥这些久经考验的经典术

语。我在【2】中说过“EABC 是语言深层的语义描述量,是句类的函数,但与句类格式无关。主谓宾补恰恰相反,它是语言表层的语法描述量,不管句类,但与句类格式息息相关。……两者从不同层面或角度对句子的结构提出分析的模式,不能相互代替”。这里需要补充的是,两者虽不能相互替代,但应该朝着相互补充的方向努力,因此有必要开展两者的对比研究。为此,我们已发出招聘相应访问学者的信息(见 1998 年 4 月 14 日《人民日报》海外版、《光明日报》)。

回到术语的使用,本论题系列由于一种习惯的力量,往往将语义块全盘代替短语,这是错误的。以本标题为例,就应该在辅块后面加上短语两字。上面示例中的括号标记‘像,跟,……一样”,用短语括号命名,更为恰当。

1998 年 6 月 24 日

论主辅块变换

HNC 理论为建立语句的物理表示式所采取的一项重大举措就是先区分主辅两大类语义块(也可以说先区分 E、JK 和 fK 三类语义块)并将辅块排除在语句的物理表示式之外。这一举措的根据,在一般理论意义上可以这样说,对陈述句,主块是句子联想脉络的主体,而辅块只是联想脉络的支脉。从汉语的实际情况看,标准格式下各主块有严格的排列顺序,都不带语义块指示符,规范格式下两相邻 JK 之间必须插入后一个 JK 的指示符。而辅块的位置则可在 E 块前随意移动,它是否带指示符与格式无关。

按照上述说法,似乎存在两个问题,第一是语句物理表示式是否只适用于汉语?第二是该表示式是否只适用于陈述句?

第一个问题应该在论题 13 或 14 中阐述,两文已提前写就,未涉及这个问题,因此在这里补说。语句物理表示式反映语句的主体联想脉络,表明该语句所隶属的句类应该配置多少个具有何种基本结构特征的 JK,这一本质与语种无关。因而基本句类的物理表示式应该适用于所有语种。混合句类的情况略有不同,其一级子类达 3192 种之多,各语种的钟爱不可能一致,因此某语种所钟爱的混合句类可能在另一语种中根本不存在,这将造成两语种互译的困难。当然,基本句类也有这种情况。这是不同语种之间的句类转换问题。将 HNC 理论应用于机器翻译,首先就要对这个问题作深入研究。基本句类的转换容易把握,混合句类和复合句类的转换可能存在难以预计的难点,它还涉及不同语种的 E 块主体构成存在巨大差异,不是简单的动词映射关系,初期的反映射库不可能提供所需要的全部信息。但是,句类转换的难点与语句物理表示式的普适性无关,两者的关系,是所谓河水不犯井水。

第二个问题涉及本文要讨论的主辅块变换问题。语句表示式是按照陈述句设计的,但祈使句和感叹句不过是陈述句的带有特殊信息的省略格式,疑问句也只是陈述句的带有特殊信息的格式变换或省略,包括辅块变换为主块。它们根本不影响语句物理表示式的意义,同样是河水不犯井水。读者从这两次"河水不犯井水"中可以体会一下,所谓深层与表层或本质与表象在思考空间将呈现出多么巨大的差异和截然不同的景象。这种差异和景象在许多情况是只可意会而难以言传的。

上面说的特殊信息包括下列特殊词语和特殊符号:

1. 疑问句的疑问词。
2. 疑问句文字文本的疑问符号"?",语音文本的疑问语调。
3. 感叹句的感叹词。
4. 祈使句和感叹句的惊叹号"!"或惊叹语调。

这些特殊词语放在“语法”概念节点 f42，f43 和 f5。

疑问句中的辅变主极为常见，英语中所谓“6 个 W”(where，what，when，why，which，who)的提问，前 5 个都会涉及辅变主。

疑问句的辅变主当然只是主辅变换的情况之一。其特点是不影响原来的主块。汉语尤为简明，根本不改变原来的语句格式。西语疑问句复杂的词序变化汉语一概不予理睬，以不变应万变，不也照样完成了提问么！

下面的例句代表了主辅变换的另一种类型：

“在中国经济发展中出现了这么一种引人深思的现象。”

它既可以说是辅变主，也可以说是主变辅。

陈述句中当然存在单向的主变辅或辅变主，但注意不要搞扩大化。

这些都是张艳红硕士论文的范畴，本文不来讨论。

值得指出的是，语义块是与语句表示式相联系的概念，向上(句子)为块，向下(词)为块素。当方向不定时，叫短语为妥。

主辅块的两可问题不属于主辅变换，并已在论题 7 中讨论过了。

1998 年 6 月 25 日

论语义块与短语

本文可用 9 个字概括它的全部内容:让语义块与短语共存!

语义块和短语这两个概念分别代表了从不同角度对词或词组的观察,语义块代表语义层面,而短语代表语法层面。两者可相互补充,而不能相互替代。semantic chunk 与 phrase 各管一个层面,双方可共存共济,但不要互相侵犯,这有利于思考和知识表示。

把应该是短语的东西叫做语义块是一种侵犯,基本概念语义块的说法就有这个情况。因为由基本概念构成的词组可以是语义块,也可以是块素,因此可对其沿用语法学的术语"短语",所以还应该引入短语符号 Ph。为什么不沿用句法分析的国际通用符号 P 来表示短语呢?因为 HNC 已将 P 用于过程句的 E 块符号。这样,知识库中目前对 HE = K(jm) 的表示方式应改为 HE = Ph(jm)或 FK(jm),而在 K 调度中引入的符号 fFK 应改用符号 fPh。

"语义块是语句的下一级语义构成单位"(参见【2】)。对语义块又定义了要素和块素两个概念。要素是语义块的核心,符号有 KH, KHH...JKH, JKHH...fKH, fKHH...,后两种分别表示广义对象语义块和辅块的核心。这种表示方式要求核心殿后,它适用于汉语,但不适用于英语,因为汉语严格遵守上述要求,而英语不遵守。汉语的这一特性为句类检验带来了无与伦比的方便。

从形式上说,符号 KQ, KH;JKQ, JKH 可以只表示块构成的前后信息,不一定要与修饰成分和核心发生联系。但由于汉语语义块构成采用核心在后的良性结构,如果不对这一重要特征予以明确标记以便于计算机的利用,显然不智。故 HNC 赋予上列符号以"核心"的特定含义。

块素定义为语义块的一部分,符号是 FK。部分不是全体,但又可以代表全体,这是符号 FK 的原定意义,它容许两可。这样,对于确定地不能代表全体的情况就缺乏表达的手段了。

语法的概念和术语必然具有表层确定性和深层不确定性的特点,这一特点在许多情况十分有用,动宾、主谓、偏正、短语等等都是如此。HNC 将动宾分成 vB 和 vC 两种基本类型,这对于确定概念联想脉络的基本类型极为重要。但是,当这一联想脉络根本不存在或这一类型的划分不重要时,就需要一个简明的表层术语了,宾语恰好适应这一需要,而 HNC 的 JK2 反而显得模糊了。符号 FK 完全可以用来确定地指示语义块的一部分,不容许两可。为了指示两可,需要另行引入一个符号,然而这是现成的,只需引进,不必引入,这就是短语符号 Ph。语义块和短语分别从语义层面和语法层面"对句子的结构提出分析的模式,不能相互替代"。【2】曾多次强调这一点,可惜后来我这个提出者也有点淡忘了。

问题当然不仅在于个别概念或符号的借用,而在于对语法知识的继承。为此 HNC 专门

设置了 f 类概念。但是,f 类概念仅能包容词汇层面的语法知识,并不能全面表述短语和句子层面的语法知识。HNC 创立了一整套从词汇层面到句子层面的符号体系,但没有全面考虑与已有语法符号的兼容性问题。这违背了改革必须与继承相结合的基本准则。对于汉语的 HNC 理解处理,由于可以直接走"中间切入,先上后下"之路(参看论题 1 及 1-1),句法分析可继承的东西不多,兼容性问题基本上可以回避。但是对于西语,恐怕要另当别论。这是一个重要的课题,我们计划邀请高级访问学者前来进行预研。

回到短语符号 Ph 的引进,它密切关系到非动词的 HNC 知识表示,关系到今后是否采用新标准和是否对当前汉语 HNC 知识库中进行调整。我的意见,立即采用新标准。非动词的 FK 表示有些要改成 Ph,有些仍需沿用 FK 表示。一个非动词词汇是否能充当要素主要决定于它的概念类别并看它面对什么句类。对此,应尽可能分别给出明确的信息。Ph 和 FK 的不同表示有利于这一信息的区分。

北京大学(俞士汶,1995)、山东大学以及国内的许多单位都曾对汉语词语(包括短语)的分类作了大量的工作,已委托陶明阳对传统的分词与 HNC 的概念类别进行比较研究。这项工作应扩展到传统的短语分类,我希望近日就这个问题"沙龙"一次。本文仅谈到引进短语符号 Ph,是否还有必要引进其他的语法符号?调整知识库是件大事,不能零敲碎打,研讨一次,集思广益,全盘考虑,慎重决策。

1998 年 6 月 26 日

基本概念短语

本论题要讨论的基本概念短语就是 HNC 前此论述中的基本概念语义块,论题 10 已对这一必要的正名作了说明,这里就不重复了。关于基本概念短语,两年半前曾写了 4 篇短文,现作为续篇放在本文的后面(即论题 11-1　11-4)。为保持历史原貌,不作任何改动。这 4 篇短文本来是‘HNC 理解处理论文选录’中【19】“基本状态句基本判断句的句类知识”一文的 4 节,都是在基本状态句的前提下,对基本概念短语进行讨论,称之为基本概念语义块是适当的。但是应该指出,这些语义块不仅可以充当基本状态句的 SC,也可以充当其他句类的语义块或其构成。因此称为基本概念短语更为适当。但文中的块符号 K 不变,块素这个术语一律照用,不必改成短语素之类。

基本概念短语应成为一类特殊的激活信息。论题 31 曾强调指出:“由基本概念构成的语义块,特别是数量、时间和空间语义块需要先行处理,不需要也不应该等到句类检验之后。……块内处理并非总是三部曲的第三步,它既可以操作于语义块感知阶段,也可以与句类检验同步进行。调度程序必须充分适应这一要求,这就是智能性表现”。

下面,对这一段引文作一点说明和补充,本文就完成任务了,因为详细的论述已见 4 个续篇,不必重复。

需要说明的是,引文中的语义块,如果改成短语,更恰当一些,但文中‘块内处理’的提法绝不能改成短语处理。因为,语义块和短语这两个术语各有严格的定义,在概念上不容混淆。但在实际使用时可以灵活一些,可以容许在它们之间出现相互侵犯或替代的情况。

需要补充的是基本概念短语先行或准实时处理的可行性。

基本概念短语具有良性(WD)、甚至超良性(SWD) 构成,这在几个续篇中屡有阐述。这是基本概念短语可先行处理的可靠基础。对拼音输入指定数字的做法是为了消除这一基础的模糊性。汉语对数的运用过于灵活,不得不出此下策。

对基本短语良性构成的揭示,即写出相应的短语构成物理表示式,必须依靠对概念层面的内在联想脉络给出相应的符号表示,并约定相应的激活因子。4 个续篇对此作了详尽的阐述。当然,它们并不符合软件设计文本的要求,但这一转换工作不应该再要求一个老人来承担了。

林杏光教授(中国人民大学)曾对我谈起有人总结的一组短语处理难点的例子,非常有趣,是非常好的典型语料。这里,我建议苗传江对这一语料进行 HNC 分析,分析结果必将成为一篇具有代表意义的论文。

1998 年 6 月 27 日

序描述句

序描述句定义为说明事物序特性的语句。它是序状态句的一个子类,因而是状态句的3级子类。其基本特征是以概念 jz00 为 SC 要素,以 jv00 为 S 要素,这两项概念约束可视为序描述句的充分条件,但 S 可省略。所以,序描述句比较容易辨识。

让我们先看一段例句。

世界＊上…21…支经济＊上…最…大…的…力量…仍然…是…国家,

美国…居于＊首位＊,

其次…是…日本,

德国…居于＊第…3…位,

瑞典…居于＊第…21…位。

这里连着 4 个描述句,都未省略 S,如果把它们都改成省略 S 的标准格式,也完全可读,如:美国第一,日本第二,德国第三,瑞典(排在)最后,第 21 位。

这些句子确实都很容易辨识,它们都满足上述充分条件,即:S∈jv00,SC∈j00,表现了典型的狭义同行优先特性。

但必须指出,这 4 个序描述句共享一个基本判断句,这个语句提供了 SC 的公共部分 SCQ 和 SB 的公共核心部分 SBH,SCQ = 世界,SBH = 经济力量。这里对序描述句的 SB 和 SC 语义块的构成模式采用了常规的"一分为二"表述方式,即我们常说的"说明 + 核心"方式。但"一分为二"只是语义块构成的分析形式,而不是实质。实质性分析必须进入到概念构成层次。

序描述句 SC 语义块的概念构成可表达为"比较的范围"+"序值"。基本概念"序"来于比较,比较又必须规定一定的范围。比较的范围和比较结果的值就是序描述句 SC 的两项概念构成,这一语义块构成知识与语种无关,是概念层面的知识。但是,这两项构成的排列顺序则与语种有关,属于语言知识,这里我们约定 SCQ ="比较的范围",SCH ="比较的值",是依据汉语的习惯。这个习惯不容违反,把握这一点,才能对"世界第一"和"第一世界"这两种排序里所体现的截然不同的含义有足够的灵敏反应。

序描述句 SB 语义块的概念构成可表达为"比较的力量"+"比较的内容"。在上面的例句中,对象是国家和公司,内容是经济力量。对象与内容的二分构成,是序描述句 SB 语义块的概念构成,同上,SBQ = 对象,SBH = 内容又是汉语的习惯,而且同样具有不可违反性。

语义块构成同语句构成一样,都需要从概念和语言两个不同的层面进行描述,这一基本

思想是概念层次网络理论的精华,将在随后的短文中反复加以阐述。

如果序描述句都老老实实地按照标准的句类格式和语义块构成格式进行表述,那可以说,即使是模糊的语音文本,序描述句的句类分析也可迎刃而解。但语言绝不会那么老实,不老实的典型表现就是语义块构成变换,这一变换的典型表现就是将 SBH 和 SCQ 合并。以例句 1 为例,标准陈述是"美国的经济力量居世界首位",但也可陈述为"美国居世界经济力量的首位",要理解这两种陈述方式的等义性,就必须具有并运用语义块构成变换的知识。对后一种陈述方式,需要多说几句。

与 j00 强交式关联的概念有 z56,55,j11,j21,11,204,这些概念里都含有"序",其中的 z56,z55 在语用上与 jz00 等价,它们也是序描述句 SC 的充分概念。但对于 z56,z55,比较的内容一定放到 SC 中,例如"李小姐是今年北京高考的文科状元",这里,对象是"李小姐",内容是"高考文科成绩",范围是"今年,北京",序值是"冠军"。这句话如果采用标准格式"李小姐的高考文科成绩是今年的北京状元",反而非常别扭。这里,语义块构成变换的陈述方式最为自然,那么,如果对后两类概念定义 SC = (比较内容 + 范围 + 序值)岂不简单明了?但不妥,因为"内容"与"范围"的顺序不像前述的二分顺序那样铁定。因此我主张持语义块构成变换的思路,尽管在某些特定情况下显得走了弯路。(如果死扣词义,"北京文科状元"的说法是不妥的,应该用解元替代状元,但当代人对解元很生疏了,只好以状元代之,这就是词义的发展。)

对序描述句的上述初步分析表明,词汇、语义块、语句、句群的概念层面研究确实大有可为。计算语言学必须把自己的立足点转过来,端正主攻方向。在这一转变中,西方语言学的语法传统是一块绊脚石,而所谓的语料库语言学则是一块误导的路标。对语法和语料库的所能和所不能要有一个清醒的认识。

最后,谈一下例句"其次是日本",它是语义块的搬移还是交换?我倾向于向空间描述句看齐,采用交换的观点,理由仍然是,搬移要加语义块指示符,而这里是永远不加指示符的。

语料:

据…了解,

在…全国已经 * …交付 * 使用 * …的…104…米…以上 * 高…楼…中,

上海 * …有…76…幢,

广州…有…19…幢,

北京 * …有…16…幢,

深圳…有…12…幢。

(其中)…北京 * …的…京广 # 大厦(以…208…米…)

…雄踞 …国内…高层…建筑 * 之首 * ,

广州…国际 * 大厦…则…以…67…层…成为 * 国内…层…数最多…的…建筑。

时 间 描 述 句

时间 j1 是有序的和一维的,这是时间的根本特性。除时间以外,其他的基本概念都是多维的,包括 j0,j0 的多维性表现为比较内容的多面性,例如 11-1 中的例句“美国的经济力量居世界第一”,是以“经济力量”为基的序值,如果换成人口或面积为基,则美国的序值就不是第一,而是第三和第四了。

时间状态句的标志是 :SC 为时间描述语义块。如果 SB 和 SC 都是时间描述语义块,就是本文要讨论的时间描述句。

由于 SB 和 SC 都是时间描述语义块,时间描述句自然具有 SB 和 SC 的可交换性。

时间描述语义块是本文讨论的重点。我们将详细说明它的构成知识,并从时间概念的表示谈起。

11-2.1 特定时间和特殊时间

首先需要引进特定时间和特殊时间的概念。时间描述句经常用来描述特定时间和特殊时间的关系,这时,如果 SB 表达特定时间,则 SC 一定表达特殊时间,反之亦然。

什么是特定时间?抽象地说,就是时间序列中一些特定的时刻或区间,其 HNC 符号是 l91/j1,但这些无益于计算机的理解。计算机需要的是一组能激活相应联想的符号,因此,我们把它定义为 wj10- 和 pj11- 。

符号 wj10- 为什么能激活概念联想?因为它包含了“年、月、日、点、旬、白天、晚上、上午、下午……”等等丰富多彩的特定时间概念。如下表所示:

年	wj10-
月	wj10-0
日	wj10-00
点	wj10-000
旬	wj10-0-
上旬	wj10-0-c31
白天	wj10-00c21
晚上	wj10-00c22
上午	wj10-00c211

下午	wj10-00c212
中午	wj10-00c210

在这些 HNC 符号里,直接的激活因子是中层符号“-;cmn,dmn;emn”,间接的激活因子是类别符号 wj1。前者激活包含性、对比性和对偶性联想,这些联想是同行优先的体现之一,是语义块感知处理中最基本、最常用的知识(这些都是老生常谈了,这里不厌其烦,只是表达我的一种渴望而已)。后者则激活周期性的联想,但这一知识目前是放在概念关联知识库里,激活的灵敏度似乎不够,值得改进。

上面的 HNC 符号不包含日常生活中常用的另一类时间概念,如“星期”、“世纪”、“年代”等,这是不能包含的。因为,上列符号是物化时间,是与日、月、地相对运动有关的时间,而“星期”、“世纪”和“年代”与上述运动无关,是与人类活动有关的时间,是人化时间,其类别符号应定义为 pj1。于是有:

星期天,星期日	pj11-c70
星期一……星期六	pj11-c71…pj11-c76
世纪	pj12-
年代	pj12-0

1 pjzz11- = 7 wjzz10-000

1 pjzz12- = 100 wjzz10-

wj10-、pj11-、pj12- 都是 HNC 对特定时间基元的定义和形式化。这一形式化的精髓在于它包含了明确的概念联想激活因子。应该指出,这个定义是对时间间隔概念 j12- 的补充。j12- 本身包含“时、分、秒”等概念,它们是一般时间间隔基元,不包含自然现象或人类活动的激活因子。

特定时间描述语义块必须以特定时间基元及一般时间间隔基元为要素,它的语义块的构成知识下面详谈。现在,先来说明特殊时间的概念。

特殊时间主要是联系于人类某些特殊活动的时间,如节日和纪念日,特殊时间应能激活相应的特殊活动。按定义,特殊时间的 HNC 符号是 j731/j1,但这个符号里没有显含应有的激活因子,应设法弥补这一缺点。

时间概念本身最基本的特性是它的顺序和间隔,这两项占用了它的两个二级节点,这种安排是天经地义的。激活因子虽然极为重要,但终究不能与上述时间概念的基本特性并列,这就是说,时间的激活因子只能安排在 HNC 时间符号的底层,这似乎也是天经地义的。考虑到特殊时间主要与“年、月、日”相联系,因此,把它与 wj10- 联系起来,显然是明智的。基于上述两点,建议特殊时间的表示采用下列框架:

特殊时间	pj10-8
节日	pj10-008
纪念日	pj10-009

斋月	gc82b/pj10-08
生日	v141/pj10-009
国庆	(pj2‖vc141)/pj10-008
春节	(((ppj21,j781)/wj10-)‖v111)/pj10-009

前三行是特殊时间的概念基元,后面各行是具体特殊时间的示例。这些具体的特殊时间是时间描述语义块的天然核心。

11-2.2 语义块的形式表示

上面,为时间描述语义块构成的阐述作好了知识准备,这里,还需要作一点符号准备,这就是语义块的形式表示。以前在这方面显得比较凌乱,需要加以统一。现建议如下:

语义块	K
核心部分	KH
说明部分	KQ
前缀部分	QK
后缀部分	HK
语义块块素	FK(KH,KQ,QK,HK 的代表)
语义块函数	K(jyy),K(fyy)

前 7 个定义是为了行文和表示之便,也许可用于规则表示。这里,语义块函数的第一变量 j 和 f 分别表示基本概念和语法概念。例如,K(j0)、K(j1) 分别表示序描述和时间描述语义块,其他类推。K(f30) 表示名称语义块,K(f42) 表示疑问语义块,其他类推。

后两项定义用于表示句式知识,以代替语义结构方程的第四级数字表示,这主要用于 C 语义块。

原来定义的语义块语义符号(即 E、A、B、C 以及它们与句类符号的组合)不变,它们仍用于句类格式及语义块语义构成的表示,新引入的符号则用于语义块的形式表述,如:

$$K = KQ + KH \qquad (K.01)$$

$$K = QK + (KQ + KH) + HK \qquad (K.02)$$

$$\langle HNC\rangle = K(\cdots)$$

$$FK = (HNC)$$

式中,〈HNC〉表示语义块的 HNC 命名,即语义块语义符号。K(…)表示语义块函数,(HNC)表示语义的 HNC 表示,即层次网络符号表示。最后的表示式 FK = (HNC)就是语义块块素的 HNC 表示。

11-2.3 特定时间描述语义块的构成知识

下面将依次转入时间描述语义块构成知识的讨论,让我们从一组例句开始。

1. 今天是 10 月 29 号。
2. 1901 年是光绪二十七年辛丑。
3. 从北京到广州的飞行时间大约要两小时。
4. 后天是张先生的生日。
5. 再过九天就立冬了。

这些例句的 SB 和 SC 都是时间描述语义块 K(j1),只是例句 5 的 SB“再过九天”略为有点特别。这些时间描述语义块可分为 4 类:

1. 特定时间描述 K(wj1),如“10 月 29 号”、“1901 年”、“光绪二十七年辛丑”、“立冬”。
2. 特殊时间描述 K(pj10-8),如“张先生的生日”
3. 时间的序描述 K(j11),如“今天”、“后天”。
4. 时间间隔描述 K(j12),如“从北京到广州的飞行时间”、“两小时”。

本节讨论特定时间描述语义块的构成。

特定时间语义块可给出下面的 HNC 表示式。

1. 特定时间宏观表示式

$$KQ = ppj1\,;f30j1 \qquad (J1.01)$$

$$KH = \sum(\sum j308 + wj10-)\,; \qquad (J1.02)$$

$$K = (\sum j308 + pj12-) + ((j3080 + j30811) + pj12-0) \qquad (J1.03)$$

2. 特定时间局部表示式

$$KQ = wj10-00c \qquad (J1.11)$$

$$KH = (\sum j308 + wj10-000) + (\sum j308 + jzz12-0) \qquad (J1.12)$$

3. 混合表示式

$$K = K1 + K2 \qquad (J1.00)$$

例句中的特定时间都是宏观表示。局部表示的例子如“下午 3 点 20 分”,混合表示的例子如“明天下午 3 点 20 分”。

在上列 HNC 表示式中,KH 的两个表示项∑j308 和∑(∑j308 + wj10-)需要详加说明,它们是特指时间表示的两个关键项。

符号∑j308 是基本数连用的一般表示式。

符号∑(∑j308 + wj10-)是数量连用的特殊表示式,专用于表示特定时间。一般的数量连用必须后跟基本命名(该数量所修饰的对象),但这里不需要,因为它是专用的,对象已经确定。这一知识是有关的处理必须把握的。

数的连用,或数的表示,有两种具体表示方式:语言方式和数学方式。计算机已能理解

数学方式。设计符号∑j308 的目的在于沟通这两种表示方式,使计算机对两者都能理解。基本数连用的语言方式和数学方式表示式分别是:

语言方式: ∑j308 = ∑(j3080 + j3081) (J3.21)

数学方式: ∑j308 = ∑j3080 (J3.22)

特定宏观时间的表示也有两种基本方式,传统方式和现代方式。传统方式必须采用 KQ + KH 构成,现代方式省略 KQ。传统方式可选取任一 ppj1 为参照系,现代方式则取唯一的 ff30j1 (公元)为统一参照系,因而它可以省略 KQ。

现代方式和传统方式各有千秋,前者方便,后者则有利于事件时间背景的联想,因此,人们常两者并用。如"苏东坡生于宋仁宗景佑三年(1036)"。

当∑j308 用于宏观特定时间表示时,各有相应的数字上限。对于"年",现代和传统两种表示的限制又有所不同,这些常识性知识不难形成规则表示。对此软件设计者可自行完成,这里就不来细说。

上面我们引入了特定时间描述语义块的 HNC 表示式,它揭示了这类语义块的构成知识,包括组合结构知识和各构成的概念优先性知识。HNC 符号里还包含各种概念联想的激活因子。在宏观特定时间的表示式中,还有一个 K 的直接 HNC 表示式,它实际上就是"??世纪??年代"语义块的形式化。这里需要说明的是:这些知识的具体表示方式(HNC 符号)肯定还有弱点甚至缺陷,然而这是枝节,重要的是它试图表达的内容,它揭示自然语言规律的策略,这才是本质,是计算语言学必须关心的本质问题。

11-2.4 宏观特定时间的模糊表示

模糊表示有数学和语言两种基本形式,模糊表示是自然语言的特长,汉语尤其特长。特定时间的模糊表示,不同语种各有千秋,但本节并不全面讨论宏观特定时间的模糊表示,而只限于函数 K(j1),即表示式(J1.02)所范定的宏观特定时间。

该表示式中的变量"j308"代表"基本数字"这一高层概念,在该式的求和过程中用具体的精确数或模糊数替换。如果替换的数字全部是精确数,函数 K(j1) 表示精确特定时间,反之,如果出现模糊数,则函数 K(j1) 表示模糊特定时间。

客观世界的模糊性表现是丰富多彩的,即使在特定时间这一局部范围内,也不是一两个具体的模糊数所能描述的。例如,"八几年"、"80 年代初"、"80 年代中"、"80 年代末",都是特定时间的模糊表示,具体含义不同,可满足不同的交际需要。这里的"几、初、中、末"都是模糊数,但它们显然含有一个共同的概念基元,将命名为基本模糊数,其 HNC 符号是 jz308,这个符号只是五元组概念的直接运用。于是,"几、初、中、末"的相应符号是:

几	jz3080
初,中,末	jz308 + g11m, m = 1, 0, 2

这里顺便一说,“几”不仅可映射成 jz3080,也可映射成 jz308,“人生几何?”里的“几”就属于后者。但它最常用的义项是 jz3080,我们将把它称为“个位模糊数”,其他称为组合模糊数。

这两类模糊数 jz308 都可替换函数 K(j1) 中的变量 j308,这是语义块函数的基本定义。但应该说明,这里的组合是线性组合,至于非线性组合,例如作用效应型组合和偏正组合,这一替换的普适性还有待考察。

这就是说,表示式(J1.02)可用于特定时间的模糊表示。这也是第三节所给出的一系列表示式的共性,表明概念层面的形式化确实大有可为。但必须指出,这类表示式一般不能表达自然语言丰富多彩的个性,特别是词汇层面的个性。这里需要从这个角度对特定时间表示式(J1.02)作两点说明。

第一,具体模糊数(即相应词汇)对(J1.02)表示式中变量“j308”的替换是有条件而不是无条件的。条件表现为对模糊数后接“量词”的限制及“量词”可否省略的限制。这里对这一语用方面的细节不来详说,它主要关系到语言生成,对语言分析影响较小,但对语音文本的解模糊处理仍有重要意义,因此,有关时间描述语义块的特殊处理仍需要把这些细节形成相应的规则。这里仅指出一点,就是模糊数后面不能直接跟量词“月”,这表现了自然语言在潇洒中的严谨,因为“月”的最大编号是 12,其模糊度在个位表示太大,在十位表示又太小,因而实际语言采用“年初、年中、年末”之类的说法。

第二,特定时间表示式(J1.02)的适用范围也是有限的,它只适用于时间模糊的数学表示,而不适用于时间模糊的语言描述。数学模糊表示一定出现在表示式的最后一位,这符合数学常识,某一位既已模糊,其低位的精确就没有意义了。但指定时间的语言描述可以违反这一常识,让高位模糊而低位精确。人们对事件的记忆,可能记住低位数而忘记高位数,例如记得朋友的生日,但不一定记得或知道他的生年。表述的需要有时只关心低位数,而不关心高位数。总之,自然语言的模糊时间描述呈现出琳琅满目的面貌,它允许高位模糊而低位精确,表示式(J1.02)是不能包含的。

对上列表示式中的求和符号 ∑,现在可以给出精确的说明,它不是一般的求和符号,而是数词与“量词”的连用符号,且“量词”由高位向低位递减。同一求和序列中的“量词”必须是包含性概念。这一陈述同样适用于内层的∑j308,由表示式(J3.21)可见,它的“量词”就是 j3081。

本节着重阐述了特定时间语义块的 KH,最后对它的 KQ 稍作交代。KQ 的概念优先子类分别由表示式(J1.01)和(J1.11)给出。

由表示式(J1.01)可知,宏观指定时间语义块的 KQ 优先于两个概念子类,一是朝代 ppj1,二是历史年代的命名 f30j1。前者已成历史,后者主要涉及带“古”字的专业。KQ 在现代语言里是经常是缺省的,缺省意味着使用“公元”,不仅如此,在口语里甚至还可以省略公元表示 4 位数字的前两位。

特定时间语义块的阐述就到这里,篇幅已不算短,然而,仅主要涉及对表示式(J1.02)的阐述。这充分表明把握语言知识海洋的艰巨性。但这里的论述方式希望能提供一个研究语

言现象的示范,一个区别于单纯语料库方法的示范。从上面的说明可以看到,不论在语句层面和语义块层面,还是在词汇层面,计算机擅长的数据加工不能代替大脑的总体性概念加工。就概念空间的分类来说,这一点更为明显。这时,你所需要的‘语料库’必须已经储存在你的大脑里。你只能期望实际的语料库在细节的明朗化方面发挥巨大作用,而不能指望它代替你的总体思考。

空 间 描 述 句

空间是有序的和三维的,这是空间的基本特征。

空间状态句的定义是:S∈jv2。

空间描述句的定义是:S∈jv2,SB∈j2。

空间描述句诱人的特点是:E-C 要素之间的关联性表现为狭义同行优先。

对宏观特殊空间描述句,E-B 要素之间也存在狭义同行优先。

同时间描述一样,"特定、特殊、序和间隔"这四个概念也是空间描述的关键性概念。下文就从这些概念的说明开始。不过,需要说明一点,对时间来说,重点是特定性的描述,特殊性已随着"公元"化而不重要了。但对空间来说,重要的是特殊性的描述,特定性则有统一的经纬度描述标准。

11-3.1 宏观特殊空间的 HNC 符号

在论题 11-2 中,我们主要是通过符号 wj1 和 pj1 对特定和特殊时间进行标记,我们曾把这两个符号叫做时间的物化和人化。这一原则当然也适用于空间。但空间的物化和人化则要复杂得多,即使在宏观特殊空间这一局部领域也是如此。下面给出一张有关概念的对比表,它清楚地表明了这一点。

	物化	人化
时间	1. 年月日时分秒	时代,朝代,世纪,年代
	2. 春夏秋冬,立春……	节日,假日
空间	1. 地域,地区,地点,地带	国家
	2. 温带,热带,寒带	城市,农村,村庄
	3. 陆地,海洋,水域	古迹
	4. 平原,高原,山区,沙漠	名胜
	5. 山,河,湖	港口,机场
	6. 岛,半岛	铁路,公路,航线

概念的物化、人化、挂靠和组合处理,已如前文所述,不是简单的分类,而是以能否提供灵敏的概念联想激活因子为主要依据。时间和空间概念都是按照这一原则来处理的。

这一原则的具体应用,对于时间概念比较得心应手,但对于空间概念则不然。表中的空

间概念,仅‘古迹,名胜,港口,机场,铁路,公路,航线’等比较容易建立特征联想因子,其他概念的联想特征比较分散。用数学语言来说,就是原有展开函数的收敛性较差,前一两级近似不足以凸现它们的主要联想特征。“山河湖”就属于这种情况,曾选用‘水’为‘河湖’的共同第一特征,‘自身转移’和‘存放’分别为它们的第二特征。但这种表示方式显然不够理想。

这就是说,部分宏观空间概念的物化需要引入新的特征空间。为此,对原设计作了两项改动,定义了两组新的概念基元。

一是:wj2 * y, y = 1 3 水域 y = 1 海洋
= 2 河
= 3 湖
wj2 * y , y = 4 7 陆地 y = 4 平原
= 5 山
= 6 沟、峡谷
= 7 沙漠

二是:wj2- 地区 wj2-00 地点
wj20- 地域 wj20-0 地带

第一项定义是对 jw 设计的模仿。第二项定义是对原设计的改动,与此相对应,原空间概念基本定义中的‘体面线点’应分别改为:

j20- j20-0 j20-00 j20-000

这些新定义的概念基元作为概念联想的激活因子,其‘方向性’和灵敏性优于原设计。例如,‘地域’和‘地区’这两个概念,现在不是按面积大小,而是按人文因素的多少为标准来划分的。人文因素的符号体现是将包含性标记‘—’前移,这一设计思想原来也有,但未全面贯彻。人文因素含量较少的特殊宏观空间,定义为 wj20- 系列(地域),人文因素含量较多的,定义为 wj2- 系列(地区)。这样,“山脉,山区,流域,三角洲,绿洲”等概念挂靠‘地域’,“根据地,解放区,开发区,基地,油田,矿区,白区,殖民地,战区,禁地,飞地”等概念挂靠‘地区’。从而改善了这两类特殊空间的联想脉络。

新定义不影响‘国家,城市’的原定义,但农村的映射符号略有变化,如下所示:

国家 pj2- 省 pwj2-0 县 pwj2-00
城市 pwj2-
农村 !pwj2- 村庄 !pwj2-00

上面定义的六类特殊空间(pj2-,pwj2-,pwj2-,wj2-,wj20-,wj2 * y)概括了宏观特殊空间的全部,当然不包含局部特殊空间。上列 HNC 符号就是宏观特殊空间的标记。

根据空间描述句的定义,如果一个语句的

E = S ∈ jv2
SB ∈ wj2;pj2

或

SBQ ∈ wj2；pj2

SBH ∈ j2

即该句为宏观特殊空间描述句。它具有引言中所说的 E-B、E-C 同行优先。

11-3.2 宏观空间描述句同行优先性的具体说明

让我们先看一组不完整的文字序列：

1.“长江三角洲位于……”

2.“北京离上海……” “……离上海约 300 公里”

3.“天山山脉横贯……”

4.“河南在湖北的……” “……在湖北的南面”

5.“乌拉尔山南北绵延……”

6.“荆州东连吴国，西通……”

7.“珠穆朗玛峰海拔……”

8.“我国海岸线全长……”

9.“湖北省面积……”

如果这是一组地理知识填空试题，完整的回答当然需要一定的地理知识（常识及专业知识）。但是，如果我们关心的是句类分析，是对后继语义块的预期处理，则所需要的知识就完全不同，常识及专业知识就退居次要地位，甚至是无足轻重的地位。本节试图着重表明这一点。

这里关键知识是：概念 jv2m 不仅可唯一确定句类（空间状态句），而且对后续语义块 SC 的要素 SCH 强加了同行优先的约束，即 SCH 必须属于 j2 * y。

上列各句出现的 E 要素“位于，横贯，在，离，绵延，东连，西通”都属于 jv2，前 3 个属于 jv21，后两个属于 jv23；“离”和“绵延”属于 jv22。因此，前六句话后续语义块的 SCH 已经是“胸有成竹”了。

那么，上述关键知识存放在哪里？回答是：就在 jv 自身，凡 jv 都有这一特性。这也是我们引入类别符号“j”的原因之一。

但应该指出，同行优先，包括语义块内部和语义块之间的同行优先，既有狭义广义之分，又有层级之别。具体运用需要在概念层面和词汇层面给出具体规定。狭义同行优先无条件适用于语义块内部，但不是无条件适用于语义块之间。然而，时间、空间、数量这三个基本概念例外，它的块间同行优先是无条件的。层级之别是 HNC 符号设计的矛盾之一，例如这里 jv23 的同行，就只到 1 级，其 C 要素优先于 wj2。这一知识存放在概念关联知识库里。那么，为什么不把 jv23 安排在 jv20？为什么从 j0 到 j6 只有 j2 和 j3 在第二级安排了“3”？说来话长，这里就从略了。

同行优先准则提供了 SCH 的预期信息,但并没有提供 SCQ 的预期信息。其次,有的 jv 需要 SCQ,有的不需要,有的可能需要,这就是语义块构成知识。这一知识在语义库中都有明确的说明。例如"位于、在"属于需要;"绵延、东连、西通、离"属于不需要;"横贯"属于可能需要。除语义块结构信息外,词义库还提供 SCQ 的概念约束信息,例如,对"位于"将给出 SCQ∈(wj2;pj2)的信息,这时,jv2 可提供完整确切的 SC 语义块构成信息。当然,不是所有的 jv2 都这么美妙,例如"在"就不能提供 SCQ 的确切信息。

"位于"与"在"的差异不仅前者是单义词,而且在于前者只用于描述狭义宏观特殊空间,而后者还可用于描述广义空间和局部空间。

上面只涉及 E 要素所提供的信息,而一般说来,在 E 的前面还有明确无误的 SB 信息。例如,上列前 6 组"不完整文字"就提供了"宏观空间描述句"的确定信息,句类知识及后续语义块的构成知识"万事俱备,只欠一用"。即使对于"在、离"这样的多义词,SB 信息的加入,使多义模糊迎刃而解。这时,你(软件)完全可以作出类似人类大脑的感知和预期处理。这里值得着重指出的是:你已经摆脱了常识海洋的困扰。这一点,难道还有什么疑义吗?我认为没有了。

当然,你(软件)不应该有万事大吉的错觉,后三组"不完整的文字"就别有天地。这里,"海拔""全长""面积"都是名词,到此,E 要素还渺无踪迹,语义块感知似乎是一片茫然,你(软件)能有所作为么?

关键在于两件"武器":

第一件是状态句可以没有 S;

第二件是 jz 的基本特性:充当状态句的 SBH

省略 S

后面紧跟数词和相应的量词。

运用这两件武器就足够了,就"柳暗花明又一村"了。

这两件概念层面的武器存放在哪里?

按照张全的建议,可在"海拔"、"长"、"面积"(它们的词性都定位为 g)的语义表示框架中添加一个特殊的 C"槽";"C:j308 + jzz"。

这里,用不着汉语自有汉字以来就使用的特种武器,即静词(这个术语是柏拉图提出的,包括名词、形容词和副词)与动词相互转换(静词之间也可相互转换)的武器。

不过,应该指出,这里仍隐含着句类转换、SB 语义块下界的判定以及"海拔"、"全长"的词性定位问题,也许在将来的适当时候加以讨论(参看论题 15)。

11-3.3 框架描述的要点

空间描述的要点包括:

1. 三维空间的一个特殊维和特殊面

2. 宏观和局部空间、微观和超级空间、狭义和广义空间

3. 三个参照系

4. 相对性与绝对性

5. 笛卡尔坐标和球坐标

下面就来分别对这五个要点作简要说明。

11-3.3.1 三维空间的一个特殊维和特殊面

抽象的三维空间是没有特殊维和特殊面的,但对语言空间必须赋予一个特殊维和特殊面,这就是语言里用‘上下’表述的维和‘地面’表示的面。

由这一特殊维形成三个特殊子空间:wj21030,wj21031 和 wj21032,汉语符号相应是“地面”、“天空”和‘地下’,其中 wj21030 是语言的面,不是几何学的面。“天空”与“地下”以此面为界对偶而存在。

对于以地球为参考的具体空间,这是最佳的描述方式。因为,wj21030 这个特殊面,是人类活动的主要舞台,是生命赖以存在的基本空间。第 1 节中所定义的全部特殊空间,实际上是对这一特殊面的划分。

语言的特殊面‘地面’wj21030 不是平面,但在局部领域,可按照平面描述。

11-3.3.2 宏观和局部空间、微观和超级空间、狭义和广义空间

语言对空间的描述有它自身的标准,了解这些标准,对于理解语言的空间描述极为重要。

语言按空间的规模,划分宏观和局部空间。

语言又按空间的功能表现,划分狭义和广义空间。

宏观空间的基本内涵就是第 11-3.1 节定义的 wj2 和 pj2。

在宏观空间内存在并可直接感知的物(包括人)及其组合构成局部空间。

宏观空间只使用面积的概念,不使用体积和重量的概念。

局部空间则使用面积和体积的概念。

宏观空间是不可移动的,局部空间大多数是可移动的。

可移动的局部空间使用重量的概念。

不可直接感知的物构成微观空间或超级空间,对这两类空间的描述需要专业知识和非自然语言符号。

局部空间实质上就是广义空间。

用 HNC 语言来说,符号(wj2;pj2)是宏观空间的标记,不带 j2 的(w;p)是局部空间的标记。符号 j2 是狭义空间的标记,符号(j01;j42)是广义空间的标记。

自然语言对不同空间的尺度描述,选用不同的词汇。例如,超级空间用‘光年’,宏观空间用‘公里’,局部空间用‘米’,微观空间用‘微,埃’。从这个意义上说,我不太同意以‘千米’

替换“公里”的标准化,因为它模糊了语言的空间描述特性,无视汉语的历史传统。

11-3.3.3 三个参照系

方位的描述是空间描述的主要内容。

语言对方位的描述采用了三种不同的参照系(坐标系):

一是“上下东西南北”的参照系;

二是“上下前后左右”的参照系;

三是“内外边”的参照系。

这里的对偶概念“上下”、“东西”、“南北”、“前后”、“左右”、“内外”是不同方位的值描述。不同语种对这些空间概念的表述都比较简洁,汉语尤为令人惊叹。语言的发源和学习都是很自然地从基本概念起步,计算机也应该如此,这正是促成我写这一组短文的原因。

第一个参照系以“地面—太阳—地物”为参照,是典型的三维参照系。多用于宏观空间描述,也可称宏观参照系。第二个参照系以“某一特定物—该物物向”为参照,多用于局部空间描述,也可称局部参照系。“内外边”以某一特定事物为参照,不管方向,可称广义参照系。

前两种参照系的“上下”都以上述特殊维为参照方向,但参照面不同,宏观参照系的“上下”有绝对意义,但局部参照系的“上下”只有局部意义。至于“东西南北”和“左右前后”都只有相对意义。但“东经,北纬”等概念例外。

三个参照系的知识是空间方位描述语义块构成的基本知识。其语义块函数可写成:

宏观描述	SBCQ = wj2;pj2
局部描述	SBCQ = w;p
广义描述	SBCQ = w;p;g
宏观参照系	SBCH = j210
局部参照系	SBCH = j214
广义参照系	SBCH = j42;j2143

式中的 SBC = (SB;SC),表示 SB 或 SC 的含义。应注意到,广义参照系也使用局部参照系的“上下”概念。关于广义空间语义块的构成知识,曾有专文讨论。

空间描述的“宏观”、“局部”或“广义”性,在一个句群或篇章甚至整篇文章里,应保持一致,因而是一项非常重要的语境知识。

语境生成模块还只有一个很粗略的设想,HNC 理论只给出了基本分类的启示性知识。这一组短文的许多论述,应视为对语境生成的预备说明。

序与数量的描述

本文标题与其姊妹篇不同。它将着重于数量描述语义块的阐述,句子的说明将放在次要地位。

数与序、量是紧密相互联系又有本质区别的概念。因此,本文的内容将包括:

1. 数的描述
2. 序的描述
3. 数量的描述
4. 数量描述句

11-4.1 数 的 描 述

数的语言描述的基本表示式是∑j308,已在论题 11-2 中作了详尽的讨论。j308 定义为基本数。另外,还定义了相对数∑j309 和潜在数 j30a。基本数、相对数和潜在数是数的语言描述的基本概念,在常识范围内通常只运用前两个概念。

数的动态概念有基本操作 jv3 和操作效应 jvr3 两类,前者如"加减乘除",后者如"等于,大于,小于"。

数的基本操作概念 jv3("加减乘除")与 v340、v390 交互关联,是一种效应概念,数操作效应概念 jvr3 与 j1vr00 交式关联,是一种比较判断概念。因此,数描述句可呈现三种类型:效应句、基本判断句和效应判断句。例如,"二加三"是效应句,"六大于五"是基本判断句;"二加三等于五"二乘三等于六"是效应判断句。

数的描述具有绝对同行性,也可称为纯净性。因为,在数的描述中不容许出现其他任何概念。因此,就数的描述来说,语义块的概念都显得是多余的。

数的世界是一个非常特殊的世界,数实际上不应该与其他基本概念并列,更不应该把它的编号安排为 j3。我一直为此而耿耿于心。读者从数描述的绝对同行性或纯净性可以体会到这一点了。

在实际语言中,数描述句不多,纯粹的数描述更是罕见,几乎只涉及很专业的纯数学。因此,自然语言理解对数描述的处理可满足于四则运算的水平。麻烦的是序描述和数量描述。

11-4.2 序 描 述

本节是对论题 11-1 的补充。先阐述序的三项基本属性:先后、正反和循环。然后说明序描述的一般表示式。

11-4.2.1 序的三项基本属性——先后、正反和循环

先后的概念来于基本概念时间。"先后"这一概念应理解为时间概念节点'j1'的基本属性。

过程是状态序列的时间表现,所以,过程有先后。

过程的对偶和对比概念都具有先后特性,这是 HNC 符号的一项隐含约定,概无例外。过程的对偶概念约定'1'先'2'后,正对比概念的约定'小数'先于'大数",反对比概念'大数"先于'小数"。这里不妨回顾一下过程概念的全部二级节点:开始 1 先于结束 2,原因 1 先于结果 2;源 5 先于汇 6,生 1 先于灭 2,生 5 先于死 6;幼年 1 先于少年 2,少年 2 先于青年 3,青年 3 先于中年 4,中年 4 先于老年 5。这些知识是人所共知的常识,但 HNC 理论把这类常识转化为概念层面的知识。这一转化是通过 HNC 符号来体现的。用汉语来表达,就是'对偶和对比型过程 v、g 类概念具有先后特性"。这里读者应注意到'v、g'的类别性约束,非'v、g"类过程对偶和对比性概念是不具有先后特性的。

从计算机对'先后'的理解来说,关于'先后'的阐述不能到此为止,上述概念层面的知识如何表达,如何进入程序的运用,都还需要进一步明确。但本文能推而求其次,把这一使命留给软件设计者。

序的先后性当然不限于时间和作用效应链的过程,作用与效应,作用在先,效应在后;作用的承受与反应,承受在先,反应在后;同行概念的 v 与 r,v 在先,r 在后;转移概念的出发在先,到达在后;问在先,答在后;增长过程的'小'在先,"大'在后,而缩减过程相反,"大'在先,"小'在后。这些广义过程的先后顺序 HNC 符号目前也未给出明确的表示。

正反的概念来于基本概念空间。空间的位置和方向(即空间的序 j21)都有正反之分。正反意味着可逆,时间的序是不可逆的(即所谓时间不能倒流),空间的序是可逆的。这是常识,又是概念层面的基本知识。

转移是状态序列的空间表现,所以,转移有正反之别。

转移的多数对偶性概念具有正反特性。入出,进退,升降,来去,买卖,借还都互为正反。曾设想过用'1,2'表示转移的正反,用'5,6'表示转移的先后。但实际上行不通。这一概念层面的重要知识只能在节点知识库(概念知识库之一)中用'j71n,n=1,2'予以标记。

正反'j710'不仅是对狭义空间对称性的基本描述,也是对基本广义空间(即作用效应链空间:作用效应空间、过程转移空间、关系状态空间)对称(对偶)性基本描述。在这些基本广义空间里,对称性普通存在,所以,HNC 符号对这一概念层面的重要知识描述用中层数字

“n，n = 0　7”予以最特殊的凸现。

在社会空间，即基元概念 6 行以后所描述的各子空间，对称性概念往往具有扩展的社会意义，即基本概念 j8 所定义的各种对偶意义。但这一扩展意义仅由对称性标记“1，2 " 5，6”予以默认。

循环的概念来于自然界的各种循环现象，从简单的旋转运动到大自然和生命运动的各种复杂循环现象。凡具有循环性的概念，例如时间的 wj10- 和 wj11c，空间的 j218，都必须在节点知识库中予以标记，标记符是“100a9”。

循环性标记当然也可在词汇层面使用，如“春夏秋冬”、“立春，雨水……”（它们都属于 wj11c）、“甲乙丙丁……”、“子丑寅卯……”、“甲子，乙丑……”、“年，月，日”等。这里顺便一说，公历的“月”因徒具循环的形式，其标记应取 j742/100a9。农历的“月”才可直接用标记符 100a9。

本节两次提到的节点属性标记，在目前的概念知识库中是用“0”级交式关联来表示的。但该库的规范化还有待商定，近期应专题讨论。

11-4.2.2　序描述的一般表示式

在论题 11-1 中，曾给出序描述句 SC 语义块的一般构成形式：

$$SC(j00) = \text{“比较范围”} + \text{“序值”}$$

这实际上是一种特殊形式的序描述表示式，其一般形式是：

$$FK(j00) = \text{“序值”} + \text{“序量词”}$$

$$= (f30j00 + \sum j308) + jzz00 \qquad (J0.01)$$

(J0.01)是序描述语义块的块素。式中的 f30j00 是“序”标记，汉语符号是“第”，英语符号是“th”。但英语的“th”在 $\sum$j308 之后。两种语言都有违例，英语是每一个十进位的前三名另设标记，汉语是“第”可省略，这都给理解处理带来了一定麻烦。

将表示式(J0.01)扩展，使之包含内容，则构成序描述语义块的 KH 的一般表示式。

$$KH(j00) = \sum(FK(j00) + \text{“内容”})$$

$$= \sum(FK(j00) + C(j00)) \qquad (J0.02)$$

式中的 $\sum$ 表示序的多重结构，例如“第十四届中央委员会第六次全体会议”就是序表示二重结构。

于是，“中国共产党第十四届中央委员会第六次全体会议”就是一个典型的序描述语义块。这里的对象是“中国共产党”，内容分别是“中央委员会”和“全体会议”，相应的序量词分别是“届”和“次”。

应该说明，由于 KQ 不是 K 的必须部分，因此，(J0.02) 也是序描述语义块的一般表示式。

序描述语义块同论题 11-2 中所阐述的时间描述语义块一样，可写出明确的表示式。我们将把这类语义块叫做 WD(well-defined 的简写)语义块。不言而喻，这类语义块的上下界容

易确定,对语义块感知大有裨益。如果 WD 语义块的块素又能以 HNC 符号写出函数表示式,如特定时间描述语义块那样,则可称为 SWD 语义块,即超级(super)WD 语义块之意。

WD 语义块都能写出明确的构成表示式,这正是语义块构成研究的目标。为了说明或理解之便,需要对构成成分给以命名。所谓核心部分 KH 和说明部分 KQ 是最基本的命名,但显然不够。对 KH 或 KQ 的进一步分解,就需要引入新的名称。但命名必须严谨,必须遵循孔夫子关于"正名"的教导。本系列说明对语义块二级成分的命名都采用对象和内容的概念,因为,它们是 HNC 理论的基本概念。这里不能不引用〈问答 32〉中一段很长的文字。

"上面我们以转移为例,说明了关于对象和内容的奇特定义方式,这个定义方式确实有些奇特,然而是关键性的。它先脱离这两个词的常规意义,仅从语义块的可扩展性给出最抽象的定义。然后参照它们的常规意义赋予两者以函数形式的范定。

回到作用和效应,它们的对象可定义为被影响者或接受者。按照这个定义,作用和效应的对象可以是任何事物,从具体的人和物到人的任何活动以至人的认识和观念。这完全符合常识,也无可非议。但理解要求建立联想脉络,我们必须把无可非议的"任何事物"的"任何"二字加以限制,否则就不能前进。施行这项限制的基点就是把作用和效应划分开来,区分作用对象和效应对象。有了这个基点,下一步的问题就是制定区分的标准,这个标准应该说比较容易选择,这就是:(1)具体与抽象;(2)整体与局部。这样,就能形成下面的定义:

作用对象:具体及整体的事物

效应对象:抽象或局部的事物

这个定义的要害是"及"或"两字,在形式上给出了两类事物或两类对象的无模糊界限。有了两类对象,对复杂 B 语义块就有了表述的手段。复杂的 B 语义块通常包括多项块素,但其基本骨架一定是由作用对象、效应对象和效应内容构成。这没有任何奥妙:因为,对于对象的充分说明不外乎具体与抽象、整体与局部这两大方面的两个侧面,而它们都已包含在上面的定义里。这就是说,复杂 B 语义块的构件清单已经明朗了。剩下的问题是它们任何排序,由于抽象从属于具体,局部从属于整体,如果给这个从属关系的双方约定一个顺序,这个问题也就解决了。

中国人的习惯是从属方在后。即整体在前,局部在后;具体在前,抽象在后。表现在语义块中就是对象在前,内容在后;作用对象在前,效应对象在后。这就是汉语 B 语义块构成的基本规则。说句多余的话,我们之所以能得到这个规则,就因为我们引入了作用和效应、对象和内容的观念。"

这段引文实际上阐述了语义块构成的基本原则,这一基本原则可概括为:对语义块的构成成分给以二维描述,而不是一维描述。这同语句构成的基本原则是完全一致的。词汇、语义块和语句构成原则的一致性,传统语言学已有深刻认识,HNC 理论的发展在于把传统的

一维线描述转变为二维面描述。这个描述面的横坐标是句类,纵坐标是语义块类型,语义块是句类的函数。语义块类型的概念与传统语言学角色的概念大体相当,但不同于“格”,因为,“格”不包括 E 语义块。自然语言总共需要多少个“格”?格语法理论提出者菲尔墨本人及后来的许多跟随者都试图回答这个问题,然而始终不得要领。从 HNC 理论二维描述面的观点来看,这个问题显而易见。格有主辅之分,$7 \times 4 - 3 = 25$,就是主要“格”总数的最终答案。辅“格”总数的理论值 $7 \times 7 = 49$,但由于辅“格”弱依赖于句类而可以兼并,确切的总数并不重要。

但是,问题的实质不在“格”的数量,而在于“格”和句类的层次分解。HNC 理论关于作用效应链和五元组的发现,关于宾语首先应分为对象和内容的思想,关于超脱具体语言进行概念总体思考的方法论,是构造二维描述面和进行层次分解的基础。

应该说,在菲尔墨的追随者当中,不少人曾闪现过二维描述面的思想火花,可惜由于缺乏上述理论准备而未得大成。

回到语义块构成,读者应注意到,我们对它的分析方法也是从“对象、表现”的二分法着手,上面的引文和论题 11-2 中关于语义块形式表示的论述都是采用这一方法,这同语句的分析方法是完全一样的(参看【2】)。从这些论述中,读者应能体会到,如果不引入对象、内容以及它们是句类函数的概念,而囿于传统的“主谓宾补”概念,我们就不能前进。

“函数”的说法意味着相对性,在上述引文的后面对此有详尽的阐述。这里就不来重复。

这里应顺便说明一下“块素”这个词,它是块素一词的自然引申,词有语素,语义块有块素,语句有句素,句素就是语义块,语义块的核心称为语句要素,也简称要素。从上面说明所知,语素一词改称词素较妥,但这就不必强求了。

从语义块的 WD 特性,引来了上面的一大段议论,目的在于表明语义块同语句的句类格式一样,可以写出各种表示式。语义块表示式应成为计算语言学的重要研究内容之一。这一研究需要语料库的帮助,但关键是理论模型。在 HNC 理论的指导下,利用语料库有针对性地对各种具体类型的语义块分别进行研究,才能取得有价值的结果。这类似于生命基因网络图的巨大工程,我们应作出作为先行者的无可替代的贡献。

如果语言都老老实实地按照各种表示式来陈述,那自然语言理解基本上就万事大吉了。实际情况当然不会这样。但是,语言再调皮,一般不过分离谱。目前受限语言的提法很时髦,但不能把受限当做急于求成的救命稻草或无所作为的挡箭牌,我个人倾向于把受限定位在语言的过分离谱表现之内。所谓过分离谱,主要指以下两点。1. 方言和某些作家对词义的超出原有联想脉络的扩展;2. 习用语或口语中一些不留线索的缺省。语言的“谱”,就是句类知识和句类格式、语义块构成知识和构成表示式。它们是理解处理的基本立足点,语境生成、合理性分析、隐知识揭示、要点主题分析(即篇章分析)以及汉语的新伪词辨识都只能在这个立足点上才能取得实质性的进展。

上面的示例可简化为“中共十四届六中全会”或“党的十四届六中全会”,这就是序描述语义块的简化表示。简化是非常规表示的基本形态,下面就来对序描述语义块的非常规表

示作简要说明。

非常规表示通常是以下四种简化的组合：

1.“第”省略。这是最常见的情况。

2. 序量词的省略。如上例对“次”的省略。

3. 内容的简化。如“中央委员会"全体会议”简化为“中”和“全会”。

4. 对象的简化。如“中国共产党”简化为“中共”或“党”。

上面的简化示例虽然“四简”俱全，但从上面关于过分离谱的定义可见，它并不过分，全部失去的信息都保留着“明显的”线索，你不难判定它仍然是序描述，因而失去的信息是可以恢复的。

序描述语义块非常规表示的麻烦在于汉语对数的活用，特别是“一”的活用。使序描述、数描述、数量描述、列举描述、属性描述、QE 描述、条件描述、语气描述的界限出现模糊。

让我们看一些示例。

一看、二慢、三通过。	序描述
三七二十一，三下五除二。	“数描述”
三国四方，五男三女	数量描述
一穷二白，一大二公。	列举描述
三不准，五讲四美，“九五”	概括描述
五老峰，九斤老太	属性描述
七下江南。三顾茅庐。	QE 描述
一触即发。一说就生气。	条件描述
一去不复返。一走了事。	语气描述
谈一谈，尝一尝	插入
七上八下，三心二意	其他

这只是汉语数活用的不完整清单。对这个清单里的“数描述”加了引导，因为它们的意义超出了数的联想脉络，属于过分离谱。这就是说，其他的描述都是不离谱和可以理解的，虽然理解处理的难度较大。

为什么拼音输入要指定基本数字？就因为汉语对数的活用几乎无所不在。

对指定数字，已有一个特殊处理系统。但从上面的上面可知，目前该系统的处理功能远远不能满足数描述的需要。

这里简单说明一下扩展数字特殊处理系统功能的策略。

数描述虽然种类繁多，但从假设检验来说，主要是以下四类：

1. 数量描述

2. 序描述

3. qv

4. 其他

目前只进行了第一项假设检验,而且检验的方式还不完备(见下文)。应按上列步骤逐步扩展,分批或集中完成皆可。但‘一’的插入检验应放在首位或由调度程序承担。“九五”之类特殊数描述则另行处理。

11-4.3 数量描述

本节讨论数量描述的一般形式和数量描述语义块的一般构成形式。

数量描述语义块有两种基本形式:‘内容数量描述语义块’和‘对象数量描述语义块’。

数量描述是上述两种数量描述语义块的共同块素,即数量描述语义块的块素。三者之间的关系是:

“内容数量描述语义块”=“数量描述”+“描述内容”

$$KC(j41) = FK(j41) + FKC \qquad (J4.01)$$

=“描述内容”+“数量描述”

$$= FKC + FK(j41) \qquad (J4.02)$$

“对象数量描述语义块”=“数量描述”+“描述对象”

$$KB(j41) = FK(j41) + FKB \qquad (J4.03)$$

=“描述对象”+“数量描述”

$$= FKB + FK(j41) \qquad (J4.04)$$

=“描述对象”+“内容数量描述”

$$= FKB + KC(j41) \qquad (J4.05)$$

=“内容数量描述”+“描述对象”

$$= KC(j41) + FKB \qquad (J4.06)$$

三者的示例:

数量描述:“80 平方米”

内容数量描述:“80 平方米建筑面积”;“80 平方米使用面积”

“建筑面积 80 平方米”;“使用面积 80 平方米”

对象数量描述:“80 平方米住房”;“住房 80 平方米”

“80 平方米建筑面积的住房”

“建筑面积 80 平方米的住房”

“住房建筑面积 80 平方米”

(J4.01) (J4.06) 是数量描述语义块构成的一般表示式,两种数量描述语义块构成的突出特征是‘内容(或对象)与数量描述‘块素’可相互交互位置。当内容或对象在后时,两者之间应该加语义块偏正结构符号‘的’。

请读者注意最后一句话里的‘应该’二字。它体现了语义块构成的一项约定。在下一节,对此有详细说明。

从上面的例子可以看到,内容或对象数量描述语义块的内容或数量既可在数量描述之前,也可在数量描述之后。如果对内容和对象同时加以描述,则将“内容数量描述”绑在一起,它与对象的排列顺序可相互交换。这就是数量描述语义块的一般规则,由(J4.01)(J4.06)式予以表达。

已有的数字特殊处理系统只适应(J4.03)所表述的情况,应根据上述知识加以扩充,使之适应上述所有情况。

以上所述,是数量描述语义块的常规构成形式。

数量描述语义块的非常规形式将在数量描述句中一起论述。在此之前,将对语义块构成知识,作一个一般阐述,目的在于引起本系列说明读者的思考。

11-4.4 语义块的构成知识的一般阐述

一般说来,语义块的构成知识比语句构成知识更难以表述。HNC 理论把语句构成知识用句类格式来描述,句类格式是句类的函数,这一函数关系是最重要的概念层面知识,是句内上下文概念联想最基本的激活因子。句类格式隐含着一项约定:当语义块的排列顺序偏离标准格式时,必须在相应语义块前面加语义块切分指示符,省略指示符(标记)的违例,只能出现在语句要素强关联的情况。上述 HNC 语句构成规则完全不同于传统的句法树规则,是对语句深层模式的阐释,是把语言分析引向人类思考模式的引桥,而且,规则的实质与语种无关。

与语句构成知识的阐释相比,语义块构成知识的阐释还不够系统和成熟。上面曾将语义块构成表示式与生命基因图相比,这容易造成现在就可以实行工程化的错觉。实际上,对语义块构成的研究还需要下大力气,才能达到像句类格式那样的成熟水准。

本节仅对三个理论问题作一些说明。

1. 在论题 11-2 的“语义块的形式表示”中,将语义块的主体构成表达成 KQ + KH,但并未指明块素 KQ 和 KH 之间的概念组合结构。对 KQ 和 KH 的命名容易造成两者为偏正结构的误解。

HNC 理论所揭示的四类概念组合结构(参看【1】)对语义块内部的概念组合依然适用。例如,“武力解决”必须纳入第二类逻辑结构,“名人效应”应纳入效应型结构。但是,正如【1】所指出的,“从理解处理的角度来看,组合概念的具体结构并不重要,重要的是它的功能表现”。这就是说,至少在当前,我们并不要求软件承担揭示语义块内部概念组合结构的任务,虽然更深层的理解是需要的。我们关注的是能否写出块素的 HNC 表示式。

2.【1】指出“在四类组合结构中,除了逻辑第一子类外,都存在正反两种形式”。这里需要补充说明,这一说法只适用于语句和词汇层面,而不适用于语义块。语句的正反形式(相应于语义块的搬移)不影响语句的意义,词汇层面也基本如此(例如“狠心”与“心狠”,HNC 理论认为它们是正反两种偏正结构,意义相同)。语义块则不同,顺序的改变必将带来

意义的改变,例如,“伟大的中国”和“中国的伟大”;“文化传统”和“传统文化”,意义是不同的。

3. 名词构成的语义块存在一种天然顺序,不容颠倒。这个现象很值得研究,我试图用准则的形式予以陈述,以便为语义块感知的局部处理及合理性分析提供一个新的手段。

名词构成语义块的四项顺序准则是:

准则 1　整体在前,局部在后;

准则 2　对象在前,内容在后;

准则 3　非 r 在前,　r 在后;

准则 4　低层在前,高层在后。

这四项准则的先后具体约定当然带有汉语的个性,但其内在的顺序性是概念层面的特征,与语种无关。

这些准则的含义都简单明了,准则 1 和 3 更为浅显,我们每一个人都能运用自如。问题是,如何让软件也能运用这些简单的准则?

我的设想是:对“整体和局部”“对象和内容”“低层和高层”给出明确的符号特征,从而把这些准则的运用转变为对相应符号的操作。

在说明具体操作之前,先给出这四类语义块的若干示例。

准则 1:湖北宜昌,亚洲国家,大楼顶层,80 年代末期,扎伊尔东部,鸭翅膀,汽车发动机,雷达天线,中科院声学所

准则 2:人民体质,学术水平,粮食产量,产品质量,中国特色,群众意见,国际秩序,企业职工,部队战斗力,城市居民生活垃圾

准则 3:政治权力,商业利益,外交政策,政治路线,自然灾害,全球品牌,精神财富,人民财产,水门事件,概念层次网络理论

准则 4:军事力量,当地势力,历史进程,国际关系,财政方面,精神状态,欧洲联盟,外层空间,当地时间,政界精英

为了下面说明的方便,这里把每一个示例都当做一个语义块(实际上可能是更大语义块的一部分)。

准则 1 主要涉及具体概念 w 和 p,示例的 KQ 部分全是 w 或 p 类概念。整体和局部的本来意义并不限于具体概念,但具体概念的整体、局部特征比较突出,其符号特征也容易辨认。因此,它应该成为软件运用的优先对象。

在 HNC 符号体系中,整体、局部特征比较明显的概念集中在以下四类:

1. 带包含性符号“—”的概念,如示例中“湖北,大楼,年代……”。

2. pe 类概念,如示例中的“中科院,声学所”。

3. jw6 类概念,如示例中的“鸭翅膀”。

4. w9 类概念,如示例中的“汽车,雷达,发动机,天线”。

显然,这四类概念按准则 1 构成语义块时,除了顺序约束之外,还需要满足狭义间同行

优先的准则，表现了最典型的 SWD 特性。

准则 2 涉及对象和内容的定义。两者的天然排序，我们已阐述多次，这里就不来重复。示例表明，第二类语义块不具有 WD 特性，软件对它的运用具有较大的困难。但应该指出，以基本概念为内容的组合还是容易判定的，如示例的前五个，这就是说，准则 2 仍具有局部 WD 特性，软件不可对此失之交臂。

准则 3 是 r 类概念的自然推论。这条准则本身即包含符号特征信息，因而具备 WD 特性，其中的 r8，ra 类概念更具有 SWD 特性。

准则 4 体现对高层基本概念和基元概念的说明。HNC 符号体系对“高层”有明确的定义，因而准则 4 也具有 WD 特性。

上述准则所约定的语义块是名词-名词型偏正组合结构的一个子集。【1】曾说过，“语法类沿用原来的命名，但实际内涵已大为缩小。缩小后的偏正还需要作二级分类，这项分类工作也相当复杂，这里就不来讨论了”。两年多以前所说的这项工作迄未启动，本节文字算一个启动的信号吧。

11-4.5 数量描述句

以数量描述语义块 K(j41) 为 SC 的语句定义为数量描述句。

$$SJ(j41) = SB + S + KC(j41)$$

但数量描述句经常采用下面的格式，第 3 节所给出的数量描述语义块表示式反而主要用于非状态句。

$$SJ(j41) = SBC + S + FK(j41) \quad (J4.11)$$

$$= SBC + FK(j41) \quad (J4.12)$$

$$S \in jv41$$

数量描述句属于 SWD 语句。

jv41 的反映射汉字主要有“达，为，有，是”。这些汉字多义性极为丰富，但数量描述句的 SWD 特性应付这里的多义困扰，显然绰绰有余。

但是，一个特殊的 jv41 值得一说，那就是汉字“占”。它专门用于相对数的数量描述。这时，FK(j41) 应取相对数，相对数之前，还要加相应的说明部分。这些细节，请软件设计者自行解决。

关 于“ l9 ”概 念

为了一个特定的概念写一篇短文,在 52 个论题中是绝无仅有的。为什么?因为 HNC 理解处理强调从语义块感知切入,一切激活语义块感知的信息都要予以特殊关注。l9 是激活信息中十分特殊的一种。

l9 和 pl9 这两个概念相应于语法学的指示代词和人称代词,这两个命名是很准确的语义说明,HNC 分别把他们转换成计算机容易理解的概念类别符号 l9 和 pl9。符号 l9 表达了两者的共性——“ 指称 ”,指示代词是一般逻辑指称,用于各种语义块(不分主辅)的起始指示,故直接用 l9 表示。人称代词专用于对人的逻辑指称,故用 pl9 表示。人称代词和指示代词有三大特点:

1. 在与其他概念组合时,它们一定充当语义块(不是短语)的头。它们前面的“ 的 ”也不改变这一规律,只不过这时它是句蜕块中的“ 子块 ”。

2. 它们都可分别充当自足性语义块。

3. 当两者同时出现时,人称代词必须在指示代词之前,而且失去了自足性特征。如果表现出伪自足(见下文),则 l9 为语气词 f50。

这是三项很特殊的语法知识。为了激活这一类的特殊联想,HNC 统一采取了概念类别符号与层次网络符号不一致(人称代词的映射符号是 p400n-)的表示方式,这个“ 不一致 ”代表概念的多元性或综合性表现,多元性表现比较简单,综合性比较复杂,有时需要激活一类局部规则去取得有关知识。这一类局部规则通过类别符号去检索。

类别符号 l9 产生的激活过程是:

l　　语义块起始或内部构成的激活

9　　如果它前面不是 p400n-,激活语义块感知。

　　如果它后面紧跟 l0、l1、QE 或 EQ 激活点,表明它自足,或者是句蜕块的迹象。

　　如果它前面有 p400n-,不另激活。

　　如果它后面紧跟上列激活点,是句蜕块的迹象。

　　如果它后面紧跟 E 块,表明出现了所谓伪自足。

注意,在上面的规则中,动词没有作为激活点来对待,因为 l9 之后的动词经常被名词化。因为,l9 具有“ qv□g ”的语法功能(参看论题 2-1)。这项功能与其自足性特征是相互冲突的,这一冲突必须通过句类检验来解决。一般来说,当音串只出现一个 E 团块,而且该团块有 QE 或 hv 的加强时,该 l9 自足;当音串出现两个或两个以上 E 团块时,先不承认 l9 后面动词的 E 资格 如果 l9 后面无 QE 或 hv 的加强,可以取消后跟动词的 E 资格。

语言逻辑概念网络的设计曾有过三个版本，第二个版本曾取消过语义块内部构成的信息标记 l4 和 l5。这样，l 类概念的出现就是单纯的语义块激活信息，具有简明的优点。后来，考虑到语义块激活信息与短语标记信息不可能截然分开，特别是汉语的短语括号信息十分宝贵，又恢复了 l4 和 l5 的布局，但具体设计方案不同于第一版本。

1998 年 6 月 28 日

论语句表示式——兼论“格”

13.1 引言:历史回顾

HNC 理论追求语句的数学和物理表示式,希望用这些表示式表述语句的深层结构,不仅如此,HNC 还试图穷尽这些表示式的类型。我始终把这一目标当做自然语言理解万里征途的第一步。第一步理论目标在五年前就达到了,但迈向征途第二步的理论努力也随之停顿了。当时我只想稍事休息一下,没有想到一休息就是五年,这些话将在论题 52 里续说。

从乔姆斯基短语结构语法开始的语法理论都以句法树为语句表达的基本框架,这种框架的结构单元是短语。一个句子由若干短语构成,每个短语充当一定的角色,各角色相互配合,构成一个结构完备的句法树。把句子的说法变成句法树的说法,为语句的分析和生成提供了可操作的意义,除此之外,几乎没有引进新的东西,短语不过是传统术语“主谓宾定状补”的统称,略有新意的是规定短语具有内部结构,而这一点在传统意义的“主谓宾定状补”里是不明确的。短语的概念不是对“主谓宾定状补”的深化,而只是形式化和结构化。

乔姆斯基从短语结构语法发展到转换生成语法,提出了所谓四元组的形式语言理论,把语言定义为集合 L = (S, Vn, Vt, P),乔姆斯基把 Vn 叫做非终止符,Vt 叫做终止符,P 叫做转换规则,使这个定义看起来似乎很玄妙,其实,只要明白 Vn 其实就是短语,Vt 就是词汇,P 就是短语结构,就一点也不玄妙了。所谓生成规则实质上仍然是语言的短语构成规则。

乔姆斯基的形式语言理论在人工语言方面的贡献本文不来评论,但应该指出,形式语言理论对自然语言理解并没有产生积极的作用,它的抽象不是减轻而是加强了传统研究中肤浅性的侧面。

后来菲尔墨试图把传统的“主谓宾定状补”概念引向深入,扩充传统“格”的概念,更具体地说,是希望对“主宾状”作出更细致的语义分类。“主谓宾定状补”虽然只是一个语法功能的粗分类,然而它是完备的。脱离谓语仅针对“主宾状”进行细分,完备性的矛盾突出了。菲尔墨及其继承者未能解决这个矛盾。

山克先生试图弥补菲尔墨脱离谓语的缺陷,从语句的语义类型出发重构短语的角色体系,然而他急于求成,仅仅理出了现在看来主要是转移句的一点头绪,就进入篇章处理,像我在【2】所说,匆匆忙忙在沙滩上建立高楼大厦,结果是半途而废。

13.2 HNC 的总体思路

HNC 在创立之初,即注意到了这一历史教训,认定必须在自然语言概念体系具有完备表述的基础上,才能建立完备的语句表述体系。因而它先从建立概念体系的完备表述入手。幸运的是,在这一过程中,我们发现了抽象概念的基元与基本之分及其五元组特征,发现了作用效应链,发现了概念基元的穷尽表述方式。在这三项发现的基础上,语句表达式的完备性问题也就迎刃而解了,这就是 7 个基本句类和 36 个混合句类的发现。

作用效应链所体现的主体基元概念,既是概念体系的基本分类,又是语句深层结构,即语句物理表示式的基本分类,同时也是各种词类(包括动词)的基本分类。

以主体基元概念的二级节点为依托,思考语句的结构,你就会获得登泰山而小天下之感。语句物理表示式的思想,语义块是句类函数的革命性结论,就会油然而生,于是,乔姆斯基关于语言不能‘well-defined’的名言需要修正了,传统语言学的‘主谓宾’,菲尔墨的‘格’,现代语法理论的句法树等概念,都有不切要害之病了。

这一系列发现过程得益于从语言层面到概念层面的升华,只有从数以万计的语言词汇中脱身出来,进入由三个超级概念基元网络,特别是主体基元概念局域网络所展示的概念空间,你的思虑才能获得必要的净化,产生 HNC 所引入的一系列新概念和新思想。这包括:有限句类的思想,语句及语义块物理表示式的思想,语言逻辑概念的思想,具体概念挂靠抽象概念的思想,层次符号的高中底三层次表达思想,概念表达的逐项展开思想,映射符号和反映射符号的思想等。

西方对自己的文化传统过于自满和自信,看不到西方文化的缺陷和不足,在知识表示问题上,特别是语义网络和一阶谓词逻辑的知识表示方面,表现得最为明显,他们总是依附于自己的母语,不曾思考过他们的母语并不是概念的最优表达方式,因而也不曾想过,自然语言理解需要为概念体系重新设计一套便于计算机使用的符号。这一局限性使得他们难以从语言层面上升到概念层面思考自然语言处理的本质。

我曾在【1】中说明,是中国的文化传统使我偶然捷足先登于概念层面的殿堂,这是一个十分有益的经验,所以愿意在这里再次向读者推荐。

上述‘有限句类、语句表示式、语义块、语言逻辑概念’是四项密切相关的思想体系,它以有限句类为基础,以语句物理表示式为核心,以语义块为表示单元,以语言逻辑概念为单元连接体。用建筑物来比方,句类是建筑物类型,语句物理表示式是建筑物本身,语义块相当于房间,但也可以扩展为子建筑物或由子建筑物蜕化而来,语言逻辑概念相当于门、过道和楼梯。

房间的类型首先取决于建筑物的类型,办公楼的房间不同于宿舍楼的房间,所谓语义块是句类的函数,就是这个意思。房间有独立于建筑物类型的自身分类标准,语义块也是如此,主块和辅块的基本分类,主块的 A、B、C 分类,辅块的 7 大类都是语义块的自身分类。A、

B、C 可称为块类符号。表征句类的语义块称为特征语义块 E,也属于主块。E 有 7 个基本子类,相应的符号是 :X,Y,P,T,R,S,D。由这些句类符号和块类符号构成语义块的物理表示式,这在【2】中已有详细阐述。

由语义块的物理表示式可构成语句的物理表示式,这在【21】中有详细阐述。基本句类的物理表示式是可穷尽的,这是关键性的结论。至于子类总数 57 之说,则不应该视为终极数字。这里应该强调的是,基本句类的每一子类,都有自己的句类知识,目前对这一概念层面知识的研究只能说是刚刚起步,随着这一研究的深入,有可能增加一些子类,但基本格局不会有大的变化。

HNC 的语义块的物理表示式是否就是菲尔墨的‘格’?

正确的回答是:‘是,又不是’。

说‘是’,因为两者都是对语义角色的表述。

说‘不是’,因为 HNC 有下列重大发展,这些是菲尔墨当年未曾想到也不可能加以解决的。

第一,菲尔墨的‘格’是不可分解的,他本人及其后继者似乎都没有想过语义角色也应有基元与复合之分,HNC 对此作了深入的探讨,其中 C 角色基元的提出具有关键性,由此产生块扩和句蜕的重要思想。

第二,菲尔墨的‘格’是动词的造句特征,仅涉及名词短语,或者更具体地说,仅涉及‘主谓状’的语义分类。他意识到独立于动词进行这一分类是有根本缺陷的,因而回避语义角色的完备性问题,表现出忐忑不安和无可奈何的心情。菲尔墨不可能跨入句类的殿堂。但是,山克先生完全有可能跨过这条界线,可惜他的冒进风格毁掉了这一线生机。HNC 的语义块是全方位的,包括谓语,是在句类层次上对‘格’的重新思考,由于穷尽了基本句类,使完备性疑难得到了彻底解决。

13.3 基本句类、混合句类和复合句类

这个题目在【21】和‘自然语言语句的 HNC 表示’一文(刘志文等,见本书附录)中都已有系统阐述,这里仅从深入研究的角度说明一些看法。

从语句三种基本类型的提出,到 57 个基本句类子类的确定和混和句类代码的制定确定,标志着一个研究周期的完成,最后一步有赖于联合攻关小组的共同努力,终于写下了一个圆满的句号。剩下的一些问题可以与下一阶段的研究结合起来进行。

那么,下一步研究的中心是什么?

1. 围绕着每一个语句子类表示式,阐发它的句类知识。在【14】到【20】中曾经作过这样的尝试,但那只是预研,现在需要向纵深发展,并最终形成可直接提供软件使用的概念层面知识。

2. 围绕着每一个语句子类表示式,列举它所优先的概念节点。【14】到【20】也这么做

过,但远不完善。此项研究当然不是当务之急,因为目前不处理新词,但理论研究应该先行,没有理论上的系统成果,动词的新词辨识及其后续处理将永远达不到大脑语言感知的水平。

3. 上面两项研究必须与语境相结合。为此,应及早开展语境分类的研究。而语境分类应从复合基元概念的局域网络入手。

本文到此结束。句类格式及格式代码,它们与句型的关系,软件如何利用格式知识等问题,放在论题 14 中讨论。

1998 年 5 月 27 日

再 论 “格”

14.1 引 言

在论题 13 里,我阐述了 HNC 的总体思路,阐述了句类及其物理表示式的来龙去脉,但觉得意有未尽。这里再作进一步的说明。

建立语句物理表示式的想法萌生于 1990 年,我当时的直觉是,对乔姆斯基关于自然语言不是一个 well-defined 的提法不以为然。从表观上看,自然语言确实具有无限和不确定性的特征,但在这无限和不确定性的背后,必有其有限和确定性的本质,否则人类的大脑怎能面对自然语言特别是语音的巨大模糊而应付裕如?幼儿的语言习得过程绝不是某些西方认知学家盲目跟随西方语言学家所想象的那样,是语法规则的习得,而是语句模式的习得。我这里说的语句模式,是指句子层面概念联想的模式,不是传统语言学的句型,也不是所谓的短语结构规则。在【2】中我把这个模式叫做句子的全局联想脉络。

为了建立这个全局联想脉络,HNC 提出了以主语义块构成语句的数学和物理表示式的概念,对语义块提出了特征语义块 E、广义对象语义块 JK 和辅语义块 fK 三分类的概念,对句类提出了基本句类、混合句类和复合句类的三分类标准,对基本句类提出了按作用效应链划分的基本标准,从而得到了 7 个基本句类和 36 个混合句类的重要结论,并进而穷尽了对基本句类一级子类和混合句类一级子类的发现,即语句物理表示式的穷尽发现。以语句物理表示式为基点,我得出了广义对象语义块 JK 是句类函数的结论,而句类可由特征语义块唯一确定(但可能存在多个句类代码),这样,我就完成了建立语句全局联想脉络的预定目标。语句全局联想脉络就是儿童语言习得的最终模式,尽管他们必然是从词语的习得起步,词语习得是音义转换的激活、扩展、浓缩与存储的初级过程。这一过程的进化将沿着从简单的词到语义块,从简单的广义效应句到广义作用句,从标准格式向非标准格式三条轨道前进,而不是西方语言学家心目中的那些语法规则。正是这一系列的思考使我认定,从语义块感知切入,进而辨识句类是对大脑语言感知过程的适当模拟,这就产生了句类分析的思想。显然,句类分析不过是这些理论结果所揭示的水到渠成的语句深层分析之路。

14.2 句类格式及其代码

对 HNC 联合攻关组的成员来说,句类格式及其代码是一个滚瓜烂熟的概念。但考虑到

本论题序列将以上网的形式广交朋友,这里还是从句类格式谈起。

句类格式定义为语句主语义块的排列顺序,也就是语句表示式的一种具体形式。一种句类有多少种格式决定于主语义块的个数,这似乎是一个简单的排列组合问题,例如 3 主块句应有 6 种格式, 4 主块句应有 24 种格式。但是,下面就会看到,实际的格式不是这么简单,语言还另有花样。

在主语义块的各种可能排列中,应该存在一种或几种语言所钟爱的排列。这并不是什么新设想,比较语言学家早就注意到了语言的这一钟爱现象。他们按照主语 S、谓语 V 和宾语 O 的排序对世界各种语言进行了统计,给出了下面的结果:

语言类型	比例
SVO	35%
VSO	19%
VOS	2%
SOV	44%

汉语被划归 SVO 型语言,但这是值得商榷的,因为汉语也经常采用 SOV 和 OSV 排序,还偶然采用 OVS 排序。另外,汉语还经常采用省略 S 的 VO 排序和更加骇人听闻的省略 V 的 SO 排序。这些情况说明, 对汉语语义块的排序问题需要进行汉语特色的研究。为此,需要提出一些新的概念,主要是语义块指示标记和语句格式的概念。

为了说明这一点,让我们暂且借用一下主谓宾的术语,拥有主谓宾三种成分的句子采用 SVO 顺序最符合效率原则。为什么?因为 V 是天然的分隔标记,这样的排序可以免除在 SO 或 OS 之间另加标记。汉语是单音节语言,这一免除的价值非同小可。我推测,凡 VSO, VOS 或 SOV 语言,都要在 S, O 或 O, S 之间加上某种标记,这一推测得有劳比较语言学家去验证,我只知道日语确实如此。

汉语基于效率原则基本采用了 SVO 排序。但它不像别的语言那样死板,也采用其他的各种排序,这时,它使用语义块标记这个法宝。汉语的标记具有更鲜明的语义信息,而不是像日语的标记那样,主要是语法信息。这一点希望精通日语的朋友去研究一下,我这里算是姑妄言之。

由于汉语的语义块可采用各种排序方式,我们认为有必要对语义块排序方式给出明确的定义,以利于计算机的理解。于是引入了标准、规范、违例和省略 4 种基本格式的概念。同时,为了便于说明这些基本格式的各种变化,又引入了语句数学表示式的概念。

语句数学表示式是语句物理表示式的抽象或再形式化(参见【21】;刘志文等,1998。见本书附录),对基本和混合句类可写出语句的统一表示式(除了 S04 句类)如下:

$$Jn0 = JK1 + E + \sum_{i=2}^{n-1} JK_i \qquad (1)$$

这个表示式就是语句的标准格式,特征语义块 E 一定在第二号位置, 广义对象语义块 JK 的

编号与物理表示式中的排序严格对应,也就是说,JK 的数字后缀 i 表示广义对象语义块的序号,而左边语句符号 J 的第一个数字后缀 n 表示语句的主语义块个数,第二个数字后缀 0 表示标准格式。对于 3 主块和 4 主块的基本句类,上式可分别写成:

$$J30 = JK1 + E + JK2 \quad (2)$$

$$J40 = JK1 + E + JK2 + JK3 \quad (3)$$

(2) 式可以有 6 种排列方式,(3) 式可以有 24 种排列方式。例如,(2) 式的另 5 种排列方式(格式)如下:

$$J31 = JK2 + E + JK1$$

$$J32 = JK1 + JK2 + E$$

$$J33 = JK2 + JK1 + E$$

$$J34 = E + JK1 + JK2$$

$$J35 = E + JK2 + JK1$$

J32 与所谓 SOV 语言相对应,J34 与所谓 VSO 语言相对应,J35 与所谓 VOS 语言相对应。这些格式的后 4 种都出现了广义对象语义块 JK 相邻的情况,这时,在两者之间加上一个仅有分隔意义的标记是不够的,因为,这将造成 J32 与 J33,或 J34 与 J35 的混淆。对这个问题,可以有两种解决方案,一种是对这个分隔标记赋予某种语法甚至语义信息,另一种是只给简单的分隔标记,同时在 E 块的形态上(包括屈折和黏着)加以补充说明。当然,这两种方案可以结合起来使用,日语似乎就是这样。汉语由于缺乏屈折和黏着的手段,采取第一种方案比较自然,这就是汉语语义块标志符号比较发达、而且不仅具有语法意义的原因。语言逻辑概念的'本体 + 挂靠'结构正是基于这一认识而设计的。本体层大体代表语法意义,而挂靠层则完全代表语义信息。

现在,可以给出所谓 4 种基本格式的定义了。

标准格式就是各语义块严格遵循物理表示式约定顺序的格式,包括这里的数学表示式(1)、(2)、(3)及表示式 J31。

规范格式就是语义块顺序违反了物理表示式的约定,但在两相邻 JK 之间都加上语义块标记的格式。

违例格式就是不仅语义块顺序违反了物理表示式的约定,而且在某些或全部相邻 JK 之间未加语义块标记的格式。

省略格式就是省去了物理表示式中某个或某些应有的语义块,省略格式本身当然又会有标准、规范与违例之分。

标准格式之外的各种格式也统称非标准格式。

HNC 对各种基本格式及其各种变形都给出了确定的编码,这个编码具有穷尽性的特征,这样就为计算机透过汉语语义块排列顺序复杂变化的表象,把握汉语语句的深层结构提供了必要的保证条件。

汉语的 E 块经常出现分离现象,非标准格式下的 JK 也经常出现分离现象,这会给汉语

理解处理带来一定困难。但应该指出,西语的 JK 分离现象远比汉语严重,在著名例句“I saw a girl with a telescope”中的歧义模糊,汉语反而是不存在的。

传统语法学关于句型的研究有很多宝贵的成果,我们应该把这些成果吸收到格式的框架里来,充分发挥它们的作用。

1998 年 5 月 29 日

三论中西语言的基本差异

——一论音节感知

汉语的每个音节都有韵母,当然,这不包括打电话和口语中常用的 m, n, ng 等音。每个音节还有不同的调,调用于区别意义。汉语的这些独特现象是人们所熟知的。

但是,对于这一独特现象背后的语言学问题,人们并没有作深入的思考,汉语理论语言学家照搬西语的音义两极的说法,丝毫不感到有什么不妥。汉语拼音方案曾试图废除汉字,走全拼音化之路。最早提出废除汉字口号的是钱玄同先生,钱先生是新文化运动的先驱者之一。这些先驱者的历史功绩已有充分的评价,但许多先驱口号和思潮的片面性和消极影响则有待今后历史学家的研讨。

汉语的每个音节通常都有多个汉字,每个汉字通常又有多个意项,因此,汉语的每个音节通常是一个巨大的意义集合,是一个超级多义模糊集。但是,人们在用汉语交流时,除了听者不熟悉的人名、地名,并不感到这个超级模糊集的威胁,大脑(仔细说并不全是大脑)的语音感知系统应付裕如。对这一关键性的神秘现象我们首先应该思考的是:语法或句法能起多大作用?

当然,这涉及语法或句法的定义问题,如果将语法说成是语言的法则,句法说成句子的法则,那就是外交词令。我认为,至少源于西语的语法体系始终回避了一个根本问题,那就是语言的实质。语言过程是"概念联想脉络的建立、激活、扩展、浓缩和存储",这五个环节缺一不可。我并没有用这个说法代替现有关于语言的各种经典"定义"的意图,但是,我确实认为,要建立计算机的自然语言智能,就必须对语言的实质作上面的表述,并以此为基点,开展计算语言学的研究。这就是 HNC 的探索之路。

回到上面的神秘现象。这个现象可命名为汉语的单音节感知,西语基本不存在这个问题,因为西语的最小可独立使用的语义单位绝大多数是多音节的。而现代汉语基本上是单音节和双音节平分秋色(古汉语则是单音节为主)。汉语的这一"单双"现象是本文要论述的中西语言的基本差异之三。

近十多年来,中文信息处理花了很大的力量进行汉语的分词处理,为此还开展了分词规范的工作。十年过去了,是否有必要作一点反思?是否应该想一想,汉语的词与西语的词有什么根本不同?规范分词是否有点多此一举?

从古汉语的以单音节词为主到现代汉语的单双并重反映了汉语的演进过程,伴随这一

过程发生了汉语‘字义基元化,词义组合化’的深刻变化。组合化的汉语词语就必然具有不定形的特点,这一特点与人类思维的创造性特征是相适应的,很难也不应该为了计算机处理的需要而加以规范。首先,规范的效应既违背汉语的特点,也限制人类思维的创造性发挥。其次,所谓计算机处理的需要密切依赖于处理的策略,源于西语的句法分析策略确实是需要的,但 HNC 处理策略则不需要。因为,分词并不是 HNC 处理过程的‘瓶颈’,而是‘瓶底’。第三,规范不可能彻底,单字词、双字词、多字词的模糊界限通过规范能够有所减少,但不可能消除,处理过程仍然必须面对这一模糊界限的消解问题。

因此,根本措施是解决汉语单字词的处理问题,对语音输入来说,是单音词的感知问题。

这是汉语特有的问题。汉语心理语言学应该开展此项研究。但由于现代汉语语言学从一开始就对汉语的根本特点视而不见,或见而不全,当然也就不能指望心理语言学界会开展汉语音节感知的研究。

音节感知包括调的感知,调在连续语音流中是有变化的,但人的听觉(当然不仅仅是听觉)能适应这一变化。但这个适应是有限度的,对这个限度很值得进行实验研究,目前当然是空白。

语音识别有音调识别的游离研究成果,但没有一个系统把音调识别和语音识别集成到一起。汉语语音识别的传统是既脱离理解,也脱离音调识别,一切照搬西方的路子。殊不知在语言这个特殊领域,照搬只能写写凑数的文章,既不能在理论上解决汉语特有的问题,也不能在技术和工程上形成自己应有的优势。

关于汉语的音调,四声的划分是对孤立单音节发音而言的,连续语音流的音调变化,应区分理论研究与工程应用。理论研究应着眼于细分,而工程应用则应着眼于粗合。粗合的具体化就是把实际的音调粗分为去声和非去声两大类。这种区分既具有技术实现的可能,又有重大的实际价值。

从音节感知来说,汉语的四声并非同样重要,去声和非去声大体上平分秋色。“重大实际价值”就是基于这一点。对这一点,我没有作具体的统计研究,只作了粗略的观察。去声和非去声之分就是传统音韵学的平仄之分,平仄在唐诗宋词中的实际应用显然支持平分秋色的推断。

在一次讨论会上曾有人问我‘HNC 能处理唐诗吗?”

我的回答是:唐诗应该比现代汉语更容易处理,因为,唐诗最规范,计算机善于利用规范知识,而不在乎文体是文言还是白话。

唐诗的语音流有明确的平仄信息,违背了平仄节律,就不是唐诗。现代诗人由于古汉语根基薄弱,有时写旧体诗又不遵守老规矩,毛泽东主席都表示不妥。我上面的回答当然隐含着平仄信息确知的前提。我认为,平仄识别、汉语的平仄节律和汉语的‘单双’现象,是汉语语音识别的三大特殊问题,应予以特殊关注。“八五”攻关期间我多次为此发出呼吁,可惜未引起应有的重视。

音节感知是口语的需要,口语除了韵律信息,还有情景信息,后者包括谈话人的姿势和

表情信息,有人为了突出口语与书面语的差别,过分强调了情景信息的作用,这些人忘记了一个基本事实,人们在听谈话录音或现场广播时,并没有由于情景信息的丧失而感到特别困难(至于录音效果不佳,那是另一回事)。韵律信息的作用大于情景信息,它类似于句类格式的变化,主要是强调性意义。基本的语音信息仍然蕴涵在各音节之中。情景和韵律信息对音节感知肯定大有帮助,但不是决定性的。听录音也许需要听两遍,那就是因为失去了这些帮助。

韵律信息中包含多少中文信息处理特别关心的分词信息是一个有趣的问题,还没有人对此进行过研究。不难想见,在规范的语音信息中当然存在有益于分词的韵律信息,但是,在不规范的随意的语音流中,可能存在不利于甚至有害于分词的韵律信息,但大脑的感知都能应付这些干扰,既能利用有益的信息,又能排除有害的信息。计算机的自然语音感知,要利用韵律信息,必须限定在规范语音流,这是毫无疑义的。

但是应该指出,所谓规范语音流的规范本身就是一个模糊概念,有点易意会而难以言传的味道。与分词规范类似,你不可能要求讲话者严格遵循这一规范。因此,我的看法是:韵律信息只可参考,不可作为硬性的规则来使用。

这里顺便说一下规则问题。软件偏爱硬性的规则,即可变成简单产生式的规则,但我希望用优先代替规则,优先是软性规则,而不是硬性规则。两者区别在于硬性规则是绝对规则,而软性规则是相对规则。绝对性表现为:如果产生式左边的条件全部满足,则右边的结论一定成立,这里要求全部满足,如果只是部分满足,结论就被否定了。相对性规则不同,即使条件全部满足,结论也不一定必然成立,它上面还有语境的约束。另一方面,如果条件部分满足,结论可暂时保留而不完全否定。

相对性规则的这种约定,似乎使软件无所适从,变成了无法使用的规则。问题在于:第一,软件要学会在无所适从时如何应付。第二,软件要学会随机应变。HNC 核心软件如果在这两点上始终不能有所进步,那么,HNC 符号系统所提供的知识是不可能得到充分应用的。

无所适从就是不能作出确定的判断,例如,对双音词模糊集"使用试用适用实用"或"使用试用"适用实用"一般不能彻底消解。这时最简单实用的应付方式就是请求用户帮助。但是,软件必须把用户的帮助当做学习的机会,即记住用户的选择,在下一次出现同样情况时,软件按照用户的示范如法炮制。这是最简单的学习方式。进一步,应该同时记住用户选择时的语境,这样的学习结果就可以进入长期记忆,变成自己的知识。而简单学习结果只能放在短期记忆里。

对 HNC 处理技术早已作了上述筹划,但迄未实行,现在已到了刻不容缓的时候了,应尽快制定实施策略。

所谓随机应变,主要是语境知识的运用,【3】中关于伪词"岳飞"的讨论就是对随机应变的具体说明,这个问题将在论题 39 中作进一步阐述。

软性规则运用的极端重要性不言而喻,目前我只能说这么一点一般性想法,不能深入

讨论,仅凭 15 年前黄河项目鏖战时的那点经验会弄出班门弄斧的笑话。

回到汉语的音节感知。上面,我强调了平仄区分对独立音节感知的重要性,同时淡化了韵律知识的作用,这是我的直觉判断。但这一点对于制定音节感知的策略十分重要,所以写了上面许多的话。

对汉语拼音键盘输入,平仄区分是不现实的,从这个意义上说,如果语音识别能够解决平仄问题,那么拼音输入理解处理的难度就不会小于语音识别了。

音节感知的本质是进行音义转换。形式上它是多义选一处理。但一个音节的义项可以达数百之多,大脑的感知过程绝不可能是对每一义项依次与前后文信息作匹配处理,而后选取最大值。

那么,大脑如何实现音节感知?理论上说,就是概念联想脉络的激活,具体地说,激活有内外之分,或自激与受激之分。自激是以某音节为激活点,向外激活联想脉络。受激是从某音节之外的激活点出发,沿着联想脉络确定受激音节中的相应概念,从而完成音义转换。

区分这两种概念激活类型是音节感知的关键。而汉语之所以必须建立音节知识库,主要就是为概念激活的类型,首先是对自激性音节提供指示信息。

音节外的激活点有三种基本类型,一是无模糊的双字词或多字词,二是有模糊的双音词,三是另一个单音词。这三类激活点构成三种受激类型,分别称为甲乙丙型激活。这三种基本类型及其组合可自由分布在待感知音节的前后,感知过程的复杂性依赖于受激类型。

自激情况主要是语义块感知的激活。汉语规范格式语句广义对象语义块 JK 的激活基本依赖于音节的自激,因为,JK 的标记都是单音词。辅块标记单音词与双音词大体相当,但是对条件辅块以及两可性的目的对象语义块和参照内容语义块,激活方式仍以单音词为主。上述两可性语义块在论题 34 中有详细阐述。

这里只简单说明一下音节感知处理的要点。自激情况的音节感知属于特殊处理问题,而受激情况的音节感知属于层选和段接的问题,都超出本文的预定范畴了。

1998 年 4 月

二论音节感知——段接处理

23.1 引　　言

汉语的 HNC 理解处理以音段为单位,这就自然提出了段接处理的要求。

所谓段接处理是指对音段两端单音词(字)的处理。这里说的处理有两层含义:第一是确认前提,即确认该端点是不是单音词。第二是确定该单音词的组合方向,是向前组合还是向后组合?这也可能是层选的任务,即音段内部单音词处理的要求,这表明层选与段接是不可分开的。

层选与段接的共同基元处理模块就是"确定单音词组合方向"。两者的性能主要取决于这一模块的水平。此模块的根本目标是保证不出现单音词的孤魂。

因此,本论题将主要讨论"确定单音词组合方向"的有关知识和规则。

这里应顺便说明一下,上述两层含义不一定同时存在,例如单音段就不存在第一层含义,而多数 1 类音节不存在第二层含义。

23.2 段接类型和单音词组合方向的有关知识

首先应该说明一下段接的类型,应区分指定字段接,单音词段接,单音词连接,混合段段接,两奇段段接,奇偶段接,两偶段段接 7 种不同情况。本文需要讨论的只是前三类段接,后四类段接是层选的自然扩展。本文不来讨论。

指定字段接是指"的了和不"的段接。

单音词段接是指对单个单音词的处理。

单音词连接是指对多个相连单音词的处理。

指定字段接在论题 23-1 中讨论。

对于单音词的段接和连接,首先要理顺总体思路。

总体思路的要点是:建立段接因子及其分类的概念。并根据音节知识库,从汉语的全部音节中,选出一些最常用的段接因子。这个要点的实质是对汉语音节 8 种义类的已有划分,按照确定组合方向的需要进行再分类。

这是一项繁重而不寻常的物理工作。没有这项物理建设的支持,段接处理就是无源之水,无本之木。必须有人承担起这项任务。

段接因子应分为前组合、后组合、并合及综合 4 类。

前组合因子包括量词、泛指基本命名、后语素、语言逻辑概念 lh5。

后组合因子包括前语素、特指基本命名和某些基本概念。

并合因子指语言逻辑概念 l4。

综合类是指含有上列两类或三类因子的情况。前三种情况可称非综合类。

这几条,就是所谓的组合方向知识。

段接处理欢迎非综合类,但有多少音节属于这一类?

对这个问题的回答必须是动态的,也就是说,必须先明确是在句类分析三步曲的哪一步进行段接处理,还是那句老话,见机行事。而不能仅仅基于那个静态的汉语音节感知知识库。例如,在句类假设阶段,往往需要作 E 团块处理。当两动词之间仅相隔一个音节时,你(软件设计者)必须优先考虑该音节是否具有并合因子。如果该音节为 bing,则它就从静态的综合类转变成动态的并合类,按汉字‘并’的义项 l4 进行局部检验。有的读者可能不同意这个例子,认为这不是静态到动态的转换,我对此不作辩解。

强调见机行事,不是说静态分析就无所作为。相反,此项物理建设必须从静态分析入手,而且可以大有作为。汉语仍然有一些音节具有静态非综合特征,首先要把它们抓住。段接处理的成长过程需要精心设计,要像抚育婴儿那样考虑自然规律,而不能拔苗助长,这是我一年来最深的感触。这个‘精心’要靠设计者自己的力行,但仅靠力行是不够的,要懂得精心之秘诀在于静心和深思。深思才能站得高,看得运,才能脱离‘只见树木,不见森林’的狭隘,才能上升到概念的高空俯视一切,让‘自然的最高立法’(康德语)活跃在你的心中。这就是我想说的话,没有更多的建议了。

23.3 段接处理规则漫谈

自然语言理解软件设计者心中的上帝是规则,对此我曾很不以为然,因为我心中的上帝是概念联想的激活,是句类知识,是同行优先。后来我萌生了软性规则的说法,也就接受‘规则即上帝’的观念了。

然而本节仍以规则漫谈为题,因为具体规则的制定和调用,终究是软件设计的本职,而不是物理分析的职责。物理分析提供制订规则的知识基础,但不能替代工程操作。另一方面,软件设计应有自己的创造,不能也不应该指望从物理分析那里拿到现成的全套规则。

这个漫谈将着重阐述规则的规则,列举下列几条。

1. 不容许出现单音词的孤魂
2. 抓住并合信息
3. 抓住特指、泛指、特泛指相互之间和自身的搭配
4. 抓住包含性概念的搭配
5. 抓住时间、空间、数量、数与数的内部搭配
6. 抓住一些最常用的基本概念

7. 抓住几个最常用的块尾标记

规则之规则应围绕着两条,一是必须抓住什么,二是必须禁止什么。我觉得,上列一禁六抓是当前段接处理的纲。它的意义当然不限于段接,也包含层选,因为如上所述,两者是不可分的。

规则 1 是一切单音词处理的灵魂,包括层选、段接和新伪(指新词和伪词)处理。这条规则是‘不容许出现概念的孤魂’这一最高规则的子规则。HNC 曾为块内处理提出过唯一的一条总规则,叫做同行优先。“不容许出现概念的孤魂”实质上就是‘同行优先’的另一种说法。

HNC 符号体系挖空心思去揭示概念的同行性,为什么?因为同行性揭示得越充分,联想歧途的孤魂就暴露得越明显。孤魂 = 不合理 = 抑制,同行 = 合理 = 激活。这就是 HNC 提供给计算机的基本判断公式,即最高规则。同行与孤魂是针对块内处理的一对概念,预期与意外是针对块间处理另一对概念。它们是激活与抑制的两种基本机制或类型。我不能说大脑的自然语言理解处理就是采用这两种机制,但我可以肯定地说,计算机要模拟大脑的语言理解处理过程,必须从这里起步。

当然,由一个词或一个简单概念构成的语义块无所谓孤魂,孤魂是复合构成语义块的孤立奇点,在块内孤零零的,没有同行的朋友。就音段处理来说,对于较长的音段,错误的层选可能造成一群孤魂,而正确的层选最多出现一个(对奇段)或两个(对偶段)边界单音词孤魂,然后靠正确的段接使之不孤,并同时保证不造成新的孤魂。这里层选的标准是奇段一个单音词,偶段最多两个单音词,如果所有层选的孤魂数多于上述约定,则表明有伪词存在。就块内处理来说,则不容许任何孤魂的存在,它只讲有无,不管数量,有孤魂就表明原句类假设有误。这就是运用规则 1 的要点。

这一切判断都要立足于孤魂的认定。孤魂如何认定?有 6 种认定标准,第一是联想脉络标准,它包括同行和预期两个侧面。第二是语境标准,它包括专业范畴、知性层次、作用效应链的内在顺序及环节侧重三个方面。第三是语法标准,包括概念类别搭配和习惯搭配两个侧面。第四是语体及文体标准,第五是情景标准。第六是常识标准。

概念知识库提供形成第一和第二种标准的一般知识及通用规则,词语知识库提供第一和第三种标准的知识和规则。前三种标准的知识都蕴涵于 HNC 符号体系之中,但需要通过概念和词语知识库的转换形成规则性知识。当前的概念知识库仅涉及第一种标准,应尽快开展第二种标准的制定。语法标准也要有脱离于词语的通用规则库,在性质上也属于概念层面知识库,但考虑到语法规则的语种个性,我们将语法规则库独立于概念知识库。后三种标准,HNC 符号表示作了一定的努力,但只是杯水车薪,主要靠今后建立相应的专用知识库。

一般读者对上面的论述肯定会产生‘标准的则使用甚易,前提确认极难’的感觉,但联合攻关组成员当不致如此。大家都在为建立这些认定标准而艰苦工作,对这个总体目标是必须昭昭而不可昏昏的。

最后看一个孤魂的例子。

zheng zhi jia men jin guan zhi dao ju shi wei xian，

政治家 | 门禁 管制 * 刀具 * 市委 *

尽管 知道 * 局势 危险 *

政治家 | 们…尽管知道 * 局势 危险 *

这 12 个音节如果“门禁”参与分词，就形成由虚切口构成的两音段，一个三字词和一个九音节奇段。如果对这个奇段分别取奇段上层或下层，就分别出现边界单音词孤魂 xian 和 men，但 men 可通过与“政治家”段接而不孤。这就达到了正确层选。如果“门禁”不参与分词，就形成三个音段，处理过程要简明多了。

规则 2 对语音文本抓之甚难（对文字文本甚易），但又势在必抓。当前最重要的是处理好“必为与不为”的统一。指定字“的了和不”，指定音 ji、yu、tong、gen、de、bing，顿号“、”是必为之首选。

规则 2 的并合包括传统语言学所说的偏正、联合和后补三类组合。上列 4 个指定字和 6 个指定音，有的只涉及一种组合，有的涉及两种或三种组合，甚至还有其他意义，如 de，它还有引导作用效应句的功能。对它们的详细阐述，另有专文。但这里应强调指出，规则 2 的“抓”着重于联合结构对仗性的利用，这个问题以前曾阐述多次，杜燕玲应承担起总结提高的责任。

规则 3 和 4、规则 6 和 7 都将另有专文，规则 5 已有专文。

1998 年 7 月 1 日

三论音节感知——偶段处理

夫兵形象水,水之形,避高而趋下;兵之形,避实而击虚。水因地而制流,兵因敌而制胜。故兵无常势,水无常形。能因敌变化而取胜者,谓之神。故五行无常胜,四时无常位,日有短长,月有死生。

孙武《孙子兵法》

偶段处理需要很好的策略。关于这一策略的主要方面,在提前成稿的论题 26 中已有阐述。这一策略的要点就是见机行事,所以篇首引录了《孙子兵法》的一段话。这段话本来打算放在论题 25 前面,后来觉得放在本论题之首更为恰当。

偶段中可能存在激活音节 l0、l1 和 QE,但被伪词掩盖了。在语义块感知阶段对这些被掩盖的激活信息一律采取按兵不动(有所不为)的策略,这就是"水因地而制流,兵因敌而制胜"。我们曾统计过,l0"把被向对为"出现在奇段和偶段的比例大约是 9 比 1,单音节 l1 和 QE 尚未统计,需要补做。但无论统计结果如何都不应影响在语义块感知阶段偶段处理的上述既定策略。

现代汉语的书面语言以双字词为主,由单个或多个双字词构成的偶段是常见音串。但偶段中可能存在伪双音词,这时就需要进行消除伪词的处理,简称消伪处理。偶段消伪意味着要找出偶数个单音词,至少 2 个。奇段也有消伪问题,要找出奇数个单音词,至少 3 个。这里需要说明两点,第一,奇段中至少存在一个单音词,找出这个单音词乃势在必行,这属于层选而不属于消伪。第二,我们约定,多字词独立成段(见论题 1-1),因此消伪只涉及单音词。所谓单音词、双音词、多字词在语言学的意义上都是模糊范畴,但工程意义上不容许模糊,其双音词、多字词以工程所依附的词典为准。这样做可以净化思路,从而有益于调度设计,把实际的模糊消解问题放到合适的时机去解决,就是在句类分析三步曲中的第三步。

这就是说,偶段不存在层选问题,而奇段必须进行层选。两者都可能需要进行消伪处理。所谓偶段处理实质上就是消伪处理。

但是应该明确,上述策略不等于说,在语义块感知阶段对偶段不进行任何消伪处理。如果偶段后紧跟非串号前的指定字"了",而该段最后的双音词不是或不含动词,那就明确无误地表明该偶段有伪词存在,而且伪词的位置之一已经确定,这时应打破常规及时进行消伪处理,也就是及时找出单音动词。因此,单音动词并非全在 K 调度中处理。

上面的阐述表明,偶段处理的见机行事之"事"就是消伪处理,而"机"就是条件。消伪及其条件有不同的类型,依条件出现的先后列表如下,列举了 6 种情况的偶段处理。

阶段	条件说明	消伪类型
语义块感知	单字词‘了’前面无双音动词	找单音动词
句类假设	音串最后为不带‘了’的广义作用型 E 团块，前面无相应的 l0，不属于常见违例。	找相应 l0
K 调度准备		找下层双音动词
语义块构成	两侧有孤魂，向偶段求援	偶段伪词处理
	一侧有孤魂，向偶段求援	同上
	偶段有孤魂	同上

第一种情况不必说明。

第二种情况的关键在于对非违例的判断。汉语的违例格式并不是完全随意的，违例类型与句类密切相关，这是一项重要的句类知识。应据此在概念知识库中建立一个常见违例表。此表即非违例判断的依据。

第三种情况在论题 26 中列为 K 调度的准备项目之一，当然也可以作为 K 调度本身的项目。

后三种情况都属于句类分析三部曲的第三部。表面上看，似乎偶段处理与句类检验无关，但正如论题 31 中曾指出的，三部曲的第二和第三两部不是截然分开的，第六种情况的偶段孤魂就可能来于句类检验的提示。

伪词处理将在论题 30 中说明。下面看一个例句。

其他…商业…银行＊、保险＊公司积肥…银行＊金融机构…

…也…要…吸收＊高效＊…毕业生，

此音串惟有双音词吸收（洗手）前面有 QE ye 和 yao，因此，对动词团块“银行、保险”和“积肥…银行”都可以不予考虑。三个主块，标准格式，形势简明。基于这一形势判断，E 要素双音词吸收＊的模糊立即可解，因为吸收要求三主块物转移句 T21J，而洗手要求两主块作用状态句 XS＊11 或复合句类。JK2 和 JK1 的要素“毕业生”和“机构”符合要素检验要求，顺利进入语义块构成处理。

“毕业生”前面的双音词高效（高校），由同行优先而选定高校。JK1 范围的顿号按 JK1 的核心并列处理（这是语法知识），从最后的机构可知，这些并列的核心优先概念类别 pe，银行和商业银行，公司和保险公司，银行金融机构都因满足同行要求而入选，于是，第一 6 音段中的积肥就成为明显的孤魂了。

对这个孤魂的处理，我想不用再说什么话了。只想提醒 HNC 攻关组一声，这里对孤魂的处理就需要运用语法学关于顿号的知识。在句类分析的基础上，一运用这项知识，此句伪词的辨识和分解立即迎刃而解。对句法分析已有成果的利用，HNC 应该十分重视，不可有丝毫的意气之争。

1998 年 7 月 5 日

论调度及 K 调度

论题 25　34 为 52 个论题之第六组课题,中心问题是语句处理的基本策略。

传统的句法分析从短语类型入手,这一处理策略的立足点是:句子由一系列短语拼接而成。不同短语充当不同的角色,建立起短语与句法成分,短语与"格"之间的对应关系,就完成了句子的句法-语义分析。

这一处理策略在形式上对西语是适用的,因为西语有丰富的短语标记,即使如此,它仍然不是模拟大脑语言感知过程的适当模式,因为,正如我们在论题 13、14 所指出的,传统语言学的语义格还不能充分反应语句要素之间的概念关联性。

HNC 理论引入的以句类为变量的语义块函数才充分反应了各语句要素之间的概念关联性,把握这一关联性是贯穿于语句分析全过程的红线。汉语述语的辨识、词的多义选一(包括汉语单音词巨大模糊的消解)、语义块构成的歧义消解都必须以这条红线为基本依托,这是语句分析的战略性依托,而不是战术性依托。这一红线的基本内容就是〈概述〉一文所指出的句类知识。

句类知识以句类格式为纲,它决定语句感知过程的开始、过渡和结束。

HNC 定义了句类格式的 4 种基本类型,大家应注意到标准格式的 E 块都安置在第二位置,这种安置有利于尽早确定句类。在省略格式中最常见的 !310 格式将 E 块置于句首,也符合这一"尽早"原则。那么,常用的规范格式和违例格式为什么反而将 E 块后移?我认为,这里面隐含着语言表达的"效率原则",语句之妙在于 E,对 E 的选词最费斟酌,放在最后有利于赢得更多的缓冲时间,而不至于明显影响整个语句的输出效率。

4 种格式的运用与句群序列有关。句群之首通常应采用标准格式,非标准格式通常拥有继承信息,因而不宜在句群之首使用。"继承"是指前面的语句已经为后续语句提供了部分 JK 的有关信息,因而后续语句可以用代词表示有关的 JK 或其块素,甚至可完全省略,这是违例格式和省略格式对语境知识的活用。

对句类格式的是否标准,对省略格式的 !310,对规范或违例格式中的 JK2(JK3),都应保持高度警惕,这是调度的要点。我们强调(E,l)联合感知,正是这一要点的体现。

对于无句蜕的基本句类,即使是不带调的拼音流,通过(E,l)联合感知,不难确定句类格式。这里说的"确定",是指对多种假设的检验。检验的含义将在下一论题中详细阐述。但这里有必要对与调度有关的假设与检验问题先作一些要点说明。

汉语的 l0 主要是单音词,使用 l0 的句类只能是广义作用型句类;l0 的出现意味着规范格式,l0 必在 EH 之前;E 存在 EQ + EH 的主体构成,且两者经常分离;在基本句类中,仅

有作用效应句存在两 E 块,否则一定是复合句类;复合句类的 E1 句很少采用省略格式,而 E2 句一定采用省略格式;复合句类的子句不采用广义效应句,这一点可作为句蜕与复合的重要判据。

以上七点,是实现智能调度的基本手段,是对"假设"与"检验"的基本约束。

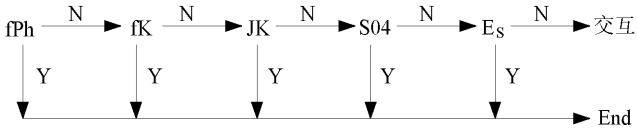
上述调度过程必须假定 E 的存在,而这里的 E 必须是双音词。因为,在语义块感知阶段,你必须先回避单音动词,否则将陷入"草木皆兵"的困境。

如果该语句确实采用了单音动词,则可能出现两种情况:(1)未发现 E 块;(2)伪双音动词干扰。

对情况 1,可直接转入 K 调度寻找单音 E 块;对情况 2,则需通过句类检验排除伪词干扰之后,才能转入 K 调度。这就是 K 调度的两个入口。

但是,无双音 E 要素的语言串(串以逗号为终止标记),不一定必须有单音 E 要素,它可以是一个插入语 fPh,一个辅块 fK,一个广义对象语义块 JK,或者是一个无 E 块的 S04 句类。因此,K 调度的操作要在排除这 4 种可能之后,才能进入单音 E 要素 E_S 的感知。

但这不等于说,K 调度就是按下列判断过程进行死板的顺序操作:



因为,当接收到串标记",",时,"黑板"上可能已有前三项之一的"记录",这时,可直接进入相应的检验。这里需要指出的是:以",",标记的 JK,多数情况是句蜕块,甚至可以说理应是句蜕块。fK 也往往是句蜕块。

本论题要点大体如上。"调度的灵魂是数据驱动",这句"老生常谈",是本论题最恰当的结束语。

论句类检验

26.0 引　言

句类检验是本论题序列第六组的的中心论题,共十个论题。这里先说明一下这十个论题相互之间的关系。论题 26 是总纲,紧接着的两个论题 27 和 28 是句类检验的两种特殊情况,即块扩与句蜕情况的句类检验。论题 29 和 30 是句类检验的两项重要目标;论题 31 是句类检验的必要延伸;论题 32 和 33 则涉及句类检验的两个基本侧面。一头一尾的 25 和 34,则涉及调度及其相关知识的运用。

本文分六节。第 1 节列举句类检验的准备操作,指出当前的弱点所在。第 2 节说明如何加强当前的层选处理。第 3 节讨论动词连见处理。第 4 节讨论假设类型的确定。第 5 节讨论句类检验的两种基本类型,其部分内容将力争以我最不擅长的通俗方式来说明,由于句类检验是当前最薄弱的环节,不得不出此下策,希望能在较短时间里改变这一危急状况,而此文能略尽绵薄之力。第 6 节说明句类检验所运用的知识。

读本文时,一定要熟悉论题 1-1 中所介绍的术语。

26.1 句类检验的准备操作

句类检验是句类分析三部曲的第二部,上承语义块感知与句类假设,下接语义块构成处理。三部曲的大体分工在理论上是清晰的,但分段层选进程中遗留的疑点,检验阶段面临的所谓深度优先和广度优先的矛盾,是两个突出的问题,本文将分别在第 3、第 4 和第 5 节讨论。

所谓疑点处理只是句类检验的准备操作的一个侧面,对这个侧面实际上要分解为一系列的具体操作,并不存在单独的疑点处理环节,全部准备操作可概括为下列 10 点:

1. E 排除处理
2. 动词连见处理
3. 假设类型处理
4. 句类格式处理
5. 句类转换处理
6. 块扩或句蜕处理
7. 语义块分离处理

8. E 块构成的精确定位处理

9. 辅块精确定位处理

10. 确定要素检验的级别

当前软件对这 10 项处理的 4 7 项都已有一定基础,弱点在一头一尾。针对这一情况,本文着重讨论 2、3 两项处理,第 1 项已在论题 2-1 中说明。

语义块感知阶段所依附的基本操作是分段,并进行必要的层选处理。但是,如何界定这个“必要”?层选处理不可能一步到位,经常遗留一些疑点,这些疑点何时处理?如何处理?

当前急需对这些问题作深入和透彻的思考,然后通过讨论找出解决方案。这个方案必须是全面的和彻底的,同时又有所不为。本文试图理出一个思考的头绪,提出一些基本设想,为正式讨论和制定最终方案提供引导。

这些问题并不是现在才提出来的,也不是过去没有提出过解决方案,问题症结在于没有形成文档,因而在理论设计与软件设计之间出现了很大的漏洞,在层选问题上表现得最为明显。从现在开始,这个漏洞必须堵住。

关于第一阶段的层选,过去约定了三条:

第一、奇段只考虑一个单音词。

第二、偶段只取上层。

第三、多字词优先,但对复杂夹层要特殊处理,必要时作出多个假设。

对上述约定,现在看来仍然不必变动,既不需要缩小,也不需要扩大。当前软件的问题似乎在于对第二和第三条约定都没有严格遵守。第二条扩大了,第三条又缩小了。

一切疑点处理的时机都应放在处理标记出现,即音串形成之后,具体的处理策略取决于疑点的类型。过去对疑点的讨论,多数是蜻蜓点水,其结果正如毛泽东先生的名言所说,“抓而不紧,等于不抓”。现在必须扭转这一“等于不抓”的状况。

26.2 层选处理的两个加强项目

第一、偶段处理的加强

偶段处理的上述约定似乎过分野蛮,但它的合理性已在论题 24 中论述,这里不来重复。该文指出,只选上层可能造成的严重后果之一是把偶段下层的 E 块遗漏了,这必须加以补救。

寻找偶段下层的动词是偶段处理的加强内容之一,处理的时机应该放在 K 调度之前。这就是说,当处理标记到来而尚未发现动词.或已有的 E 假设都被否定时,就应该回头检查一下 4 音节以上的偶段下层是否有双音动词。对偶段下层双音词的确认必然涉及两个单音词的安排,因此局部处理何时进行就是一个问题。这里只提出来而暂不讨论。

偶段加强的第二项内容是寻找偶段中的单音语言逻辑概念,特别是 l0 和 l1,此项处理属于典型的见机行事,必须通过智能调度来实现。如果 E 块位置显示出该音串优先标准格

式,此事不必考虑。如果 E 块位置显示出该音串优先规范格式,而未曾发现 l0,就必须从偶段中重新寻找。E 块位置如何显现上述信息?这是一个句类格式知识的运用问题,需要专文阐述,由有关读者来完成更有利于工作的推进。这里应该说明的是,此项处理属于上述准备操作的第三项。

第二、夹层处理的加强

夹层处理即多字词处理,讨论次数甚多,但并未形成成熟的处理方案。我希望这一次讨论划上一个圆满的句号。

多字词优先,但保存夹层的全部信息,以备回溯。这就是夹层处理方案的要点。对于四字以上的词和三字词的人名地名,回溯是小概率事件,但一般三字词不能当作小概率事件。1.0 版必须考虑回溯处理。

夹层和虚切口的术语(见论题 1-1 的术语介绍)都是为了这一回溯处理的需要而引入的。回溯的需要与虚切口是否变成实切口(这时多字词两侧的音节都不能向两旁形成组合词)无关,因为多字词内部可能还存在双音词,即使不存在,也可能多字词本身就是伪词,例如‘夜总会’实际上应该是‘也总会’。

虚切口可能有左右皆虚、左虚右实、左实右虚、左右皆实 4 种情况,在每一种虚切情况,又有是否存在多字词重叠之分,这当然是小概率事件,1.0 版可不考虑。但这一类的复杂性主要是数据存放问题,是否考虑由软件设计者决定。但上述虚切口的 4 种情况则必须记录在案,以备类型假设时使用。

26.3 动词连见处理

让我们先想象一下音串刚形成时的景象:简单地说,它是由一些激活点构成的随机序列。激活点有 7 种类型:l0,l1,l9,v,hv,qv 和 QE。后 3 种不能独立存在,必须与邻近的 v 合并。这一合并处理随着分段处理过程进行,在音串形成时已基本完成,这时的激活点序列只包含 4 种激活点,l0,l1,l9,v,将简称 lv 序列。

由于在语义块感知过程不考虑单音节动词,因此,可能‘残存’一些 hv、qv 和 QE 激活点。这些信息必须保留,以备 K 调度时使用。所谓‘残存’,包括不曾出现 v 的极端情况,所以打了引号。

音串形成以后,调度的第一件事是进行论题 2-1 中说明的 E 排除处理。但这里作一项重大修正,取消准则 4(后见排除准则),将它改为动词连见处理。

所谓动词连见处理,仅有处理之名,并无处理之实。什么是处理之名?就是把所有连在一起的动词当做一个激活点。但是,应强调指出,这里说的连在一起是广义的,除一个音段内的同层连见外,还包括下列情况:奇段的上下层动词的交叉连见、两音段的段接连见和动词之间的有插入成分(u,uv,uu,vu)的连见。这样,一方面可以减少随机激活点序列中的动词数量,从而方便 lv 准则的利用;另一方面,可以避免对连见中的疑点过早作出不适当的

处理。

按照我的写作风格,现在可以转入假设类型的讨论了。但此处不行,因为我对上面的文字也很不满意,句子里的专业性隐知识太多。所以,宁可罗嗦一点,补充下面的话。

语义块感知阶段的中心目标是寻求 E、JK 和 fK 的激活点,具体操作过程是分段层选。但此阶段不可能对层选进行彻底处理,只能以基本满足上述中心目标为原则。同时根据音段的特性,偶段根本不作层选处理,奇段也只是标出奇号位置上的 l0、l1、l9、hv、qv 和 QE 信息,并不真正作出层选处理。重要的是,对夹层一定要进行信息的特殊记录,以备后用。多字词本身永远是独立的音段,其两侧一定形成虚切口,这样,音段的范围和数量是确定的,这对处理过程的思考(它关系到数据格式的设计)非常重要,特别是音串形成以后。

当前层选处理的根本问题在于没有遵守层选的前述 3 条原则,过头和不及的错误都有,必须改正。它们的共同严重后果是抛弃了必须保留的信息,古语‘过犹不及’的名言在此得到了验证。

对激活点的随机序列要进行合并和排除处理,它们都是所谓局部处理,可以在感知过程中进行,不必等到音串形成以后才算总帐。合并处理的要点已如上述,不再补充。但排除处理则需要作进一步的说明。

排除处理的 4 条准则要区别对待。“的”准则和‘是’准则是绝对的,因为‘的’和‘是’是指定字。l9 准则和 h□g 准则是模糊的,不能在音串形成之前使用;音串形成以后,也要随机应变。

音串形成以后,如同下围棋一样,要作一次形势分析,以决定下面的处理步骤。形势主要决定于动词的数量(所以应设置一个 v 计数器)。若动词数大于 3,就必须先作上述随机性排除处理。否则可跨过动词连见,直接进入假设类型处理。

动词连见也属于局部处理,可在分段层选过程中同时进行。当然,对上面提出的 4 种类型要区别对待。这里需要注意的是,当两音段之间插入单音段时,要不要作动词连见处理?这个问题暂不作定论,而留待讨论。

语义块感知应区分音串形成前后两个阶段。形成前的关注点是不遗漏动词;形成后的关注点是当动词数大于 3 时,尽可能减少动词,排除和连见处理都是服务于这一目标。排除还有潜力可挖,这就是‘的’前面的动宾结构。这个问题同样也暂不作定论,而留待讨论。

连见处理不仅是为了减少动词数量,也是为了保留信息。奇段上下层连见的处理办法主要是一个概念的升华,读者对此应有所体会。

26.4 假设类型处理

假设类型处理是准备操作中最关键的一步。而这一步本身的关键在于物理,而不在于技术。也就是说,关键在于把问题想透,而想透的关键在于对句类检验本质的领会。

不能把句类假设理解为句类代码的选定。不是的!一个假设中可以包括多个代码。同

理,也不能把一个句类假设理解为仅针对一个动词。不是!它可以针对多个动词。这些是物理问题或概念的升华问题。首先要把这一点想透。

为什么可以这么做?

因为下一步的句类检验是链式关联处理,是现场信息与预期信息的匹配,无论是现场信息还是预期信息都允许多个。不仅如此,由于绝大多数情况现场信息的位置(JK 的要素)与产生预期信息(动词)的位置分隔两地,一个句类假设所针对的动词可以不要求位于同一位置,这就是上述奇段的上下层动词交叉连见可以合并的依据。当然,如果 JK 要素与上下层交叉动词紧挨着,则不能作连见合并处理,这是需要提醒的。

由于作为假设依据的动词可以是上述种种复杂情况,这就提出了所谓假设类型的问题。假设类型就是把各种情况加以分类并制定相应的代码,如此而已。

当前软件测试中所出现的绝大多数问题,都可以归结为没有建立假设类型的概念。所谓疑点及其回溯问题,所谓两头之弱,问题都出在物理方面(另见下文的句类检验)。这个问题解决好了,就会一通百通。这里我故意把话说得绝一点,以期引起争鸣。

至于假设类型的具体分类,只是一个技术问题了,这里不作讨论。

下面将转入句类检验的阐述。但在此之前,有必要说明两点。

第一,紧跟在类型假设之后的句类格式处理,其中包括 lv 准则的应用,晋耀红的已有程序表现不错,仅在语句的数学表示式与现场表示式的转换之间有一点误会,属于小故障,这里就不讨论了。

第二,所谓夹层处理,即寻找夹层内外动词的处理。何时进行为好?这里建议把它作为 K 调度的预处理之一,另一项就是前面已提到的偶段处理。

26.5 句类检验的两种基本类型

我坚信大脑的理解处理不需要复杂的数学运算,初期的 4 比特量化想法即基于这一点。那么,大脑的基本运算是什么?【1】实际上回答了这个问题,就是预期匹配和同行优先,下面简称预期和同行。

预期和同行的具体运算方式都是比较,不过预期是现场与现场的间接比较,一般类别信息与层次符号并重,也有以类别信息为主的情况;同行是现场与现场的直接比较,主要利用层次符号,类别信息仅起参照作用。

句类检验主要属于预期,少数属于同行。具体来说,对 JK2 和 JK3 的检验绝大部分属于预期,而对 JK1 的检验,部分属于同行。这是概念层面的一项宝贵知识。苗庄组正致力于此。

本文以下所谈将限于预期型检验。

所谓句类检验的两种基本类型又有两种标准。一种是并重与偏重之分;一种是要素与总体之分。

所谓并重,就是类别信息与层次符号并重;所谓偏重,就是以类别信息为主。

所谓要素标准,就是只检验要素;所谓总体标准,就是不但要检验要素,还要检验要素的说明部分,或要素的各项构成。

搞清楚上述区别至关重要,因为检验的计算必须以这些区别为基础。

一般的假设检验有一套成熟的数学方法。句类检验也是一种假设检验,但必须采取不同于一般假设检验的方法,下面就来说明这一点。为了联合攻关组成员都能看懂此文,这里不能不先作一点科普性介绍。

假设检验的问题是:面对一个考察对象,它可能有 N 种模式,需要通过检验的方法确定它属于 N 个模式中的哪一个。什么是模式?从下面的对比,你就不难明白。

考察对象	N 个模式	确定具体对象属于哪个模式
音串	N 个动词	确定哪一个动词充当述语
	(N 种句类)	(确定该音串属于哪个句类)

这里的考察对象是音串(或文字串)。假设音串中有 N 个动词,由于每个动词都有可能充当述语(中心动词),因而有可能形成 N 种不同的句子,这每一种可能性就是一个模式(pattern)。一般假设检验的任务是确定具体的考察对象属于哪一个模式。对一个含有 N 个动词的音串或文字串来说,用传统语言学的说法,是从 N 个动词中确定哪一个动词充当述语;用 HNC 的语言来说,是从 N 种句类中选定一种句类,我们把这个选定过程叫做句类检验。

那么,假设检验如何进行具体运算?

它对每一模式进行匹配运算,结果是一个得分。N 个模式就有 N 个得分。从得分作出判断有两种基本方法,一是最大似然法,另一种是阈值法。最大似然法认定,得分最高者是待考查对象所属的模式,通俗地说,就是冠军入选,这容易发生滥竽充数的错误。阈值法设定一个得分标准,叫做阈值,得分超过阈值就入选,这容易发生以假乱真的错误。两种方法各有不足,通常的做法是把两者结合起来,即最大似然 + 阈值,这样可以较好地防止上述两类错误。

从上面关于假设检验的简单介绍,读者应能理解,得分的具体计算方法和阈值的选定是假设检验的关键。这是完全正确的。

那么,句类检验如何选定得分的计算方法和阈值?

HNC 提出了两种方法,就是上面所说的预期法和同行法。

预期法将计算方法和阈值的选定合而为一,简化成下面的七字诀:

合则留,不合则去。

这是苏东坡先生《范增论》里的一句话,这里把它引用来作为预期法的基本计算公式。

预期提供一种标准,符合这个标准的留下,不符合的就去掉。就这么简单。

预期标准从何而来?句类代码是第一个来源,现场数据提供的格式信息是第二个来源,要素的句类知识是第三个来源,由动词提供的 JK 要素的概念优先知识是第四个来源,语义

块构成的具体信息是第五个来源。

这五项预期标准的运用次序不容颠倒。

这五项标准的前三项是绝对预期标准,后两项是相对标准。

相对的意义是:既可作预期标准,也可作同行标准。

在 HNC 知识库的@S 或@K 栏目中,概念优先序列最后加‘;’者表示不能用于预期标准,不加‘;’者表示可用于预期标准。至于用汉字表示的优先,通常都用于同行标准。

同行法只提供得分的计算方法,不涉及阈值。同行法主要用于解模糊,而解模糊可使用最大似然法,用不着阈值。

预期法实质上是一种排除法。句类检验实际上是一个逐级排除的过程。

还需要举什么例子来说明 HNC 预期法的实际应用吗?

我认为最好是读者自己来做。

如果有的读者由此体会到,“什么阈值,什么得分计算,对句类检验来说那都是书呆子观念”。那就表明,你达到‘蓦然回首’的境界了,你就能体会下面的论述了。

句类检验不应该涉及复杂的数学计算,如果涉及复杂的运算,就不是对大脑感知过程的适当模拟。

这里的全部奥妙就在于概念联想脉络的激活。所谓‘激活’,就是预期的东西与现场的东西相符合,现场的东西把预期的东西激活了。被激活的东西留下,未激活的东西抑制掉,这就是七字诀。

这里的‘东西’就是信息,它有两种基本形式,一是概念类别符号,二是 HNC 符号。

最近组建的知识库‘两特种兵团’,就是为了加强这两种形式的预期信息。

本文到此暂告结束,欢迎争鸣。留下的一节,特种兵团会有更详尽的论述。

1998 年 6 月 7 日

论块内处理

31.0 引　言

块内处理是语义块构成处理的简称。这里的语义块只涉及 JK,然而仍是一个大题目,一篇论述难以容纳,所以,本文仅讨论块内处理的准备操作和基本联想脉络,处理细节和有关专题则分散到其他的论题里,这包括:

1. 块内处理的依据,见论题 34。
2. 音段的内外处理,见论题 22、23 和 24。
3. 块内处理的模糊消解,见论题 35、36 和 37。
4. 两种特殊语义块——块扩和句蜕处理的专题论述,见论题 26 和 27。

31.1 再谈准备操作

在具体说明块内处理的准备操作之前,应澄清两点容易发生的误解。

第一,句类分析的三部曲是一种理论上的概括,实际的句类检验往往与块内处理同步、交叉进行。先做完全部句类检验、然后才做块内处理的理想情况当然也经常遇到,它要求全部检验都可以采用要素标准(见论题 26)。但必须清醒地认识到,当句类检验需要采用总体标准时,实际上也就要求同时进行块内处理。不同句类检验的不同项目往往需要不同的标准,这就必然造成上述的同步交叉现象。

第二,由基本概念构成的语义块,特别是数量、时间和空间语义块需要先行处理,不需要也不应该等到句类检验之后。先行处理可带来处理空间的净化,理解处理过程也可以说是一个对音串的逐步净化过程。先期净化越多,对后面的处理越有利。九个基本数字(不包括零)的指定,就是为了能够先行处理数量语义块和某些时间语义块(如 5 点半,6 月 9 日,98 年中秋节)。这里说的‘先行’就是信号处理里的‘实时”,此项先行处理主要是数字指定字特殊处理系统的任务。

上述两点表明,块内处理并非总是三部曲的第三部,它既可以操作于语义块感知阶段,也可以与句类检验同步进行。调度程序必须充分适应这一要求,这就是智能性表现。

现在转入块内处理准备操作的讨论。实际上,句类检验的准备操作也就是块内处理的准备操作。对这些操作在论题 26 中开了一个 10 项清单,其中的 7　9 项直接与块内处理有关。对这个清单需要作总体和分项说明。后者散见于有关论题,这里仅作总体说明。

准备操作的主题是形势分析,论题 26 说:"音串形成以后,如同下围棋一样,要作一次形势分析,……"。所谓形势,就是指 lv 序列,并主要决定于动词数量、分布和模糊状态。这里的数量和分布都应以"团块"(动词连见形成团块)计,包括动词的上下层连见。如下面的例句:

我…从来…没…想到 * 会…翻译 * …1…部…德国军事 * 著作。

我…打算 * …把握…在…第二次|世界大战…的…亲身…经理 * …写出来。

这两个例句反映了两种复杂形势的典型。

第一个例句无 l0,只有两处动词,第一处是一个团块,第二处是无模糊低档动词"著作"(根据萧老师最近的分类,我以后将把动词分为四档,高档、中档、低档和档外,"+"为低档,"++"为档外,零级词为高档)。这个低档动词位置在最后,又没有 l0,而且很可能知识库中标明了这个动词要求标准格式,因此,"著作"可以排除在述语候选之外。剩下的团块则比较复杂,关于动词团块的处理将在论题 3-1 中阐述。

第二个例句有四个动词:"打算"、"把握"、"经理 *"和"写出来",它们构成两个团块。第一个团块的第一个动词"打算"(其伴随词"大蒜"可先不考虑)要求块扩,后面容许紧跟动词。但是,如果以动词"把握"作为块扩的 E 进行检验,将立即被否定,因为,在其要素位置上是一个无模糊的纯动词"写出来"(亮点)。这样,调度程序应作出进入 K 调度的准备操作的决定,而且目标就是从偶段中重找 l02,于是,"把握"就成了第一怀疑对象,而问题也就迎刃而解。这就是基于句类假设检验的智能调度过程,难道说还有什么神秘么?

第二动词团块中的"经理"需要排除,这利用了"的"的排除作用,参看论题 2-1。

为了取得比较简明的形势信息,HNC 采取了两项重大措施。

第一,在语义块感知阶段,我们对 lv 序列的两形势要素 l 和 v 分别采取了"宁缺毋滥"和"宁滥毋缺"的不同方针。前者是通过特殊处理,后者是通过知识库的概念类别栏目及其执行过程来贯彻的。对这两条方针的贯彻有松紧之分,目前采取了前紧后松的策略,即容许 l 缺少而保证 v 不滥取。这样做,与初期的理解处理能力比较适应,随着处理能力的增强,将来可考虑逐步向松紧适度的策略转变。宁缺毋滥方针的具体贯彻就是当前的指定音特殊处理系统只处理一二级 l 激活音,对于口语中常用的由实词转化而来的激活音基本上不予考虑。我认为,1.0 版本的这种策略是明智的。就 l0 来说,它用于规范格式和部分违例格式,这时 E 块一定后移,由句类信息的提示,不难在 l0 信息暂时丢失(例如隐含在偶段中)的情况下把它再找回来。

第二,充分利用音串中指定字"了、是、的、在、不"所提示的形势信息。关于这些形势信息的论述不属于本论题系列的范畴。这里只再次强调指出,上列指定字的设置主要是为了取得简明的形势信息。

总之,形势分析是实现智能调度的生命,是具体安排准备操作的依据。

31.2 块内处理的基本联想脉络

本节将首先给出块内处理基本联想脉络的清单。这个清单是制定块内处理调度程序的依据。

块内处理基本联想脉络清单

联想脉络类型	信息来源(理论)	信息来源(现场)
1 带内容 C	语义块表示式	
2 带 C 不块扩	句类知识库	
3 不带 C“块扩”	@S	现场
4 良性	语义块表示式	
	@S 构成表示式	
5 强约束	句类知识库	
	@S,@K	
6 必带逻辑标记	句类知识库	
7 对仗性	句类知识库	
	语义块表示式	
8 带 HE	@K 构成表示式	现场
	句类知识库	
9 分离变换	@S 构成表示式	现场
10 K(jm)修饰		现场
11 多重修饰		现场
12 多重内容		现场

清单所列是块内处理时必须考虑的 12 个联想脉络。这里要强调的是其激活信息的生成不同于句类假设。句类假设的激活信息来源是单一的,主要依靠对动词或动词团块的分析,仅 S04 句类需要通过 K 调度。块内处理则不同,一些联想脉络的激活完全或主要依靠对现场信息的综合分析。清单将信息来源分为理论和现场两列,就是为了表明块内处理的这一特点。

清单的 1 到 4 项是必须激活的首要联想。它们表明了语义块构成最本质的特征:是否含内容 C。

带 C 语义块具有块扩和句蜕的天然特性。这一信息由语句物理表示式给出。但不应忘记语义块是句类的函数,少数句类的带 C 语义块不具有块扩特性,这必须在句类的概念层面知识库中予以标注。语义块构成的这一最重要的共性和个性知识分别由 1、2 项给出。在它们的知识来源说明中只标示了概念层面信息,未涉及词汇层面。那么,在词汇层面是否还要另行给出有关信息?目前实际上采取了双管齐下的策略。

带 C 语义块意味着构成的非良性表现,因此,特定词汇的良性表现应在 @S 中加以标

注。但应该指出,BC 型语义块的纯 B 或纯 C 表现只是一种简化,被简化的内容或对象变成了隐知识,必要时需要加以恢复,这当然不是当前的任务。

不带 C 的单一对象语义块不具有块扩和句蜕的天然特性,但有例外。少数例外可在词汇层面予以表示,清单的第三项——不带 C“块扩”即表示这一例外,块扩加了引号,因为它包括句蜕。这种情况通常出现于 JK1,知识库中很少加以标注,主要靠见机行事。

单一对象语义块必须遵守良性构成规则。若有例外,应在 @S 中有所标示。此项未标现场信息,表示不要求见机行事,若有违反,应作为纠错的依据。1.0 版本则以请求交互的方式处理。

清单的 5 到 7 项不是必须激活的联想脉络。但其激活信息全在概念或词汇层面的知识库中,一旦出现,就必须抓住不放。这三项联想脉络涉及某些句类语义块的特殊属性,对句类检验和解模糊特别有效,软件绝不可失之交臂。在 @S 或 @K 中,强约束的标记是“。”,放在概念优先性序列的最后。

清单的后五项都属于语义块构成知识。其中的 8、9 两项的激活信息可从概念或词汇层面取得,但最后三项则完全依靠现场信息,只能见机行事。

后　记

清单的最后两项是语义块构成处理的难点,对它们的讨论需要立足于语料。请联合攻关组成员协同收集,并及时与我沟通。在本论题的续篇 31-1 中对这些及其他难点将有所阐述。

1998 年 6 月 15 日

论块内组合结构

直觉是纯粹的专注的思维的可靠概念,它仅由理性之光产生,而且比演绎更可信一些。

笛卡尔《思维指导法则》

31-1.0 引言

本论题的主篇讨论了 JK 块内处理的总体方面,本文着重块内的组合结构。从理论上说,块内的组合结构无非是 HNC 所定义的 8 种基本结构的再现或复合。而难点在于复合,更准确地说,在于复合的顺序。

仅从音词转换来说,这个问题有时可以回避,但从音义转换来说,这个问题是不能回避的。

8 种基本组合结构的一般复合规律是一个理论上很有趣的课题,4 年前就曾想结合汉语多字词的组合结构进行研究(见〈问答 34〉),迄未如愿。其中的主要原因是始终抽不出时间系统了解语言学家在这方面已有的工作。本文仍然要回避这个问题,仅从工程角度对一些常见的困扰进行讨论。

31-1.1 对组合结构基元的简单回顾

词语的组合结构是传统语言学最感到自信的方面,以联合、偏正、动宾、主谓加上后补为组合结构的基元,似乎是放之四海而皆准。我在发现作用效应链以后,对此产生了怀疑,于是修正了原来的形式标准(其中,国人命名的后补最为"形式"),而代之以语义或概念标准,增加了作用、效应和语言逻辑三种结构。并进一步指出,除了联合结构外,每一种组合结构都存在正反两种形态。这个问题值得写一篇严谨翔实的论文,这就需要按传统习惯做很多琐碎的工作,而我的性格注定了永远提不起这份精神,自然是永付阙如。

HNC 的 8 种结构中的联合、偏正和主谓沿用了原来的定义,但对联合和偏正缩小了它的使用范围,惟有主谓原封不动,但心里始终存有怀疑。最近就萧友芙老师整理"概念类别"之机,对主谓结构采取了对动宾结构同样的一分为二方式,增加了 Cv 类型。但仔细想来,主谓结构应一分为三,再增加一种 BC 类型,这才符合 C 的两可特性。

但主谓结构的三种类型可视为它的三个子类,不必像动宾结构那样必须分成两个独立

的类别,因为前者对语句总体结构的影响不像后者那么显著。

8 种组合结构基元都有自己的子类,其中偏正和语言逻辑的子类比较复杂,尚未进行具体研究。这个问题在理论上是比较重要的,但工程应用上并不迫切,所以一直拖延着。知识库的组合代码栏目目前只约定了一组数字,实际上应该有 3 组,分别表示 8 类组合和它们的正反形态及子类。

上述组合结构是指概念的组合,适用于对词、短语、语义块、句子和句间各层面组合结构的表述。这个现象汉语语言学家早就注意到了,是中国学者的一项贡献,但思考的深度远远不够。例如,作用效应结构在汉语的词汇和句子层面都有突出表现,词汇如‘提高,击溃,改善,水压……’,句子如兼语句和‘使’字句。在西语里,由于词的扩展仅采用简单的前后缀方式并注重规则性形态变化,没有汉语‘字义基元化,词义组合化’的便利条件。另一方面,句子的扩展有发达的关系代词可以利用,因而作用效应结构在词汇和句子层面都比较隐蔽。西方人没有发现这种结构是可以理解的,而中国人也迟迟没有发现,则不能不说是《马氏文通》的消极影响。对语言逻辑结构,也可作同样解释。汉语采用这种结构的词汇如电焊和点滴等,在西语往往用短语或另造一个新词来表达。句子层面的语言逻辑结构表现,中西语言倒没有什么差别,都以辅语义块的形式出现,不过我们叫做辅块,传统语言学叫做状语或状语从句罢了。

不同语种组合结构的个性很值得进行系统的研究,特别是在语义块层面,对于机器翻译将大有裨益。我曾把它叫做语义块构成的变换,是实现 HNC 翻译方案的两大关键之一(另一关键是句类转换),应该争取尽快开展这项研究。我热切期望着联合攻关组成员中有人响应这一历史性(在工程意义上绝不夸张)的召唤。

31-1.2 块内组合结构的分类和处理策略

语言的 8 种组合基元虽然适用于语言的各个层面,但各层面又各有自己的特色。例如,汉语有不少动宾结构的词汇,这是汉语的特色之一,但词汇很少采用反动宾结构。语义块层面则不同,它经常采用。如‘农民最低生活保障制度’这个语义块里就存在反动宾结构‘农民最低生活保障’,它构成 KQ,然后与 KH‘制度’以偏正结构形成 K。

从工程角度研究块内组合结构,传统语言学的 4 类型说,即联合、偏正、动宾和主谓(这里去掉了国人所加的后补)的分类,具有净化思考方式的优点。下面将采用这些术语进行讨论。

与词汇相比,块内组合结构的最大特点在于它经常综合使用多种结构。软件设计首先要适应这一特点。下面给出块内组合结构的工程分类。块内处理应根据这一分类和主篇中所说的 12 个联想脉络形成调度的思路。

1. 无组合,指语义块仅由一个词构成。这个情况似乎很容易处理,其实不然。麻烦出在单音词。1.0 版本必须限制一下具体音节。

2. 联合,包括多级联合。联合标志有:指定字“和”,音节“ji,yu,tong,gen,…”。多级联合标志为“、”,但最后要用指定字“和”或音节“ji,…;deng”加以补充。麻烦在于,第一,“和”还有 102 及其他义项;“ji,yu…”也有其他义项。这都需要检验。第二,联合有对仗与非对仗之分,对仗联合是强有力的解模糊及纠错武器,但前提之真需要检验。检验标志之一是后面跟 QE 音节 dou。第三,单音词的联合结构往往不加标志,最后以数词+基本命名结束。

以上所述,需要形成规则。

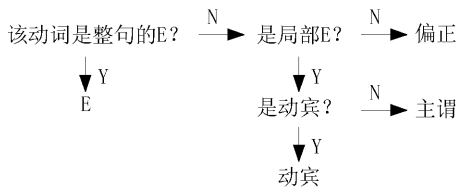
3. 简单偏正,指修饰部分不带动词的单级偏正。

指定字“的”是汉语偏正结构的醒目标志,但是,“的”也是汉语实现句蜕和偏正语义块中心语省略的语法手段。

大量的偏正结构不存在“的”标志,这时需要通过概念类别或同行来验证。此项验证对双音词或双字词比较简单,麻烦的是单音词的参与。

4. 复合偏正,指修饰部分带动词的偏正。“农民最低生活保障制度”就属于这一类。

文献中大量出现的所谓令人畏惧的偏正动宾歧义大多数属于这一情况。对这一歧义现象的 HNC 处理策略是:



HNC 凭借完备的语句物理表示式及其背后的句类知识才能实行这一策略。展示这一处理结果应列为下半年的工作重点之一。苗庄组应先行准备好例句。

在这一类里,实际上包括了简单的动宾和主谓结构。

5. 多重偏正,包括修饰部分的再修饰和中心语再修饰两种类型,即:

$$K = KQQ + KQH + KH$$

$$K = KQ + KHQ + KHH$$

这里的任一块素可含动词。因而,后两种结构多数情况是句蜕块,属论题 28 的范畴。

我认为,从调度的角度说,块内处理就是对上述五类结构之一的确定过程。应基于这一点并结合主篇所说的 12 种联想脉络来制定块内处理框图。

对上列五种类型,都应该制定相应的规则,并形成规则库。我建议有关人员集中讨论一次,并落实分工。这次应吸取教训,从最简单情况做起。也就是说,先搞好前三类。

在制定此类规则时,收集语料固然重要,但更重要的是思考,所以引录了笛卡尔的一段话放在篇首。

1998 年 5 月 16 日

概念类别符号及其运用

33.0 引　　言

概念类别符号或简称类别符号与层次符号是 HNC 符号体系最早提出的两个概念(见【1】),类别符号也称字母串,层次符号也称数字串。这里应该指出的是,【1】中所定义的概念类别与 HNC 知识表示菜单中的概念类别并不等价。后者是对前者的工程化和扩展,下文有详细说明。

概念类别同概念本身一样,有基元与复合之分,那么,概念类别到底有多少基元?本文将首先对这个问题作系统阐述。

33.1 概念基元与概念类别基元

在 HNC 的文献里,概念基元与语义基元经常等价使用,在【1】中有这么一段话:"这三个超级语义网络,将用来对自然语言概念体系进行总体描述,也就是说,我们试图通过它们对语义场的说法给出具体的表述。……产生这一想法的来源有两个,……第一个提出了语义基元的杰出思想并暗含着总体表述的宏伟目标"。这是一段有大量潜台词的陈述,这里我想敞开窗户说一下。第一,它表示对义素和语义场之说评价不高,打个比方,它们很类似于我国乡镇企业的转口加工,附加值很低。对义素和语义场作了比较深刻思考的实际上只有山克和菲尔墨两位先生,思想是杰出的,但宏伟目标是暗含的,而且两位都半途而废。第二,国内外曾有二十几位专家从事过语言分类体系的研究,但学界公认没有一个体系令人满意(陈群秀、张普,1995)。然而,这项探索随着 HNC 概念体系的建立已划下了圆满的句号。

为什么要说上面一段话?因为,如果没有一个完整的自然语言概念体系,概念类别基元就如同义素说和语义场说的实际表现那样,基本上是空中楼阁。应该强调指出:HNC 概念符号体系同概念类别基元体系是同步建立起来的,后者虽然只是前者的一部分,然而这个局部对整体和全局的形成起了决定性的作用。

概念类别有基元与复合之分,这只是划分概念类别的角度之一,属于语义分类的角度。此外,还有形态与多元性表现以及工程需要的角度。张普先生(张普,1992)曾指出,概念属性分类"必须把多层次,多类型,多关系,多变化这些性质综合加以考虑"。这个意见是正确的,这里的三类区分角度正是对张先生这一意见的具体化。

"概念类别有基元与复合之分"这句话里的"基元"与"复合"相当于数学里的公理与推

理。这里的复合是复合基元,每一个复合基元都有确定的定义。

HNC 约定:概念类别的基元用单一字母表示,复合基元用多个字母连写或字母串表示。但这项简明的约定曾受到语义结构方程思路的严重干扰。

HNC 把自然语言概念先分为抽象与具体两大类,对抽象概念设置了四种类别基元:ϕ, j, l, f,分别命名为基元概念,基本概念,语言逻辑概念和"语法"概念;对具体概念设置了两种类别基元:w, p,命名为物和人。这里一共是 6 个概念类别基元,它们是否满足概念类别基元的完备性要求?这是一个很值得深思的问题,HNC 的回答是:否!这个答案是在 ϕ, j, l 三个超级语义网络的设计基本完成以后才明朗的,在 ϕ, j, l 之间,在抽象与具体之间,还需要分别引入一项综合性概念类别基元,这就是 s 和 x,分别命名为综合概念和物性。到这个时候,我们觉得("觉得"一词比较贴切)概念类别基元的完备性有了保障。这样,总共是 8 个概念类别基元,"8"是一个吉祥的数字,这使我感到愉快,与此相对照的是,当发现基本句类总数是"7"的时候,使我忐忑不安达数年之久。我自认为不迷信,但心理现象就是如此奇特,所以人工智能还是暂时避开心理现象为妥,HNC 坚持这一宗旨。

精选概念类别基元是建立自然语言概念符号体系的需要,同时也是为构造复合概念类别基元奠定基础的需要。在 HNC 理论体系中,概念类别基元和概念类别的复合基元都是概念联想的激活因子,因此必须具有确定的意义。在这个问题上,由于上述语义结构方程的干扰,曾发生过临时抱佛脚的失误。

复合概念类别基元的设计不是一项轻松的任务,但经过这一年来联合攻关组的大规模实践活动,现在已具备进行完备设计的充分条件,我深信此项任务能在短期内胜利完成。

下面来说明概念类别的形态及多元性表现,这个关键性问题在传统语言学里完全被忽视了。在【1】中对此有详尽的阐述,这里只补充四点:

第一,概念类别的形态表现仅限于抽象概念,HNC 把抽象概念的形态表现概括为五元组(v, g, u, z, r)。五元组是抽象概念的形态类别基元,由它们可构成各种形态复合基元,构成方式就是五元组字母的连用。这一形态复合基元的定义同样也曾受到语义结构方程的严重干扰,需要进行彻底清理。

第二,形态基元和形态复合基元都可用来表达抽象概念的多元性表现。但表达方式绝不能与类别复合基元的定义方式相混淆,即绝不能采用小写字母的连写方式,而应该采用 HNC 所引入的两种组合结构符号":"和"+"。

第三,概念的多元性表现不限于五元组,也可以是其他概念类别基元。

第四,概念的多元性表现一律放在知识表示菜单的"概念类别"栏目里。概念本身的 HNC 符号表示式中只使用类别基元或类别复合基元。

本节最后对概念类别的工程需要作一简要说明。这里的工程需要是指软件的需要,具体说,是指理解处理过程中某些特定环节的需要,例如,E 假设,E 排除,E 排队,JK 或 fK 感知,局部处理等等。考虑这些特定环节的需要,在概念类别栏目中给出简明的表示,是一项意义重大的举措。

为了满足这一需要，必须引入“字母＋数字”的类别表示方式，这种表示方式有的只是对词语 HNC 符号的简化，如 l0，l1，jw6，j3，有的则赋予了新的意义，如 w9，wj2，pj01。

短语表示方式，如 vB，vC 等，即小写大写字母混合使用的方式，也属于工程需要。

33.2 关于语言逻辑概念的复合表示

语义块感知是 HNC 理解处理技术的突破口，在这一处理环节，语言逻辑概念的作用最大，可是，目前恰恰是对这一类概念的理解十分薄弱，这促使我写这一节短文。

让我再次从类别复合基元谈起。

概念类别符号基元 φ、j、l 可形成 6 种复合基元，但 HNC 只选用了其中的 4 种：lφ、lj、jl、φj，另两种组合 jφ 和 φl 未予选用。

在选用的 4 种复合类别里，符号 φ 约定略而不写，但这不等于对 l 和 lφ 不加区别。l 表示语言逻辑自身的概念，lφ 则表示以基元概念为依托的语言逻辑概念，lj 表示以基本概念为依托的语言逻辑概念。例如，lg02 表示对象，lφg0200 则表示作用对象，lφg0220 表示转移对象。由于 l 本身的层次符号约定为最多两层，因此，l 类符号的层次数就可以指明它是类别基元还是复合，上述约定省略即基于此。

这一类的简化约定在 HNC 符号体系中屡见不鲜，我希望软件能适应这一类的简化，而不是反其道而行之。

语言逻辑概念之首 l0 定义为主语义块指示符，接着就是辅语义块指示符。读者由此可以感受到 HNC 概念符号体系设计的总体意识。这就是说，HNC 六层理论模式是统一设计的，前两层的结合更为紧密，但自 1995 年以来，陷于尘世之喧嚣，思虑之锐气日减，内心之惶恐已难以用言语来形容了，这是插话。

与(l0，l1)相对应的(l2，l3)是为汉语专门设计的，我不知道是否还有其他语言存在类似的巧妙语义块标记方式。但是，不管是否存在，都值得为汉语的这一“创举”设置专门的 l 概念节点，这也许有点民族感情的狭隘性吧！

l2 一定成对出现，分别指示两个语义块，其中至少有一个是主语义块。另一个可以仍然是主语义块，也可以是辅语义块。典型的 l2 搭配有：

给……以	lq2200	lh2000
以……为	lq21	lhv2040
	lq22	lhv2002
	lq11	lh20

(给，以)搭配是汉语特有的 X01 句类转换的变形，因此，它一定构成作用句。如“我们要给敌人以毁灭性打击”。(以，为)搭配有 3 个义项，义项 1 常用于句蜕的形式，如“以江泽民同志为首的党中央”；义项 2 用于构成 !11 格式的语句，如“我们要以大局为重”；义项 3 常见于古汉语今用，如“我们要以攻为守”。

同 l0 一样，l2 的作用当然也是使语句格式规范化。对 3 主块句，就是 E 块移至最后，因此，上面的 HNC 映射符号应在意料之中。熟悉句类格式的读者可能会问，对于 4 主块句，是否存在指示两个 JK 的 l2？人们不难想到，“从……到”似乎是天然的一对，但是，我们不把它列入 l2，而列入 l30，因为它标志的是范围的起止，以构成基本概念短语 K(j42)。而 K(j42) 是一种很特殊的短语，既可充当 JK 或 fK 的块素，也可充当 QE(如从里到外，从头到脚等)。

(以，为)搭配的 HNC 映射符号，采用了 l 类概念少见的五元组表示形式 lhv20，lv 代表语言逻辑动词，具有双重作用，既是语义块指示符，又是动词。lv 在构成 E 块时，必须采用 EQ + EH 的构成方式，它本身充当 EQ，就这里的 lhv20 来说，它优先“动静”搭配(见论题 3)，如“为首、为核心、为榜样、为标准、为左右手……”等等。

传统语言学早就提出联想(association)与搭配(collocation)的概念。联想主要是纵聚合的体现，搭配则专指横组合的约束。这两个概念起源于现代语言学鼻祖索绪尔的联想关系(或聚类关系)和组合关系(或搭配关系)。这两类概念关联性大体上相当于联想脉络的纵向与横向延伸。但是，应该指出，这些思想不过是分析与综合这两个概念的派生品，关键在于明确概念联想脉络主干的内涵，HNC 的理论认为：概念联想脉络的主干是作用效应链，是由此导出的 7 个基本句类和 36 个混合句类，更具体地说，是 57 个基本子类和大约 300 个常用的混合子类。这些句类构成概念联想脉络的主干。句类之间的关联性则构成概念的超级联想脉络，属于 HNC 理论尚待探讨的句群模式和篇章模式。

在每一个句类内部，语义块要素之间的关联性，语义块内部各块素之间的关联性都属于“搭配”，HNC 称之为链式关联性；而语义块各块素自身的概念集合属于“联想”，HNC 称之为交式关联性。链式和交式的提法比搭配和联想的提法，我认为更恰当一些。在 HNC 符号体系里，我们试图通过概念节点对这两类关联性予以详尽的表述，并力求使表述的形式易于为计算机使把握。这就是 HNC 理论前两个理论模式的精髓。

“详尽”的具体表现首先是(ϕ、j、l)三大语义网络的设计，它们涵盖了自然语言的全部概念基元，任何自然语言概念都可以通过这些概念节点及其组合予以表达。

以三大语义网络为基础建立概念的联想脉络是 HNC 理论的新思路，人们对这一新思路充满疑虑是可以理解的。一方面，“HNC 理论概要”一文本身对这一新思路的阐述不够透彻；另一方面，HNC 的软件可实现性，在总体框架上虽然比较明朗，但实现的深度仍有待探索，HNC 符号知识的运用、句类知识的运用都存在深度方面的严峻挑战。本论题及其一系列姊妹篇都希望从不同侧面阐述 HNC 符号的深层含义，以期有助于软件深度的进步。

“以，为”搭配的 HNC 符号表明，它通常构成关系句或反应句，作为 E 块指示符“为”是模糊的，具体的句类取决于其“静”搭配 E。

“以，为”不仅是 l2 的反映射符号，也是 l3 的反映射符号，具体表示式为：

以……为　　lq3　　lh3

　　lq34　　lh34

第一个映射符号本来包含了第二个映射符号，但后者不可省略，原因是对具体辅块类型的辨识，除 Re 以外，都能从“为”后面的词语得到明确的信息，而 Re 往往不能，例如“以 A 点为圆心划一个圆”。因此，应该把这里的(lq34，lh34)当作诸葛式的锦囊来使用。

“以，为”同是 l2 和 l3 的反映射符号，显然为句类分析带来了一定困难，考虑到 yi，wei 两音节的巨大模糊和 wei 音的特殊重要性，这一困难就更加严重了。因此，我对 wei 的特殊处理寄以厚望。

回到本节的主题，上面的大段文字表明：

(1) l 类概念绝大部分是复合概念；

(2) l 类概念中也存在 v 类概念，但 lv 类概念往往只能充当 EQ。

1998 年 5 月 26 日

语义块表示式及其应用

本文是一篇短论,不必搞引言分节那一套,信笔写来,把想说的意思说清楚就是了。

语义块表示式有物理与数学之分,下面先把各种语义块的物理数学表示符号的有关约定归纳成一个表格。

语义块类型	物理表示	数学表示
特征语义块	X,P,T,Y,R,S,D,jD	E
广义对象语义块	A,B,C	JK,JKm,JKn
辅语义块	Ms,In,Wy,Re,Cn,Pr,Rt	fK,fKm,fKmn
两可语义块	RtB,ReC	fJK,fJKm,fJKmn

对这张表格应作两点说明:

第一,表格中的物理表示符号只是语义块类别基元的符号,主语义块的实际物理表示式,除极少数情况外,都是这些类别基元的组合形式。

类别基元的组合形式有:

1. 类别基元与数字的组合。
2. 类别基元的相互组合。
3. 上面两种组合的再组合,用于表示 JK。

类别基元与数字的组合又有两种类型:

1. 特征语义块类别基元加数字,表示句类子类,如:

   X1,X2,T1,T2…

2. 广义对象语义块类别基元加数字,表示广义对象的不同侧面,如:

   TB1,TB2,TB3,PBC1,PBC2 …

类别基元的相互组合又有三种类型:

1. 特征基元的相互组合,表示混合句类。
2. 特征基元与广义对象基元的单向组合,表示 JK 的类别特征。
3. 广义对象基元的相互组合,表示 JK 的构成特征。

若在语句表示式中采用后一种表示,表明该 JK 的构成具有非良性特征。

以上所说,不过是【2】中有关内容的说明书式翻版。

第二,表格中的新面孔。

首先是两可语义块的 RtB 和 ReC,顾名思义,它们的中文名字是:目的对象和参照内容。你看,我又用对象和内容这两个最基本的语义块块素概念把两可语义块一分为二,这是演绎

的结果。将来大规模真实语料的统计也许会表明需要引入新的两可块组合符号,但可以坚信,这两种一定占绝大多数。

RtB 的约定指示符是 l18,ReC 的约定指示符是 l19,两者对应的典型汉字是"为(wei4)"和"从"。

表格中的 JKm、JKn 似乎是新面孔,其实不是。下标 m 是语义块数学表示式所用的符号,下标 n 是 JK 子类所用的符号,实际上仅用于物理表示式。这里关于下标 m,n 的意义约定同样适用于辅块和两可块。

应该说明,辅块 fKm 的下标 m 与辅块指示符 l1m 的 m 一一对应。辅块 fKmn 的下标 n 也与辅块二级子类的指示符相对应。这个指示符的形式是 l1m * n,但 n 的具体安排尚未设计。

fJKm、fJKmn 的意义类推。但目前实际上只使用 fJK0,它表示对象内容两可兼主辅两可的语义块。如以"关于"打头的语义块。

语义块物理表示式的意义在论题 7 中以答问的形式提前作了说明,这里就不重复了。下面谈一下它的应用。

这些表示式的应用需要从不同的角度去阐述,例如工程和理论的角度,现实和未来的角度,局部和整体的角度等。

这里仅从工程、现实和局部的角度作提要式说明。

第一,根据 JK 的物理表示式确定句类检验的类型。对复合基元表示的 JK 必须作总体检验;对单一基元表示的 JK 可先作要素检验。

第二,根据 JK 的物理表示式确定语义块构成分析的类型。对单一基元表示的 JK 调用良性处理函数,对复合基元表示的 JK 调用非良性处理函数。

第三,对含 C 的 JK 要自动考虑块扩、句蜕和分离。

第四,对广义作用句处理违例格式时,假定含 C 的 JK(可简称 C 块)在前,不含 C 的 JK(可简称 B 块)在后。对 C 块进行自足性检验,或逆向对 B 块作自足性检验,也可两者并举,相互印证。

第五,对 4 块广义作用句处理标准格式时,假定 C 块在后,B 块在前。对 B 块作自足性检验,或逆向对 C 块作自足性检验,也可两者并举,相互印证。

上述两项自足性检验即论题 1 所说的 BC 佯谬处理。逆向自足性检验时,以遇到 p400 或 l19 为自足性标志。注意,当遇到 l19 时,要向前再检查一步,看是否存在 p400。

第三条应用准则与知识库的填写如何配合?希望听到回音。

1998 年 6 月 7 日

第 四 部 分

HNC 理解处理问答

HNC 理解处理问答话题目录

问 1：

汉语自然语言处理应从分词做起，这是常识。国内有关高校和研究单位已在这方面做了大量工作，可是，先生对他们的工作成果似乎不感兴趣，为什么？

答：

问题在于工作的基础和目的。现有工作的基础是文字文本，目的是为分词而分词。对后一问题这里不来多说，一说就难免节外生枝。以后有机会再谈。

HNC 将从语音文本入手，这两种文本面临的分词处理有天壤之别，这一点人们似乎并不清楚。所以，这里不妨多说几句。让我们从一个句例开始。

“刘嘉玲×正式—向⌒上海×中级—人民⌒法院—起诉×汕头×雅丽丝—实业—公司”

这句话有 12 个词，在两词之间给了三种符号：×、—、⌒。最后一种符号表示文字文本出现的交叉组合（即‘向上’；‘民法’也是词）。第二种符号表示语音文本增加的交叉组合，符号‘×’则表示不存在交叉组合。我们看到，音调模糊把该句交叉组合从 2 个增加到 7 个。这不仅是一个量的变化，而是出现了组合爆炸的潜在危险。

然而，这种潜在危险还不是质的变化，质的变化在于，文字文本的分词处理也许可以先避开理解，而语音文本的分词处理则绝对离不开理解。

这个认识不会引起太大的争论。但是问题在于：理解处理如何着手？分词是不是必需的第一步？

在回答这个问题之前，让我们把音调模糊带来的后果列出两个清单。

1. 交叉组合出现的词

shi-xiang	事项 试想 识相 实象 食相
xiang-shang	向上 相商
ji-ren	继任 己任 级任
min-fa	民法
yuan-qi	怨气 元气 缘起 远期
si-shi	私事 死尸 四时 巳时
ye-gong	夜工

2. 各词的模糊集

正式	证实 正视 正是 正事 政事 正史 正室
上海	伤害
法院	发源 发愿
起诉	泣诉 耆宿
实业	事业 失业 事业 师爷
向	相 象 想 ……（总计 31 个汉字）
汕头	山头

这就是说,在 9 个双音词中,只有两个词‘人民’和‘公司’是无模糊的。

两个专用三音词‘刘嘉玲’和‘雅丽丝’,这里不来讨论,即使是人的听觉,也不能给出准确的汉字。

应该指出,上列两个清单并不能表示音调模糊的全部后果。实际上,汉语的每个音节都对应着一个意群,这里只考虑了两音节的组合意群,并没有考虑各自的意群。让我们假定,你仅仅面对着上列清单中诸多概念,你应该采取什么对策,对它们进行取舍处理,最后组成一个合理的语句?

我认为,唯一的对策是建立概念的联想反应和要素反应。

在这句话里,两处联想反应是:

zhong-ji ren-min fa-yuan, qi-su

shi-ye gong-si

联想反应必须有要素反应的配合,这里的要素反应是:

qi-su 与 fa-yuan

这里说的要素反应是联想反应的升级或深化,联想反应是理解的初级形式,它对于消除双音词的模糊集有一定的效果。但是,如果我们要进一步判断语句的完备性和合理性,仅有初级形式的联想是不够的。就这个句例来说,从“起诉”和‘法院’,还要联想到‘原告’和‘被告’,这不是指联想到这两个词,而是指联想到人类的一项活动——法律活动,这样,才能把两项初级联想有机地联系起来。

用概念层次网络的术语来说,有了‘法院’—‘起诉’这一项要素反应,就能假定,这将是一个作用句,而且特征要素 X∈10-5(法律),这种语句的另外两个要素 A 和 B,必须属于 p 类概念。这个句例中有三个 p 类概念,刘嘉玲、法院、实业公司。至于它们之中,是哪两个充当 A 和 B 的角色,则需要运用语法知识了。

关于模糊文本理解处理的策略,我的设想,大体上都说到了。细说起来,可以归纳成下列几点:

(1) 利用现代汉语以双音词为主的特点,找出每一对双音节的概念集。

(2) 对每一个概念作前后联想,形成一些语义块的雏形。

(3) 从概念集或雏形语义块中取得语句要素的信息,并由此作出语句类型假定。

(4) 从语句类型出发,充分运用要素关联知识,进行各项模糊的解模糊处理。

上述四步只是一个分析步骤的要点,还有三个重要问题没有涉及,它们是:隐知识的揭示、语境及主题的生成、伴随成分的判定。所谓语句的伴随部分是指该语句中有关条件、手段及动因目的等的说明部分。对于这三方面的知识处理,层次网络理论都提供了明确的线索,但已超出了当前的讨论范围,这里不再多说。

问 2:

先生刚才的回答似乎有意回避了分词问题。从汉语的特点来说,分词的要点是区分单

音词和非单音词。如果能够把一个语句中的单音词先分离出来,先生前面说到的一系列分析处理就有了可靠的基础,为什么先生似乎有点故意不承认分词是理解处理的起点?

答:

你误会了。我刚才说过,语音文本与文字文本有天壤之别,对于后者,把分词作为理解处理的第一步或许是可行的,因为交叉组合率很低。而在语音文本的情况,由于交叉组合率很高,理解处理的第一步与其说是分词,不如说是分段。分段之后,在进行联想反应和要素反应的同时,要插入音节感知、选词等项处理。随后的语句要素分析、解模糊处理等也要伴随着这两项处理。所以,不能把分词处理作为一个前提性的独立步骤,它是贯穿于理解处理全过程的。

问 3:

先生刚才的说法似乎是用分段代替分词,又加入了音节感知的提法,这些虽不是新名词,但似乎有特定的含义,先生能否说得详细一些?

答:

自然语音流中存在天然的段,段信息表现为抑扬顿挫。有个别人说话缺乏这种信息,听起来非常吃力,就像一片混响。语音流的段大体上对应于语义块,人的听觉有音节反应(对于汉语)、词反应和段反应三级,以段反应为核心。我刚才说的分段不能说同天然语音段毫无关联。不过,这里的音段是一种工程定义,指两相邻音节不能构成双音词或多字词的分隔点。如〈问答 1〉例句中"×"号所示的各点。上面说的分段处理就是找到这些分隔点。这是语音文本处理可靠的第一步。

当然,这一步不过是万里长征刚刚迈出了半步,不过,它是非常结实的。下一步,就可以进行奇偶判断。所谓奇偶,指的是音段的音节数。特别是在书面语中,奇段中存在单音词的可能性远远大于偶段,所谓奇偶判断,就是指利用这一项先验知识。

再下一步,就可以利用搭配知识并进行音节感知处理了。

词的常用搭配是概念关联性的一种表现,我们正在词库和字义库里加入这种知识。不过,应该指出,这种知识是死知识,只能作为概念关联性知识的补充,而不能代替它。概念关联性的表达,从根本上说,要依靠层次网络的符号体系,即字义和词义的 HNC 映射符号。这是语言理解的核心问题之一,以后有机会再详谈。

现在回到音节感知处理问题。也许可以说音节感知是汉语特有的现象。中国人学西语口语,就有一个从音节感知到音节串感知的习性转变问题。可以设想,远古年代的汉语,音节感知起主导作用。因为汉语的基本命名都是用单音节来表达的。现代汉语虽然是以双音节为主结构,但单音节表义的历史痕迹并未全部消失,集中表现在下列五种概念的表述上:

1. 逻辑概念
2. 基本概念

3. 基本命名

4. 人物标志

5. 物性概念

对于上述分类,这里不来解释,它们都相应于概念层次网络符号的一类节点。

所谓音节感知处理,就是有针对性地在音段中寻找上述五类概念,特别是逻辑概念。

逻辑音的使用频度最高,它们又是语义块切分的重要标志之一,可以说,逻辑音的判定,是理解处理的基石。我们曾一再呼吁语音识别系统要对逻辑音予以特殊关注,原因就在于此。我热切希望这一呼吁能得到充分的反响。

这里,我愿意把汉语中的重要逻辑字列表于下,以引起更多的关注。

把 被 不 并
从 出
但 的 对 到
而
个(各) 过
和
就 将 及(即) 仅
可
了 来
么 每
那(哪) 能
起(其) 却
是(使)
同
为
向(像) 些
要 也 以(已,一) 应 于(与) 有(又)
在
这(着)

上列 40 个音、51 个字的辨认,既是基础性又是工程性的一项研究课题,汉语理解处理要更上一层楼,有赖于这一课题有所突破。但这一课题又不可能脱离其他课题而独立进行,要协同前进。

上面我们列举了五类单音节概念,这几类概念的同音混淆是首当其冲的问题。例如,“把”与“八”;“就”与“九”;“仅”与“近”。这是音调模糊造成的混淆,针对这一类语义混淆,先验知识大有可为。这就是我反复强调要测定语音识别系统稳定模糊区的原因。

现在我们可以给出音节感知处理的定义了。它就是对上列五类常用单音节概念的辨认,特别是逻辑概念的辨认。(注:后来实际上分成 8 类)也就是说,首先是概念类别的辨识,

而不是特定概念,更不是特定‘字’的辨识。这个步调很重要。

问4:

先生对于音节感知和处理的设想,我已有了初步的印象。我的体会是,先生把分词的侧重点先放在单音词的义类感知。但据我所知,现有的分词系统对于词库中没有收录的词或者新出现的词往往无能为力。先生对此有何设想?

答:

辨认新词的困难,只要你不拘泥于仅仅运用语法手段,而是采用概念层次网络理论模型的句类分析法,对于文字文本,是不难解决的。但是,要把这一点说清楚,就不能不涉及该理论的核心思想,那就不是三言两语的事了。所以我们还是先把话题限于理解的外围问题。

词典里收集的词都是符合语言学定义的词。不过词的定义本来就有一定的模糊性。所以同一词典的不同编纂者的收录标准有所不同是不奇怪的。例如,“的话”作为助词被收录了,但类似的‘来说’未被收录;“越来越”也未作为副词而被收录。类似的例子还有‘愈益’、“其人”等等。所以,供理解处理使用的词库不能只是词典的翻版。它应尽可能收录不符合词的标准定义,然而却在语言中经常使用的‘搭配’。

粗略地说,可划分两类搭配,一类是词一级的搭配,它反映概念之间的关联性,另一类是“字”一级的搭配,它主要反映语法知识。例如助词‘着、了、过’与动词的搭配,同样的搭配实际上还有动词与‘到、于、成、出、得、来、开、起、去、住’等的搭配。无论是词典或词库,不可能也不应该把所有这些搭配都收录进去。于是它们就成了你所说的‘新词’中的一大类,对于这一类‘新词’的处理,显然需要对于搭配知识的运用。

搭配知识可分为规则性、启发性和习惯性三大类。对于前两类搭配的科学表达,是语言知识表达的根本问题。层次网络理论正是从这一问题的反复思考中萌发出来的。

语法学早就对规则性搭配知识给出了形式上的表述,这就是:形容词与名词、副词与形容词或动词搭配形成偏正结构;名词与动词搭配形成主谓结构;动词与名词搭配形成动宾结构;动词与形容词或副词搭配形成后补结构等等。这些规则的包容度太大,对于我们当前实际碰到的问题来说,它们所能提供的判断知识远远不够。我曾经作过初步统计,大约只能减少12%的模糊度。

上列每一条规则的包容度都可以进一步减少。例如,“肤浅”或‘深刻’主要是用于描述理解;“激烈”主要是用于描述冲突或竞争;“残暴”主要是用于描述人的性格或行为;“快”或“慢”主要是用于描述过程;“教育”或‘诽谤’的对象一定是人;“种植”的对象一定是植物;“驯养”的对象一定是动物;“探索”的对象一定是未知的事物、现象或问题。不仅如此,所有这种行动的主语都一定是人。所有这一类知识能否以一种简明的形式给出比语法规则更精细的规则呢?

这就是概念层次网络模型试图达到的目的之一。它把所有这一类知识,或概念关联性,归纳为两条规则:

一曰:同行优先搭配

二曰:交链式关联优先搭配

第一条规则主要是对偏正结构给出语义约束,把它细化为:概念矩阵中同行不同列(由字母串表示)的概念优先搭配,其符号表达形式为:下列组合优先

(ui-k,gi-k)　(ui-k,vi-k)　(gi-k,zi-k)

(vi-k,ri-k)　(xi-k,wi-k)　(gi,gi-k)　(vi,vi-k)

同行判断简化为概念表达的数字串是否相同。

当然,同行搭配不能表达全部的偏正结构,跨行"偏正"结构大量存在,但它主要涉及 j 类概念,这是单设 j 类(原 15 行)的原因之一。

第二类规则对另外三种语法结构给出语义约束,把它细化为:某些概念节点之间存在强关联性。其符号表达仍是某些组合优先,但这里不可能一一列举,仅举其大者:

(p,vi i≥6)　(v10-i,p10-i;p12-10-i)　(v10-i,g10-i)

(v↑,w;p)　(v↓,g;r)

最后两项优先组合表示:作用型动态概念优先 w、p 类概念为宾语,效应型动态概念优先 g、r 型概念为宾语。

请原谅,我有点违反先谈外围的约定了。不过既然谈到"新"词问题,就不能不涉及概念关联性。通过网络节点去表达关联性,显然是一条捷径。当然表达符号或方式值得仔细推敲,或者说,要精心设计吧。

来源于语法现象及概念关联性的"新"词,只是"新"词的一部分,但终究是主要部分。为表达这部分知识而专门建立一个字知识库是很有价值的。这个字知识库的重点是表达〈问答 3〉中提出的五类概念字或字的义项。

这个知识库为判定语句中的单音词及"新"词提供规则性及启发性知识。单音词和"新"词的辨认,实质上是同一问题的不同形式。

"新"词的范围,还应该包括人名、地名、单位名、各种简化名称,新出现的词,或老词新意(例如"走穴"和"下海"等等),这些"新"词由于个性太强,规则性或启发性知识都作用不大,只能主要依靠死知识了。

问 5:

先生刚才谈到要建立一个表达概念关联性的字知识库,我猜想,先生将用层次网络符号来表示字义,因为这种符号蕴涵了某些规则性知识,但启发性知识如何表达?

答:

所谓"规则性知识"和"启发性知识"之间并没有严格的分界线。一般说来,前者可转化为一个判断推理程序,而后者主要用于引导总调度程序的运行路线,也就是作为调用什么规则或查询什么知识库的判据。这是第一项差别。

例如,对于音节个数为奇数的音段,应优先检查其中是否有单音词,这就运用了一项启

发性知识——“奇段中存在单音词的可能性远远大于偶段”。

在检查单音词的过程中,首先是判定重点对象,这就要查询音节知识库,这个库主要是提供启发性知识。

规则性知识与启发性知识的第二项差别是:前者形成确定性判断,后者形成可能性判断。当然,这两种判断又是相互转化的。例如,对〈问答 1〉中示例,依据〈问答 4〉中两条基本规则,可将模糊音节对 fa-yuan 和 qi-su 优先转化为“法院”和“起诉”,并从而产生该句为作用句的假定。这个判断推理过程就具有双重性。

所以,不能硬性规定知识的规则性与启发性,同样对层次网络符号更应持这种观点。请注意,在〈问答 4〉中陈述的两条规则,都用了“优先”这个词,这就留下了“可能性”的余地。不应忘记,现阶段理解系统程序的设计,要充分发挥人机交互的作用,要时刻想到:该系统只是人的助手,而不是全部代替人。对于你所能运用的全部知识,都应当作为启发性知识来对待。

问 6:

语言知识、语言规则都具有启发性的一面,即不确定或模糊的一面,这一点是语言界的共识。正是基于这一认识,语言界开始引进统计和概率的方法,产生了语料库的思想。我曾听到有一位国内学者自称“我是语料库派”,这既表明人们对语料库的重视,也表明对语料库的看法有争论。而先生对统计的方法似乎基本持否定态度,是这样吗?

答:

关于统计方法或语料库的问题,首先应该看到汉语和西语的重大区别。西语需要通过语料库统计出语词的相关性,而汉语的词库就是这一相关性相当完备的表达。但词库是以词典为蓝本的,并不需要语料库,如果你的目的是寻求词与词之间的常用搭配,当然可以利用语料库。然而,常用搭配主要是概念关联性的体现。那么,为什么不从建立概念关联性理论模型这一根本问题着手?

当然,词的搭配有它的个性,这不是理论模型所能解决的。个性是习惯和约定俗成的问题,例如我们说“毕业,完工”,不说“毕工,完业”;说“新年旧岁”,不说“新岁旧年”。这就要把握搭配的个性。个性对语言生成或表达至关紧要,所以,有些机器翻译学者不大赞成搞什么语言理论模型,是可以理解的。但个性对于理解和解模糊的影响是比较小的。

语料库对于语言个性的揭示,无疑有较大意义。但我们当前工作的重心不应该放在个性方面。至于效果显著的常用搭配知识,本来就存储在你的大脑皮层里,通过词典的诱导,可以比较完整地检索出来,不必借助于语料库和统计手段。

我们正在词库里增加搭配部分,以反映词的常用搭配。这一部分死知识绝大多数是〈问答 4〉中两条基本规则的示例,这样,以层次网络符号表达的概念关联性知识(活的)和音序码表达的词搭配知识(死的)是重复的。但这项重复是必需的。存储方式的“浪费”可以换来

判断过程的效率。更重要的是,理解系统的完善化,必须经历从简单到复杂,从低级到高级的学习过程。还是让人先来充当妈妈,扶着这个刚学走路的婴儿吧!

基于扶婴儿走路的观点,我对于语料库有极大的兴趣。需要设计多种多样的语料库,供未来的理解程序使用和学习(检验)。

这种语料库不是语料的堆积,要配备一系列的程序,以配制各种类型的语料。打个比方来说,语料库里不仅应具备山珍海味的好原料,更要配备优秀的厨师,否则做不出美味佳肴来。

语料库的选材要有全面性和代表性。当然也要注意到重点和雅俗兼有等等。

从文体来说,四种题材的语料库都要选一些,首先是叙述型语料,纵的方面包括报导、传记、历史、游记等等,横的方面包括叙境、叙事、叙人,其次是论述性语料,纵的方面包括政论、论文、专著,横的方面包括论事、论理、论人。最后是抒情和对话,这主要是从小说取材。

这些原始的语料以原始的形式存放,只不过把汉字改成音序码。当然,编目是要精心设计的。

但更重要的是'厨师'程序。这些程序分为两大类。一类是把原始文本变成各种模糊文本。例如,如果要检验理解程序对于 de(的,地,得)、ji(及,即,几)、xiang(向,像)、yi(以,已)、jin(仅,近)的辨认能力,就将文章中有关各字的调序码去掉,只保留基本音号。而要模拟博士卡的输入方式,就将常用字(认识字)的调序信息全部去掉,但是也可以保留一部分最常用逻辑字(如:的、了、在、和、是、有)的音序码。要模拟语音识别系统的输出,可以只保留音码和调的信息,同时送入候选音及其加权信息。

第二类'厨师'程序是把原始文本变为各种标准分析文本,基本的分析内容包括:语义块切分、要素判定、语句类型判定、伴随成分判定、语境判定、主题判定等。这些分析工作都基于文字文本来进行,当然这些都是理解本身或其预处理程序,但这些程序最终应纳入语料库的'厨师'程序库。

问 7:

先生关于语料库的设想,我已经比较清楚了。但我从先生刚才的谈话中更感到先生轻视统计的作用。难道语料统计的国家级成果——汉语词频统计不是很重要的资料吗?

答:

讨论学术问题不必采用外交辞令,当然措辞恰当仍然十分重要。一些不必要的学术争论不能说与此无关,这一点请你们注意到。

从理解工作的现阶段状况来说,急需一些语法现象的统计知识,可惜得很,词频统计里没有反映。现有的词频是无条件概率知识,而一些条件概率知识更有实用价值。

我不是说现有的词频统计没有价值,当前市面上各种语言处理系统就利用了这项知识。

办法也很简单,就是凡有待选的词或字,就取词频最高者。有的产品出于宣传的需要,冠以智能二字,也无可厚非。因为,智能产品本来就没有严格定义,它终究用了一点知识。

汉语的常用字绝大多数有多个义项,但不同义项的运用方式,在现代汉语里却有明显差别。有的义项可以用该字独立表达,有的义项则不能,需要与其他有关的字联合起来,少数甚至不能独立表达。例如"经"字的"经营"义项是必须联用的。由此义项构成的词,如经纪、经理、经商、经销、经售都是同一网络节点内部的组合。如果与其他网络节点联合使用,则不能省去"营"字,你不能对"经"字的这个义项独立使用。例如,不能说经(木材、五金、土产),另一搭配字(营、销、售)绝不能省去。但"经"字另一义项"经过"却可以独立运用。例如:"这件事要经过某人办理",你可以省掉"过"字。因此,在说明字义时,要引入"独立性"的概念,在各义项里给出相应的标志,这对于新词及单音词的辨认极有参考价值。

字的总频度与其独立义项的频度没有必然的联系,总频度高的字,其独立义项的频度可能很低。如"实、发"两字,是常用字。由它们构成的常用双字词(按词频统计)都多达 29 个(总共 4200 个),与"不"字一起雄居群字之首,但其独立义项的频度很低。

理解处理不仅关心字的独立义项的条件频度,也关心字在句首或句尾出现的条件频度。例如"频度最高的字"的"的(de),在句首的条件频度为零。这虽然是一个特例,但说明了这两个条件频度也需要另行统计。在论述及叙述(非小说)型的文本中,单字词与双字词的比例大约是 1:2,即平均每 5 个字有一个单字词。但在句尾,单字词的比例要小得多。如果仅对句尾的单字词作字频统计,将会得到一个更集中的分布。这一项统计知识也是十分有用的。

以上所说的各项条件频度,当前都没有可靠的资料。我们不得已自己动手做了一点统计。所以,我们是以自己的认识来推进统计工作的。如果说全国性的语言统计工作有什么需要改进的地方,那就是不要再陷在太小的圈子里,应多听取各方面专家的意见。

问 8:

请原谅我在上一次提问中用词不当,但我的本意是希望听到先生的尖锐评论,因为,我对于语言界低水平的重复工作也深感失望。我读过先生写的内部报告"关于语法学的贡献和不足",先生的看法确实与传统语言学大相径庭,很想听一听先生对语法在理解处理中的地位或作用的看法。

答:

你刚才提到传统语言学一词,我想补充一下。你所说的传统只是"五四"以后引进学派的传统,不是传统一词的常规意义。中国传统语言文字学,原称小学,是从音形义三者的联系去研究语言现象的。当然是以古汉语为主要对象。这个可贵的传统可惜引进学派没有继承下来。

中国二千余年的封建文化传统究竟不同于西欧中世纪的黑暗时代,儒学也不同于宗教

统治。15 世纪开始的欧洲文化革命,基调是复兴和创造。我们在本世纪开始的新文化运动,基调则是打倒和模仿(引进)。这个基调在当时是有其巨大历史功绩的,但也有片面性和消极性影响。语言学是深受这一消极影响之害的重灾区。

语法学的基本概念是主语和谓语,这个概念是公元前四世纪亚里斯多德提出来的。在那个时代,语言学、逻辑学、自然科学和哲学是不分家的。主语和谓语实际上是命题逻辑的概念。然而,语言的内涵和表达方式不能用命题这么一个框子去概括。你总不能把‘寻寻觅觅,冷冷清清,凄凄惨惨戚戚”;“两情若是久长时,又岂在朝朝暮暮’这样美妙的语言称为命题。

语法学在这个框子里脱离语义去探求语言的形式规律。对于印欧语系,这种研究方式仍有较大的意义。因为,印欧语系各语种之间的语法现象大同小异,纯形式规律十分发达,且蕴涵着较多的语义因素。例如,时态和格的概念里主要是语义的内容。汉语没有姊妹语种,形式规律又极不发达。因此,如果说汉语也要进行语法研究的话,那本来就不应该走脱离语义的道路。

不幸的是,汉语的语法研究不仅走上了这条道路,而且由于受到‘打倒’和‘模仿’这一基调的支配,它比西方语法学更墨守成规,更不敢迈入语义的雷池。因此,本来与中国传统语言学的思想一脉相承的格语法、概念从属理论、功能语法理论等等都产生于西方。中国语言界甚至连跟踪的步子都很慢。

许多语言信息处理工作者抱怨汉语语法学未能为汉语信息处理提供足够的知识。实际上这种抱怨是不公平的。抱怨者不了解汉语的特点,他们所期望的那种语法学知识根本就不可能存在。由于语法学的名声太响,许多人是中了迷信语法之毒而不自知了。

理解的根本问题是把握概念之间的关联性,主要是语义问题。特别遵守脱离语义这一宗旨的汉语语法学能有多大作用?

我说得够尖锐了。关于理解的外围问题,今天大体上都谈到了。下一次我们转入理解的核心问题。不过,我希望你先仔细阅读有关概念层次网络理论的内部报告。否则,我们缺乏共同语言,讨论起来就比较困难了。

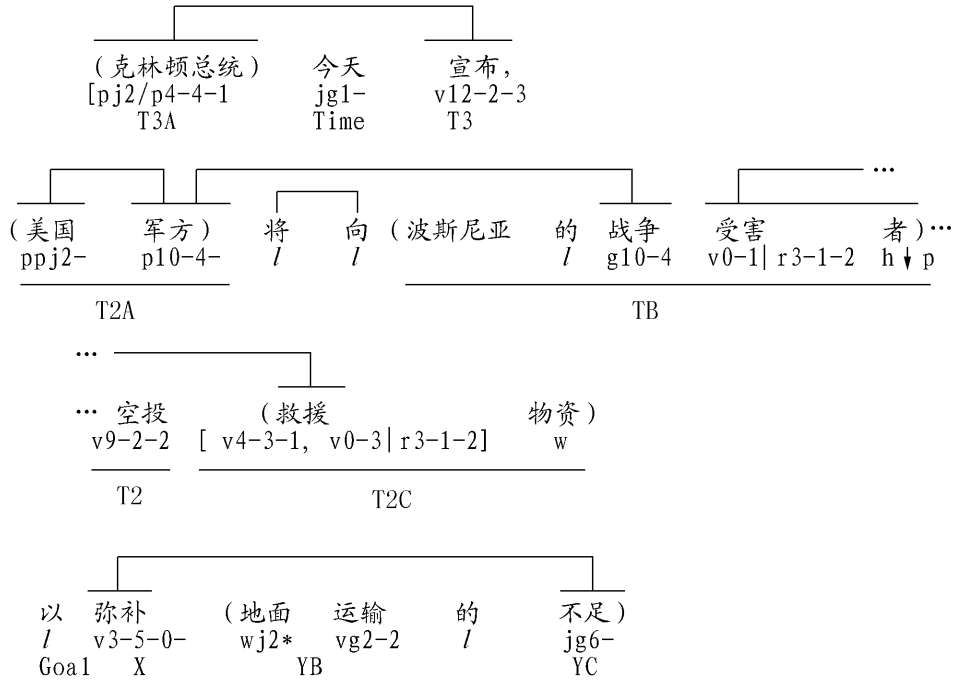
问 9:

我认真阅读了概念层次网络理论的有关文章,对于这一理论模型的雄心有所领会。但是,并没有豁然开朗的感受,相反倒有一种‘如堕五里雾中’的迷惘。我很想再看到一些概念层次网络分析的样板,从感性认识起步。

答:

让我们先分析一段具体的语料:

(1)



在以下的讨论中，将基本采用上面的书写格式。第一行，汉语文字文本；第二行，层次网络符号；第三行，语义块符号。多个单词构成的语义块用括号表示。规则关联性概念用连线表示。

这段语料，有三个语句。前两个是转移句，后一个是作用句。

转移句有下列 8 项要素：

1. 特征要素 T
2. 转移发动者 TA
3. 转移接收者 TB
4. 转移内容 TC
5. 转移起点 TB1
6. 转移终点(目的地) TB2
7. 转移工具 In
8. 转移途径 TB3

这 8 项要素中，前 4 项和第 6 项是基本要素，其他 3 项是附属要素。

转移网络有 5 个二级节点，相应于 5 类转移句。但节点 2-4(交换与替代)在表现形式上更接近关系句，句类分析时，将按关系转移句处理。从内涵看，有 6 类转移句，如下面所示：

一般转移句 TJ = TA + T0 + TB + TC

传输句 TJ = TC + T + TB2

自身转移句　　TJ = TA + T + TB2

转移接收句　　T1J = TB + T1 + T1C

物转移句　　T2J = T2A + T2 + TB + T2C

信息转移句　　T3J = T3A + T3 + TB + T3C

在形式上,也可以说转移句有 3 类,它们分别以 TA、TB 或 TC 为参考点:一般转移句以 TA 为参考点,转移接收句以 TB 为参考点,传输句以 TC 为参考点。参考点大体上相应于语法中的主语。转移句提供的要素约束条件是:

T2C　　优先于物(w)

T3C　　优先于信息,并经常扩展为语句

TA　　优先于人(p)

TB　　优先于人(p)、广义空间

现在来作语句分析:

第一句　由　克林顿总统　∈p

今天　∈jg1

宣布　∈v12-2-3

知　该语句为 T3J(信息转移句)

但　TB、T3C 为空集,J1 = T3A + T3 + 空

推知　T3C = J

第二句　由　美国军方　∈p

空投　∈v9-2-2

物资　∈w

受害者　∈p

将、向　∈l

知　该语句为 T2J(物转移句)

J2 = T2J = T2A + l + TB + T2 + T2C

第三句　由　弥补　∈v3-5-0-

运输　∈vg2-2

不足　∈jg6-

以　∈l

知　该语句为 XJ(作用句)

J3 = XJ = X + YB + YC

分析过程就是如此,"五里雾中"的感觉是否有所减轻?

问 10:

在第一句中,只有一个动词'宣布",句类的判断是唯一的。但在后面两句中,都分别有

两个动词,“空投”与“救援”;“弥补”与“运输”。如何决定舍取?

答:

切中要害。这正是所谓汉语理解难于西语的根据之一。在西语,此处的“救援”和“运输”两词将在词性上与“空投”和“弥补”有所区别。

这里的问题似乎是:两个动词,谁是语句的重点或特征要素?但这样提出问题,并不恰当。

关键还是要从概念关联性去考察语句的各个成分。对句例 2 首先要看到,“空投”优先于“物资”;“救援”优先于“受害者”。“空投物资”和“救援受害者”是两个更完整的组合概念,都表达了人类的一项活动,前者是转移活动,后者是作用效应活动或作用关系活动。这两项活动互为因果,转移的作用是救援,救援的需要引发空投行动。这两种因果观导致两种表达顺序:

1. 空投物资,救援受害者
2. 救援受害者,空投物资

但不论是哪一种顺序,强关联的概念对“空投—物资”和“救援—受害者”本身仍然是按照自然顺序(这是就汉语来说的)排列的。

现在第二句打破了正常的自然顺序,将“救援”与“受害者”分离,而插入到“空投”与“物资”之间,并与“物资”形成紧搭配,在形式上“救援”成了“第三者”。这种“插足”现象是表达有所侧重的标志,表明该语句强调了“空投”的中心地位。但这种重点的倾斜并不影响内涵。该句在形式上可按转移句分析,但其内容仍是两项人类活动:“空投物资”和“救援受害者”。我在〈问答 2〉中谈到隐知识的揭示,这就是内容之一。

例句 3 的情况与例句 2 类似,“弥补”与“不足”是强关联性概念对,“弥补不足”是一项作用效应活动。现在插入了“运输”一词,但该句中并无“运输”的强关联性概念(w),而作用“弥补”尚缺少对象 YB,YB + YC 只是作用对象的一种自然顺序。这一句因为涉及概念较少,而动词之一无强关联概念存在,问题就比较简单了。

问 11:

先生刚才的论证,我认为,关键在于关联概念的把握。不能抓住“救援”与“受害者”、“空投”与“物资”、“弥补”与“不足”的强关联性,则往后的分析就无从着手。但是,与任何一个词强关联的词往往不胜枚举,计算机如何掌握这浩如烟海的知识?

答:

如果仅仅运用统计手段的结果,形成词关联的死知识库,这个问题的解决,几乎没有尽头。当然,我不是否定死知识的必要性。这在〈问答 6〉中已经谈过了。

然而,概念关联性不同于词的关联性,它基本上以层次网络的二级节点为关联单位。对象的数量不是以万计,而是以百十计。更主要的问题还不是数量,而是只有通过这种表达方式才能形成理解的基础。

概念关联性的表达主要是通过节点的定义和运用结构符号规定节点之间的关系。具体分为三种形式(判断的形式):

1. 同行优先。如军方与战争,运输与空投,美国与军方。
2. 链式关联或要素约束。如总统与宣布,空投与物资,救援与受害者,弥补与不足。
3. 交式关联。如克林顿与总统。

你看,上列关联性在层次网络符号体系里表达得清清楚楚,计算机仅仅通过相应的数字串和结构符号就能作出与人相近的结论。我希望你沿着这条思路往前想。

问 12:

计算机把握第一点和第三点似乎比较容易,但第二点难点甚多,愿闻其详。

答:

问题在于,这三条规则是否展示了一条把握概念关联性的途径以及如何沿着这条途径步入理解的殿堂。

这个步入过程将是漫长而艰难的,但重要的是,要把握步入过程的顺序。这个过程的前两步就是行内关联(同行优先)和行间关联。在这一问答中,就着重讨论第二步。

让我们先解剖一个麻雀:二级节点 3-5,句例中的词'弥补'属于它。

节点 3-5 的关联节点如下表所示:

	交 式	链 式	混合式
行内	3-3(扩展与缩小)	3-1 (利与害) r9-3-0 (功与过)	
行间	0-3(作用的免除) 0-4(约束) 4-7(协调)	1-0 (过程) 7-2-1 (能动性)	j6(度)

我先说一个小故事。当初设计 3-3,3-4,3-5 节点的时候,最早的出发点是,它们分别与基本概念'量、质和度'相对应,这三个概念的最初编号是 15-3,15-4,15-5,现在变为 j4,j5,j6,数字串的协调性失去了。我曾为行间关联性的数字表示花了不少心思,往往顾此失彼。现在看来,此项设计的优化,不能过于追求形式,不必现在又将 3-3,3-4,3-5 加 1。(注:最后还是实行了"+1"的改动。)

3-5 节点的定义是调节与控制,它有两个对偶性三级节点(所有效应网络节点都有此项特性):推动与抑制,分别记为 3-5-1,3-5-2。它们与基本概念"度"的关联性是明确的,度不足,需要推动,度太过,需要抑制,这就是链式关联性。另一方面,"度"的动态表达 jv6,本质

上就是调节,所以与 j6 又存在交式关联性。

调节的两个对偶性节点,推动与抑制,分别与“作用的免除”ð-3、“约束”ð-4 交式关联,这是同一核心概念的两个不同表达角度。同样,调节本身 3-5-0 与节点 4-7-0(协调)交式关联。表达角度与语句的侧重点密切相关,这里不来讨论。

调节的对象可以是作用效应链的任一环节,因此,3-5-0 在作用效应链内部的链式关联节点似乎是无所不包。但从它的两个对偶性三级节点来说,主要是过程 1-0 和人的能动性 7-2-1。这一点,可作为计算机把握 3-5 链式行间关联的起步知识。

在行内,3-5 与 3-1 是链式关联,“趋利避害”这句成语中肯地揭示了这一点。3-5 与 3-3 是交式关联,因为,度的调节与量的增减是紧密相关的。

如果你希望进一步获得这一概念节点的感性认识,不妨联想下列组合概念:

推进(3-5-1)改革进程(1-0-10)

发扬(3-5-1)成绩(r9-3-0-1),克服(3-5-2)缺点(r9-3-0-2)

发挥(3-5-1)聪明才智(g7-2-1)

弥补(3-5-0-)不足(jg6-0-1)

调动(3-5-0)积极性(g7-2-1)

鼓励(3-5-1)好人好事,打击(3-5-2)歪风邪气

发展体育运动,增强人民体质

严禁吸烟

壮大己方力量,削弱对手实力

奖励有功人员,鞭策后进分子

从这一系列联想中,应能得到下列两点重要认识:

1. 一个概念节点必然与另一些概念节点强链式关联,形成语句内部要素之间的约束条件,如节点 3-5 的关联矩阵所示。它是语言现象中最普遍、最重要的规则性知识。中国传统语言学的研究风格本来为这一规则性知识的揭示敞开着大门。是“引进”学派用语法学的大简化风格把这个大门堵住了。这是代价重大的历史教训。

2. 但是,上述规则性知识仍然不过是“引玉之砖”,更重要的是隐藏在规则性知识后面的启发性知识。比 3-5 关联矩阵更深层的知识是:

适度—积极—利, 失度—消极—害

这是一对链式概念组合,它不仅揭示了节点关联矩阵的内在联系,还提出了一项重要启示:对规则性知识不能仅当作死规则对待,要从联想规则作进一步的联想,上面的示例就是为了对这一点给出一个初步印象。

问 13:

先生在〈问答 4〉中谈到语法规则的包容度太大,先生用概念节点关联矩阵的方式去表达语言现象中最普遍、最重要的规则性知识。对此,我已有了清晰的印象,从我熟悉的信号

处理概念来看,可以说是倍感亲切。但是,概念要通过词来表达,而词的个性往往很强,“弥补”就是很有个性的,而它的层次网络符号与“调节”相同。如何表现它们之间的差别,我的直觉是,延长数字串的方法并不可取。

答:

你的直觉是完全正确的,我以前关于扩展层次的设想需要修正。

个性的表达主要是通过组合结构符号。现在就以“弥补”为例,详细说明一下这个问题。首先,需要给出“度”的基本内涵。“度”的衡量应分三级,“不足、充分、超过”,分别记为 jg6-0-1、jg6-0-2、jg6-0-3。“补”字作为“补充”的义项,精确的表述是:“通过增加数量的方式,产生一种调节的效应,变‘不足’为‘充足’”。层次网络符号把这一表述简化为:

[v3-5-0,v3-3-1]↑jg6-0-2

在这个表达符号中,“方式”和“变不足为△△”的概念是隐含的。不难把隐含变为显含,但当前不打算这么做。

“补”字在“将功补过”一词里的义项,其精确表述是:“人们通过再一次活动,取得积极的效应,取代原来活动的消极效应”,其层次网络符号是:

v9-3-5-0↑r9-3-0-1

这里,精确表述中的第三句话隐去了。“弥补”一词的语义,是以上两者的“或”:

[[v3-5-0,v3-3-1]↑jg6-0-2; v9-3-5-0↑r9-3-0-1]

这里顺便说一点,“弥补”中“弥”字的本义是 ju6-0-2-。“弥补”一词是语法学的偏正结构,“弥”字的作用与后补结构词“补足、补充”里的“足”字、“充”字完全相当,它本身并无“补”的语义。可是,汉语词典因有“弥补”一词就给“弥”字加了这条语义。由于西语无“字”,引进学派也就不注意字义与词义的区别。这种把词义误作字义的错误并非罕见,作为一部工具书是不应该的。这种错误将导致错误的音节感知,例如,对“mi”产生 3-5-0 的联想。因此,我们在建立音节知识库时,应注意防止这种错误引导的发生。

回到个性的表达问题,个性也是分层次的。每一层次的概念都有个性,因而,可分出下一个层次。但是,如果这些个性主要表现在与其他节点的链式关联上,则不如直接表达这种关联性,而不采用扩大层次的方法。上面,对“弥补”这个概念的表达就是如此。

为了表达概念节点之间的链式关联性,我们引入了四种符号:↑ ↓ → | 。每个符号的左右两方分别记两关联的节点。↑与↓的左方节点为作用,右方为效应,构成一个作用效应对。前者突出作用,将构成作用句的 X 要素;后者突出效应,将构成效应句的 Y 要素。→的左方也是作用,但其右方是作用的对象。如 3-5-1 的“鞭策”一词,可记为 v3-5-1→p,表示这一概念的对象一定是人。符号 | 的定义是:左方给出一个概念,右方给出该概念的有关内容。引入这个符号,是为了与→相区别。在语法学,把四种符号的右方统称为宾语,这实在过于宽泛。于是,有人从语义的角度将宾语细分为十余类,甚至数十类,这又不得要领。我认为,宾语首先要把效应、对象和内容三者分开,这是最本质的区别。让我们用四个例子来表明这一点。

鼓励上进　　是作用—效应
种植蔬菜　　是作用—对象
获得成功　　是效应—内容
禁止吸烟　　是作用—内容

链式关联是层次网络理论的中心概念之一,很难用三言两语把它说清楚。我这个始作俑者,对它的认识,也还没有达到华罗庚先生所说的由厚到薄的境界,以后有机会再谈。

链式关联的结构符号可用于表达概念节点之间的关联性,当然包括词与节点的关联。至于字与字或词与词之间的特殊关联,我们现在打算在词库或字义库中直接以音序码(汉字)的方式装入,不考虑具体的概念组合结构。这个缺陷如何弥补,我现在还拿不定主意。

问 14:

关于链式关联性以及效应、对象与内容的区别,我已有了大体的轮廓。请允许我暂时不求甚解,回到句类分析问题。我已经体会到,先生提出句类分析,而舍弃句法分析的传统做法,是为了把握语句要素之间的语义约束,主要是链式关联性,为理解和解模糊处理提供判断的依据或准则。但是,各要素在语句中的位置是灵活多变的,语义约束可能出现交叉或模糊,因此,仅仅依靠语义是不够的。语法学的大简化、大包容特点,虽然脱离了理解,但对于判定主谓宾成分仍然简明有效,因此,我认为,对语法学的成果不能一笔抹杀,而要加以合理地利用。

答:

如果我关于语法的评论产生了"一笔抹杀"的印象,那确实十分不幸。在〈问答 1〉中有这么一段话:"这个句例中有三个 p 类概念,刘嘉玲、法院、实业公司,它们之中,是哪两个充当 A 和 B 的角色,则需要运用语法知识了"。因此,如果我说,"我完全同意你的看法",你不应该感到奇怪。

语言的最大个性是它的语种性,具体表现就是语法。脱离了语法,语言理解几乎不可能。语义与语法的关系是内容与形式的关系,两者不可能独立存在。

语法的"介质"是"虚词"和形态变化,汉语只有第一项。我在〈层次网络理论概述〉第 14 章对这项"介质"已作了较系统的介绍,这里结合〈问答 9〉的句例,具体说一下"虚词"的作用。句例中有 4 个"虚"词:"将"、"向"、"的"、"以"。这几个词都是高频词,值得详谈。

"将"字的"虚"义有下列几项:

1. lvq4-0-1-1-0-3
2. [l0-2-0-0,l0-3-2-0]　　相当于"把"
3. lv4-0-12-7-0　　相当于"用"

义项 1 的逻辑层 lvq4-0 表示:它是基元概念的简化(由 4 决定,0 暂无意义),并用在动词之前(由 vq),其语义由基元层 1-1-0-3 给出,表示"未来",如下图所示:

1 - 1 - 0 - 0　　顺序
↓　↓
过程 序　　- 1　　过去
- 2　　现在
- 3　　未来

义项 2 的逻辑层表示它是语义块指示符,具体指示是作用的对象或转移的内容。

义项 3 的逻辑层与义项 1 仅相差一个字母 q,也表示基元概念的简化,但基元层的意义是手段,如下图所示:

12 - 7 - 0 -　　手段
↓　↓
人与事物　手段　　-1　　方式、方法
-2　　工具
-3　　途径

句例第二句中的‘将’字应取义项 1。如何选择义项(多义模糊),这里不来讨论一般原则,但就这个句例来说,问题似乎比较简单。因为它后面的‘向’字是一个语句要素指示符(l0),而且 l 义项唯一(见下文说明)。根据矛盾律(思维三基本规律之一),“将”字的义项 2 和 3 都可排除。具体说来就是:“将”的义项 2 也是 l0,而两个 l0 不可能连用。义项 3 是语句伴随成分指示符,也不可能与 l0 连用,连用就违反了矛盾律,因而只能选用义项 1。

“向”字的 l 义项是:

[l0-2-2-0;l0-5-2-0;l0-4-1-3;[lv0-,jg2-1]]

在 l0 的意义下,有四个选义,分别指定转移的对象、转移的终点、过程的趋向,空间的位置或方向。由‘向’∈ l0 可推知‘将’∈ lvq4-0-1-1-0-3,已如上述,而‘向’的具体选义,则由语句的特征要素 E 确定,由于该句的 E∈ v9-2-2,根据同行优先规则,“向”的义项应取 l0-2-2-0 或 l0-5-2-0。

我的这些推理性陈述,已高度‘专业’化,如果你不熟悉层次网络符号,那肯定如堕五里雾中,你能适应我这样的叙述方式么?

问 14 续 1:

我基本上能适应先生的叙述方式,不过,正好有一个小问题。先生在〈概述第 14 章〉中,只写到 l0-k 的 k 取值 0,1,2,3 分别指示特征要素 E、作用者 A、对象 B、内容 C。但刚才出现了l0-4、l0-5,我猜想,对转移而言,4　7 相应于〈问答 9〉中所列 8 项要素中的后 4 项。对吗?

答:

你可谓举一反三,已窥门径(注:语义逻辑符号 l0k 的设计最终只保留 k = 0,1,2,3。)。这里再来说一个小故事。作用、过程、效应、状态的对象都比较明确,或者说,对象与内容可以给出比较明确的划分。转移则不然。试看下面的四句话:

1. 张三送李四一本挂历。
2. 货物已运往上海。
3. 张三告诉李四王五快要结婚了。

4. 王五快结婚了,张三把这个消息告诉了李四。

按照直觉,把挂历和货物理解为转移的对象比较顺,把‘王五快结婚了’的消息理解为内容比较顺。但从转移句的整体来看,这四者是处于同一地位的语句要素,现在统称之为 TC。句例中的张三,没有疑义,充当 TA 的角色,但将李四命名为 TB 似乎有点不协调,所以曾采用过 REC 的符号,但这个 REC 的语法地位与其他句类的 EB 相当,用 TB 是完全合乎逻辑的。语法的双宾语,就是指 TC 和 TB。

关系句也存在划分 RA 和 RB 的困难,特别是对称性关系。我们将把关系双方统称为 RB,分别标记为 RB1 和 RB2,与其他句类的 EA、EB 相对应。

插话到此为止。下面还是回到例句的句类分析。

对于‘美国 军方 将 向 波斯尼亚的战争 受害者 空投 救援 物资’这句话中的 11 个词,已分析了其中的 10 个词,已确定该句为转移句 T2J(见〈问答 9〉),其标准格式是:

$$T2J = T2A + T2 + REC + T2C$$

但这个句子采用了

$$T2J = T2A + l + REC + T2 + T2C$$

的格式。要得到这个分析结果,需运用下列判断:

1. 由于‘向’∈ l0,其两侧各为一个语义块。
2. 由于‘空投’∈ v9-2-2,它本身构成一个语义块,两侧各为一个语义块。
3. ‘向’的右侧语义块就是‘空投’的左侧语义块,因为,在‘向’与‘空投’之间无其他语义块切分标志,相反,却有语义块聚合标志‘的’。
4. 1 号语义块(美国 军方)符合 T2A 的优先条件。
5. 2 号语义块(波斯尼亚的战争 受害者)符合 REC 优先条件。
6. 4 号语义块中的‘物资’符合 T2C 优先条件。
7. 该句为 T2S 的假设成立。
8. 该转移行动尚未进行(由‘将’字)。
9. 该行动目的是(救援 战争受害者)。
10. ‘向’字的 l0 义项应取 l0-2-2-0,而不是 l0-5-2-0,因为 2 号语义块的主体是(战争受害者),而不是(波斯尼亚)。如果该语义块仅由(波斯尼亚)构成,则应优先选取 l0-5-2-0,而且,作出波斯尼亚为地名或国名的猜测。

上列 1 3 步判断就是运用我多次讲过的 lv 准则。4 6 步就是运用语句要素之间约束规则,这里涉及的知识都蕴涵在层次网络符号之中。

这是步入理解殿堂的关键一步。

在以上的说明过程中,我故意避开了三个属于局部性然而又是关键性的问题,就是:

1. 关于‘的’字的知识运用。
2. 关于‘者’字的知识运用。
3. 关于‘救援’与‘物资’之并为一个语义块。

我已经讲得太长了,就此打住吧。

问 15:

先生在〈问答 12〉中谈到,步入理解殿堂的过程将是漫长而艰难的。在上一问答中详细说明了步入理解殿堂的关键一步。我的感觉是复杂的。一方面我能跟随先生的思路,整个推理过程似乎明朗而自然,几乎无懈可击。但另一方面,又感到先生的思考方式过于希腊化,有自设陷阱而不自知的可能性。问题可能就出在词的层次网络符号本身。我有点怀疑,先生是先有了例句,而后写出了句中各词的层次网络符号,因而,一切都显得那么自然。这个怀疑对先生有些不礼貌,请原谅。不过我可以举一例子来说明这一点。"向"字的层次网络符号,先生在上一问答中说得非常精彩。但是,在〈问答 1〉中的"向"字,似乎就得另有解释,这不就是一个陷阱么?

答:

"陷阱"肯定会存在的。问题是"陷阱"能不能填平。词或字的每一义项中的层次网络符号在数量上是不受限制的,而且可展开。因此,如果你说的陷阱没有更深的含义,而是指义项的表达是否完备,那显然总可以弥补。

至于〈问答 1〉句例中"刘嘉玲向上海中级人民法院起诉××实业公司"的"向",倒恰好是 10-2-2-0。法院是起诉信息的接收者。当然,要让计算机拥有这项知识,不是一件容易的事,然而却是可以办到的,因为以语义块表达的上列语句表示式可以给出这一关键信息。这是理解的基本功。你刚才提出的问题,深入探讨的时机还不成熟,因为基本功不够。

回到〈问答 9〉句例的第三句,其中的"以"字还没有给出具体的义项表示。

"以"字是〈问答 3〉中所列 49 个逻辑字之一。请注意已讨论的"将"、"向"二字也在其中。这 49 个字中的(以、一、为、不)4 字极具汉语特色,很值得写一篇专文进行论述。

鉴于"以"字的特殊性,这里把它的主要义项列表于下:

	义项 1	义项 2	义项 3	义项 4
层次网络符号	lv6-0-12-7-0	lv6-0-12-8-2	lv6-0-1-2-1	[l1-0,j4-2]
独立性	强	强	中	弱
句例	以××方式 以××为△△	以求得×× 以达到××	以××的为人	
词例	以理服人 以攻为守	学以致用 以便、以免	以貌取人 物以类聚	以上、以下 以前、以后
关联性概念	v12-7-0	v12-8-2	lv8-0-1-2-1	jgu2- g1-1-0
词例	采用、采取	旨在、为了	按照、依据	

在上表中,特别安排了独立性的说明。独立性强的义项,才能在语句中独立使用。这里需要说明三点:

1. 条件、手段、目的是密切关联的三个概念。手段依赖于条件,服务于目的。这一类的

启发性知识如何表达,显然饶有趣味。但层次网络符号当前做不到这一点,只能用‘并’的组合形式给以极粗略的指示。例如:[lv6-0-12-6-0,lv6-0-12-7-0]或[lv6-0-12-7-0,lv6-0-12-8-2]。实际上,“以”字在实际使用时,往往是义项1与2相并。例如成语‘以儆效尤’中的‘以’字就属于这种情况。在〈问答9〉句例3中的‘以’字也是这样。对于这种情况‘以’字的释义,用lv6-0-12-8-2或[lv6-0-12-7-0,lv-0-12-8-2]皆可。前者属于不求甚解,后者则深入一步,知道‘以’字是手段与目的的双重指示,其语法格式是:

以××手段达到△△目的

但手段的内容‘××’往往是已在前面说明而在本句中省去,上面说的成语例和句例就是如此。

“以”字在义项1与义项2相并使用时,一般语法格式是:

〈以〉〈手段〉〈目的〉

但汉语也有下面的格式:

〈手段〉〈以〉〈目的〉

成语‘严阵以待’、‘全力以赴’就属于这一格式。

2. 义项3实际上应写成lv6-0-1-2-1与lv8-0-1-2-1的或,前者表示自然的因果关系,不涉及人的能动性或主观意识;后者则加入了人的因素。“物以类聚”中的‘以’是lv6-0-1-2-1,而‘以貌取人’中的‘以’则是lv8-0-1-2-1。实际的语言并不十分注意这种差异。因此,在建立字义库时,采取这种不求甚解的方式可能利大于弊。

3. 在〈问答14〉中我提到语义块切分的lv准则。应用此项准则的诸多问题需要一篇专文才能说清楚,这里可以先谈一点。l类概念中独立性强的义项一定是语义块的切分点。这是引进独立性概念的目的之一。但是,一个字的义项独立性有强有弱。例如这里的‘以’字,它在语句中是否充当语义块的切分标志,需要首先确定它是否作为单字词使用。这一点,可以通过查询词库作出初步判断。

“以”字的义项还有很多,这里谈一下引入空义项的必要性。我们将约定空义项的义项序号为零,非空义项的序号从1开始,不论该字实际是否具有空义项。这样的约定对于用语义结构方程(规则)表达词义较为方便。

“以”字的空义项用得很活,如下列词及成语中的‘以’:

常用双字词:　以为、给以、加以、可以、所以、予以

非常用词:　得以、何以

四字词:　夜以继日、坐以待毙

空义项不能理解为无意义,也不是集合论的空。它是出于用字义表达词义的需要。我在〈问答4〉中曾谈到语法学关于词的五种结构的不足,用结构方程的形式可给予更精确的表达。而结构方程8:$Z=\sum X_m Y_m$就必须引入空义项的概念,因为X_m与Y_m的义项数可能不等,需要补零。其次,有些词的意义与字义的联系已很淡薄,不如直接从词义来说明,特别是一些现代的重要概念,如经济、政治、军事等等。这时,可以把词的意义赋予某一个字,而视另一字为空,纳入0号结构方程。这类词可称之为假连绵词。但这样处理又容易同汉语固

有的真连绵词相混淆，如囫囵、蹁跹、蝌蚪。把空义项序号固定为 0，就容易区分这两种情况。它们的结构方程分别是：

假连绵　　(0-1-x-0)

真连绵　　(0-1-1-1)

语义结构方程是汉语特有的一种语义现象。黑格尔曾因为德语中的一些反映辩证法的语义现象大为惊异与赞赏。其实，德语与汉语相比，恐怕是小巫见大巫了。我将在适当的时候，系统地谈一谈语义结构方程问题，这里只说一点，真假连绵的区别依靠结构方程中 3、4 两层的数字就表现得十分清楚了。

区分真假连绵只是语义结构方程一项微不足道的功能。我只是想通过这个小例子说明：对于汉语，层次网络理论除了提供节点的层次网络表示和节点关联性的结构符号表示这两种手段之外(这两点对任何语种都适用)，还提供了一种语义结构方程的手段。我不同意计算语言学界关于汉语理解难于西语的结论，主要就是因为汉语多了这一个重要的手段。你所担心的陷阱，有些是可以通过这第三个手段来解决的。

问 16：

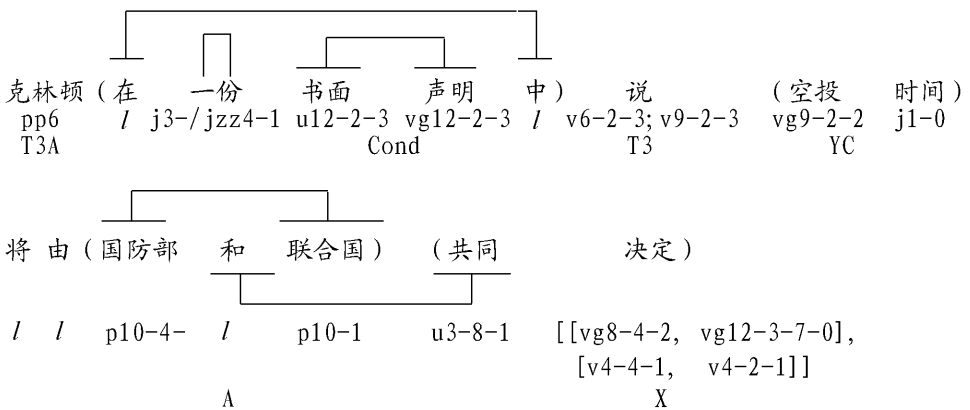
我原来只是把语义结构方程看作一种数据压缩技术，听到先生刚才的谈话，才初步认识到它同时也是描述复杂语言现象的三大手段之一。现在我开始对概念层次的妙用有所领悟。从层次的观点来看，“陷阱”是不存在的。不知先生是否同意。我对转移句已有了一点体会，但转移句似乎比较规范，可总结出几种标准的格式，其他句类是否也能如此？

答：

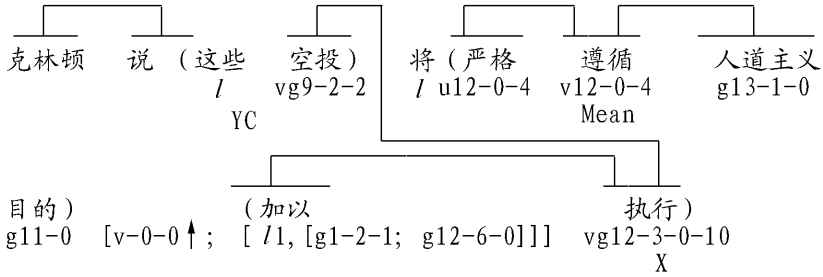
陷阱问题依然存在。层次的概念只是揭示概念关联性的钥匙之一，早已有之，我的发现在于句类。把层次的作用过分夸大，就真要陷入希腊化的局限性里了。

让我们回到具体的句例分析。继续〈问答 9〉中句例的下文，并将句子标上序号：

(2)



(3)



对于这两个句例,首先可得到下面两点观察。第一,它们同〈问答 9〉的句例(将编号为句例 1)一样,都是 T3C 用另一个语句表达的信息转移句。

$$J = T3A + T3 + (T3C)$$

第二,语句中的多数概念存在规则关联性搭配。按〈问答 11〉的方式,列举如下:

1. 同行优先:书面与声明、国防部与联合国、严格与遵循、加以与执行。
2. 链式关联:克林顿与说、遵循与人道主义、执行与空投。
3. 交式关联:和与共同。

这里,需要对'加以'这个词稍加说明。它是一个多义词,其第二个义项表示它可以充当两种非要素指示符,分别指示原因(1-2-1)或条件(12-6-0)。第一个义项 v0-0↑则是汉语特有的表示方法,在英语里不存在对应的词。我们把这种表示称为概念的高层表示,其作用是多方面的。在修辞方面起强调的作用,在语法方面起调整特征要素位置的作用。通常汉语的特征要素在句中或句首,通过高层表示可移至句尾。第三,它表示具体的作用有待用其他词加以补充。这三点就是高层表示 v0-0↑ 所暗含的启发性知识。上面将'加以'与'执行'列为同行优先,是就第三项意义来说的。补充高层表示的词包括作用型概念以及所有与作用有交式关联的节点。

下面重点谈一下两个构成 T3C 的子句,它们都是作用句,其语义块顺序分别是:

$$YC + l + A + X$$

$$YC + l + goal + X$$

在说明这两个子句之前,让我们对作用句作一个简要的回顾。

作用有 5 个二级节点,从内涵说,应有 5 种作用句,但从形式说,分为 4 种较为适当,分别命名为:

一般作用句: $A + X + B$

承受句: $X1B + X1 + X1C$

反应句: $X2B + X2 + (X2A) + (X2C)$

免除/约束句: $XkA + Xk + XkB + XkC \quad k = 3,4$

你应该注意到,4 种标准格式作用句的第一项要素(它相当于语法的主语)的符号有所

不同,有的用 A,有的用 B。后者表示:虽然反应者或承受者是语法的主语,但它们是原作用的对象。我希望用这种表示方式区分所谓施事和受事的概念。

作用句的要素约束条件是:

(A,B)　　优先于(w,p)
X1C　　优先于(vg0-0,g,r,w)
X2C　　优先于 Clause(子句)
XkC　　优先于 vg9-0-0(人类活动)

效应有 11 个节点,从内涵说,应有 11 种效应句,但从形式看,分为 3 种比较适当,分别命名为:

一般效应句　　YB + Y + YC
两对象效应句　　YB1 + Y + YB2
效应说明句　　YB + Y
　　Y + YB
　　Y + YC

不同的效应节点优先于不同形式的效应句。

例句 2,3 的子句分属一般作用句,但都不是标准格式。前一种由逻辑符'由'字,后一种由高层表示 v0-0↑给出了非标准格式的要素指示。

问 17:

先生曾反复强调,作用与效应是'你中有我,我中有你',但又强调它们的区别。我对这一点始终体会不透。从句类分析来说,承受句与反应句以 B 为主语,而作用效应句以 A 为主语,如果将前者划归效应句,不是更为简明么?

答:

你提出的分类方案当然可以选择,如果比较两种方案的优劣,很难抉择。

承受和反应本来就是效应的形式之一,但这两种效应的对偶特征不如效应网络节点那样鲜明。作用—承受—反应又是一个天然的普遍顺序,这就是我的依据。

至于作用效应句,本来就是混合型句类,放在作用句中。只是由于这种混合型句类比较普遍,而重点在于表达效应。

比较一下反应句和作用效应句的相似性和差异性倒十分有趣。两者的完整形式都是四要素句,但前者可简化为两要素句 X2B + X2,后者则不能。这是就形式而言。从内涵来看,反应句的主语是受事,作用效应句的主语是施事。看下面的句例就一清二楚:

张先生怕李小姐生气
张先生劝李小姐学英语

如果熟悉层次网络符号,"怕"∈v7-1-1-,"劝"∈v9-2-3↑v3-0-9,可立即判定前者是反应句,后者是作用效应句,从而可知前句的张先生是受事,而后句的张先生是施事。你看,复杂的施

事受事语言现象,在层次网络符号里不就变成如此简明的规则性知识了吗?

问 18:

先生刚才举的两个句例都是语法上的兼语句。反应句和作用效应句是兼语句的两种类型。用兼语句模型来分析,"李小姐"都属于宾语。从语义的揭示来说,不如层次网络的要素分析,但如果把第一句话改成"张先生怕李小姐的脾气",语法分析没有什么困难,但层次网络符号就发生困难了。张先生所"怕"者,究竟是李小姐这个人,还是李小姐的脾气?如何在李小姐与脾气之间划分 X2A 和 X2C?

答:

在转移句里,我谈到过对象与内容的混淆,最后,我们把转移的物和信息统称为转移的内容 TC。反应句中的 X2A 和 X2C 可以都是引起反应的因素,但 X2C 又可以是反应本身 X2 的内容。另外,反应句还有三种简化形式和一种并合形式:

X2B + X2

X2B + X2 + X2A

X2B + X2 + X2C

X2B + X2 + X2A + (的) + X2C

这些深层的语义结构当然不是统一的标准格式所能表述的,但它终究为这些变化形式提供了一个思考的基础。

这里又涉及"的"字的运用。让我们再回避一次,到适当的时机再谈。

问 19:

先生在谈到效应句时,没有提及效应要素的约束条件,是不是也属于回避?

答:

不是!效应要素的约束条件依赖于所采用的效应节点,给出一般表述意义不大。

我在〈问答 12〉中提供了节点 3-5 相关矩阵的样板,它是要素约束的第二层知识,第三层知识则要通过结构符号对具体的词予以表达。第一层知识表现为句类的标准格式和要素的 A、B、C 之分。层次越高的知识,模糊性必然越大。不过,要素的 A、B、C 正好可作为理解的 ABC,这一点,确实也就是我采用这个符号的本意。

问 20:

理论模型通常有两个着眼点,一是揭示具体的规律,二是为规律或真理的揭示开辟道路,层次网络理论是二者并重。这样的说法符合先生的本意否?

答:

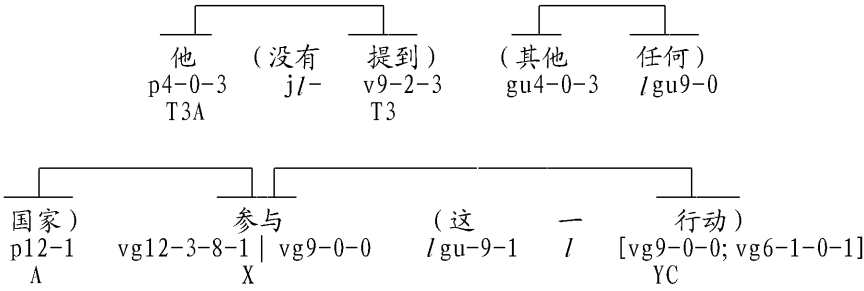
愧不敢当!

在句例 2,3 的分析中,还有一些遗留的问题,如多义词‘决定’的多义选择问题,“将”字的逻辑意义似乎超出了〈问答 14〉中给出的义项范围问题。我希望你自己思考一下。

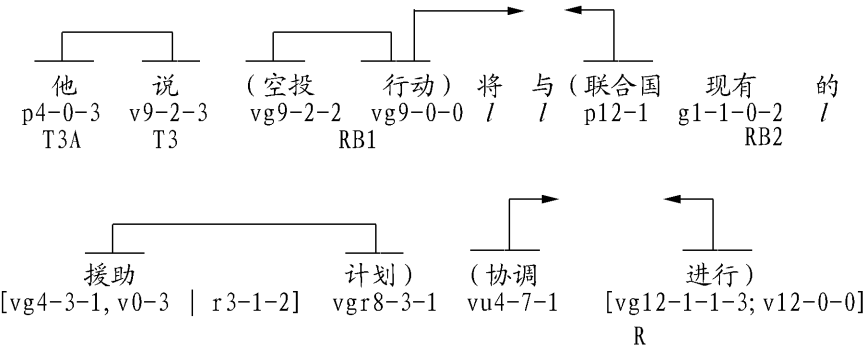
下面继续例句 3 的下文:

(4)

(5)



(6)



这三个例句都涉及关系,其中两个就是关系句。所以,需要对关系句作一个简要的回顾。关系网络有 8 个二级节点,从内涵说,有 8 种关系句。但关系的双方具有特殊的约束性。从形式说,只有一种关系句。其标准格式为:

RB1 + R + RB2 + (RC)

关系双方 RB1,RB2 的特殊约束性表现在以下三个方面:

1. 对应性

它表现了 RB1 和 RB2 的相互约束。关系的双方主要涉及概念矩阵的 p、w 两列(主要二

字显然是简化策略)。不过,这里的 p 应包括人类的活动 vg9-0-0。不同关系节点的双方存在如下表所示的优先性:

节点编号及命名	关系双方优先搭配
4-1 结合与分离	p-p,w-w
4-2 依存与排斥	p-p,w-w
4-3 支持与反对	p-p
4-4 主宰与从属	p-p
4-5 使用与舍弃	p-w,p-g12-6,7,8
4-6 拥有与失去	p-w,p-g12-6,7
4-7 适应与干扰	p-p,w-w,p-w

这个表所提供的知识应视为 1.5 层知识。完整的第二层要素约束知识由关系二级节点的关系矩阵表达。

2. 双向性与单向性

双向性也可称为相互性。呈现双向性的节点有 4-1,4-2;呈现单向性的节点有 4-4,4-5,4-6。其他节点则介于两者之间。我以前在讲课时,在单、双向性之前加了本质二字,这也是一种简化策略。

双向性关系句的标准格式可变为:

$$(RB1) + (RB2) + R$$

“张先生、李小姐结婚了”;“母子相依为命”就是典型的例句。

3. 主动性与被动性

这一属性表现在关系的 3 级节点。属于被动性的 3 级节点有 4-4-2(从属),4-6-2(失去);属于主动性 3 级节点有 4-3-1(支持),4-3-2(反对),4-4-1(主宰),4-5-1(使用),4-6-1(拥有),其他介于两者之间。

主动性关系节点与作用存在交式关联,被动性关系节点与效应存在交式关联。关系句往往难以与作用句或效应句严格区分,原因就在于交式关联性。而关于主动性和被动性的 1.0 层知识应有助于处理这种句类模糊。

问 21:

关系是一个很广泛的概念。据我所知,有些国内学者总结出来了几十种关系,并试图以此作为语言处理的理论基础。先生对此有何评论?

答:

语言理解处理缺乏坚实的理论基础,这是大家共同的感受。为摆脱这一困境,人们不得不走上创新之路,这是十分可喜的现象。

关系的概念虽然很广泛,但从语言的角度来看,主要是 3 类关系:

1. 逻辑性关系。

2. 概念之间的组合结构关系。

3. 作为作用效应链条的环节之一的关系。

我赋予关系网络的内容限于上列第 3 类关系。我坚持语言的理解必须有 7 个基本角度或基点,即 6 个作用效应链条加判断推理。“万物为一”是一个十分卓越的哲学观点,但一定要配合层次网络分析的方法,否则就会陷入大简化的泥潭。

语言理解模型要吸取两方面的教训。一是所谓“有限领域,有限词汇”,二是“无所不包”,都是急于求成的表现。科学的本质是永恒的探索,“欲速则不达”的哲理是探索之路的明灯。我常常想,如果乔姆斯基和菲尔墨当年晚一点发表他们的著名论文,他们对语言学的贡献可能要大得多。

我的意思不是抛弃有限和无限的方法,语言理解模型需要把两者结合起来。层次网络理论模型是一种弹性结构,从符号设计到网络节点的设计都充分体现了弹性,便于模型的补充甚至修改,这是它的无限性。另一方面,作用效应链的不同环节又是局部。我期望着通过后起之秀的共同努力,逐步完成各局部环节的专家系统。这里列出有关课题的清单,作为引玉之砖:

1. 逻辑词辨识系统
2. 语义块及新词辨识系统
3. 条件、手段、目的成分辨识系统
4. 作用句专家系统
5. 过程句专家系统
6. 转移句专家系统
7. 效应句专家系统
8. 关系句专家系统
9. 状态句专家系统
10. 判断推理句专家系统
11. 语境生成系统
12. 主题判定系统
13. 隐知识揭示系统

这 13 个题目是解模糊及纠错处理的基础。13 是一个不祥的数字(按西方习俗),但我不想由于这个缘故而增加或删去一个题目。按语法观点,理所当然应有句群关系分析系统,但这个题目不宜与上列 13 项并列。

7 个句类分析的题目我都用了“专家系统”的名称,这不是赶时髦,而是由于这 7 个题目是计算机容易发挥其长处,从而达到人类专家水平的领域。其他题目反而更难一些。

但是,这 7 个专家系统密切依赖于前 3 个系统的成果。我们的谈话按先易后难的步调,相继涉及到转移句、作用句、效应句,现在正在谈论关系句。关系句我只谈到一种标准格式,你不觉得有疑点吗?

问 22：

疑点不仅是关系句的标准格式，还有关于关系双方特殊约束性的提法。

我们非常熟悉的一些语句，例如“团结全国人民”，“全世界无产者联合起来”，“张先生是李小姐的男朋友”等等，从内容说都是符合关系句的条件，然而，却不符合关系句的标准格式。

答：

在〈概述 3 章〉（指〈层次网络理论概述〉第 3 章，以下都采用这种简化表示）中有这么几句话：“3-8 的‘合与分’与关系节点‘结合与分离’4-1 存在交式关联性，所以，由 3-8 节点构成的效应句和由 4-1 节点构成的关系句在形式上有相似之处，但实际的语句仍不难区分”。你举的第二个例句是典型的效应说明句 YB + Y。虽然“联合”∈［vg9-3-8-1，vg9-4-1-1］，但它后面跟着的“起来”，却是效应的标志。因而，此句的句类可唯一确定。

“起来”是一个多义词，其逻辑义项在词典中有多达 100 余字的解释。用层次网络符号可表示为：

起来 ［lvuh，3-0；5-0］

你可以对照词典，检验一下这一符号映射的优点和不足。

“团结”与“联合”意义大体相同，在英语中就是一个词 unite。不过，汉语的“团结”更强调互相支持的意义。如果表达得精密一些，可写成：

团结 ［vg9-3-8-1，lv0-0-4-0｜vgu4-3-1］

“团结全国人民”这句话，没有其他的要素标志，分类会出现模糊。但这种模糊是一种两可性模糊，只要引入两可的概念就可以解决。

至于第三句话是典型的判断句，以后再谈。

问 22 续 1：

两可处理是一个有趣的想法，但如何保证两可处理的结果一致？另外，为什么不把这种“两可句类”另外命名？例如，称之为效应关系句？

答：

“结果一致”的提法对于语言处理可能不太恰当，人们太习惯于确定性处理的思考方式了。句类是语言表达的不同角度，有的问题，例如这里的“分合”问题，可以从效应和关系两个不同角度去进行表达。“团结”作为效应，有效应的双方，其作用者和对象都一定是人；作为关系，有关系的双方，是 p-p 对应。从这个意义上说，两可处理的结果是一致的。但从句类分析的模式来说，效应说明句不要求找出有关的 XB，而关系句则必须找出关系的双方，这又有所不同了。

7 个基本句类的交叉混合，有 $7 \times 6 = 42$ 种混合句类。混合句类是指一个句子里包含了至少两个基本句类。至于一个词兼有交式关联节点意义的情况，应与我所定义的混合句类

相区别。是否对这种情况取一个名字是另外一个问题了。(注:这里说的混合句类,后来正式定义为复合句类,而当一个兼有作用效应链两个环节含义的动词充当语句的 E 要素时,则命名为混合句类。)

问 22 续 2:

先生谈到关系网络的三点特殊约束:对应性、相互性、主动性。然而,这些属性不是关系节点所特有,作用就有主动性和相互性,转移也有相互性,对应性则更为普遍,EA 与 EB 总存在某种对应性。因此,这三项约束作为一种规则性或启发性知识来运用,似乎困难很大。

答:

这确实有些困难,它涉及有关逻辑符号的设计和表达问题。就汉语来说,主要是(关于,对于,相,互)等几个虚词的层次符号表达问题。先把它们的映射符号写在下面:

关于	lv1-0-4-0
对于	l0-2-4-0
互,相	lv0-0-4-0

这里,数字层次的最后都没有延续符"-"。表示它们都是关系网络 0 分行的高层概念。每个层次网络都设有 0 分行,曾将 0 分行戏称为"不管部"。这个戏称实际上深刻表述了 0 分行的功能,凡 0 分行的高层概念,不仅与本网络的其他分行,也与其他网络存在着难以具体表述的交式关联。这是一项符号的约定。

这里顺便说一下层次网络理论的符号设计。由于逻辑概念和基本概念原来放在抽象概念的 14、15 行,现在改为 l、j 类,原 w 类现在一分为三,变为 w(物)、p(人)、x(物性)三类,符号体系有较大变化,符号设计的优化比较复杂,要考虑工程因素。就后者来说,我依然倾向以半字节为单元的方案,这不仅可节省一半存储量,也符合推理运算的精度需要。

半字节单元表示就要求对层次网络符号作出适当的分类。王华根据原来的符号约定作过一次总体设计(参见他的硕士论文"通用汉语知识库的理论模型、设计及实现",中国科学院声学所,1992.7),这里根据符号约定的变化,提出下列建议:

1. 将半字节(即 16 进制)的 e、f 两个数定为特殊符号。e 用于表示一类符号的结束,f 用于类别指示。

2. 这样,数字值域限制在 0 — d,这个限制实际上是对于概念矩阵实际维数的利用。

3. 概念矩阵为 5 列,其字母命名为 v,g,u,z,r。

4. 结构符号分述如下:

a) 并与选的符号",;",它们是逻辑概念 l4-0-3-8-1,l4-0-3-7-0 的符号表示,使用频度极高。它们一定是两义项的间隔符,因此,似乎可采用特殊约定,例如用 f0,f1 表示,省去结束符。

b) 4 个特殊的组合结构符号"↑,↓,→,|":它们可以相互连用,但意义有所不同。↑与↓连用仅表示该组合概念不突出作用或效应,后随的义项不论多少,都可

视为一个。其他的连用则各有相应的内容,例如△△↓→(×,×,×△)表示:△△的对象是×△,效应×。

这4个符号右侧的具体内容在词义组合规则中,多数用二级规则序号表示。到了形成理解文本时,采用上面的形式,还原出具体的内容。

c) 属性指定符"/":这个符号的具体含义最为模糊。二级规则序号在理解文本中是否必须还原出具体内容,可暂置不论。

d) 括号:在书写形式中有3种"{}()[]|",分别表示组合块、并、选、交、逻辑概念介入。但在机内表示,可并为1种。

e) 本源语义及次源语义符号"∮,∫"。这两个符号仅用于汉语的字义表示。在理解文本中视具体情况或还原出相应的内容或省去不管。

这样,结构符号总计为10个,还有4个余量,供机动之用。

理解文本的符号体系设计,这里不来讨论。这个问题涉及到许多我完全不熟悉的领域,不能班门弄斧。在适当的时候,应专门研讨一番。

问22续3:

并的符号",",似乎可以省略,因为,任何概念都以数字串结束,因而,这个结束符号,自然就是概念单元的间隔标志。符号","一定与括号连用,它的意义与概念间隔符没有质的区别。先生说过,两概念之交记为(XY),中间不插入间隔符",",但交的使用频度远远低于并,何必不交换符号约定?

答:

概念单元不一定以数字串结束,也可以是字母串,所以,我倾向于给"并"一个特殊的明确符号。但你的看法是完全正确的,这是一个两可问题,由你们自行决定。

问23:

结构主义语言学将语言分为综合语和分析语两类。汉语是典型的分析语,俄语是典型的综合语。对于以词序和虚词为基本语法手段的分析语,句类标准格式的提法才有意义。这样,层次网络理论所建立的句类分析程序对于综合语或分析综合兼备的语言(如英语)是否不能适用?

答:

前述各句类的标准格式是以汉语的表达习惯来表述的。不同语种应选取不同的标准格式。例如日语,特征要素就放在最后,这是日语的标准。在这个问题上,不能也没有必要去追求合理的或最佳的标准格式。不同语种之间标准的转换是不可避免的麻烦。

为了实现不同标准的转换,就需要有统一的层次网络符号。在〈概述14章〉曾经提到,综合语的形态变化也是一种逻辑符号,形态的本来意义就是如此。只要形态和虚词用统一

的逻辑符号表达,句类分析方法自然是相通的。至于具体的分析程序又当别论。这是程序模块的通用性(适合于不同语种)问题。分析方法的同一性和分析程序的通用性不能混为一谈。

这里顺便说一点历史。先于结构主义的历史比较语言学曾将语言按形态变化的发达程度分为孤立型、附着型、屈折型 3 类,并认为它们三者代表语言的三个进化阶段。孤立型的代表语言——汉语,按这一观点,属于原始的落后语言。这种观点是典型的西方自我中心主义,在学术上是幼稚无知的表现。

问 24:

仿照先生对例句 1 的分析步骤,依据词的层次网络符号,运用 lv 准则和句类标准格式,不难对例句 4 6 作出相应的分析,并得到下列结果:

J4 = RJ = RB1 + R + RB2

J5 = T3J = T3A + T3 + (T3C)

T3C = YJ = YB + Y + YC

J6 = T3J = T3A + T3 + (T3C)

T3C = RJ = RB1 + l + RB2 + R

我刚才用了"不难"二字,这是从"蒙"的意义来说的,蒙起来不难,实际是很难的。蒙的成分包括:

1)"空投行动中"的"中"字,置之未予分析,是蒙的。

2)"参与"一词并未指定后面跟 YC 而不是 YB,也是蒙的。

3)"计划"一词属于 vgr 型,现与"援助"合并,而不与"协调"合作,当然也是蒙的。

4)还有两大蒙,是"的"字的语法功能甚多,这里蒙作语义块并合符;二是高层概念"进行"的运用,方式甚多,这里将它与"协调"合并,也是一大蒙。

5)最后,lv 准则的 l 还没有确切的说明,到底哪些 l 是语义块分隔符,哪些不是,我现在还不得要领。

答:

对于"蒙"的提法我很感兴趣。"蒙(meng1)"的词典释义是"胡乱猜测"。如果把"胡乱"二字稍稍淡化一些,那么,可以说,理解处理的入门阶段就是也只能是蒙处理。知道蒙比起不知道是一个质的飞跃。

层次网络理论与其他知识表示方案的最大区别之一就在于,它为蒙处理提供了各种知识。让我们来分析一下你刚才说到的各项蒙处理。

1)关于"空投行动中"的"蒙"

由"使用"∈ vg4-5-1	该句为关系句
由 R∈4-5	RB1,RB2 优先于 p-w 或 p-g12-6,7,8
由 飞机∈ w10-	满足 p-w 约束右要求
由 空投∈ vg9-2-3	满足 p-w 约束左要求

行动∈vg9-0-0

于是,疑点仅在于“中”字。我在〈问答 14〉中说“语言最大个性是它的语种性,具体表现就是语法,脱离了语法,语言理解几乎不可能”。这是一个很好的例子。这个句子的标准格式可以有以下 5 种语法形式:

我们在空投行动中将不使用……

在空投行动中我们将不使用……

在空投行动中将不使用……

空投行动中将不使用……

空投行动将不使用……

后 3 种语法形式是汉语的特有现象。在关系对应性的说明中有这么一句话“p 应包含人类活动 vg9-0-0”。这就是说,作为关系的一方,人与人类活动可以是等价的。这是汉语的语法观,也是上列最后一种语法形式的理论依据。另一方面,把人类活动作为人在关系活动中的条件,也是理所当然的。这是语法形式 1-4 的理论依据。西语恪守后一种语法观。这个例子是汉语语法观的辩证性和西语语法观的形而上学性的生动表现。

从语言表达所提供的信息来说,“我们”本身并不带来任何信息,它的实际内容由前面的语句提供,而其形式存在则由本句的“空投行动”暗含,所以,从信息表达来说,“我们”二字是冗余成分。既是冗余,就可省略,这是汉语常用的语法手段。

替换及省略不限于关系句,将来我们还会遇到更有趣的句例。

所以,第一蒙属于两可处理,将“空投行动中”视为条件成分,或关系的 p 方皆可。前者应用汉语的省略原则,后者应用汉语的替换原则。

2)关于“参与”的“蒙”

你没有注意到“参与”的层次网络符号 vg12-3-8-1|vg9-0-0 的全部含义?结构符号|之后的高层表示要求(这是约定):补充 vg9-0-0 的具体内容,而不是对象,因此,这里完全没有“蒙”的色彩。

3)关于“计划”的“蒙”

“计划”的层次网络符号是 vgr8-3-1-。与 8-3 链式关联的节点是 9-0-0,与 8-3-1 链式关联的节点是 1-1-0,与 8-3-2 链式关联的节点是 5-4-0,与 8-3-1、8-3-2 分别交式关联的节点是 jg1,jg2。根据这些关联性知识以及 1-1 与 jg1 的天然强关联性,运用优先搭配原则可知,“计划”与“援助”(同行优先),“计划”与“现有”(交式关联)具有优先搭配权,有理由将“现有”、“援助”和“计划”三者搭配起来,构成一个语义块(这里还有“的”字的补充根据),再利用 lv 准则同前面的“联合国”搭配起来组成更大的语义块,这个子句是关系句的结论就容易作出了。因为最后一个词是高层概念,它必须有具体的 v 概念予以补充,最靠近的“协调”与它搭配乃是同行优先的又一次运用。由此可见,人们曾视为汉语理解一大难关的动词连用现象(如这里的从“援助”到“进行”),只要运用概念关联性知识和汉语特有的高层概念运用规则(我们并入同行优先准则),再加上句类分析,是不难解决的。对于这个子句来说,假定它是

关系句以后,要素约束条件又完全吻合(p-p 约束),因而,所谓确定中心谓词的困难已经迎刃而解了。不仅如此,作为一个句类分析专家系统,它还不难作出判断:(1)“计划”一词可用“活动”一词代替,也可省去。(2)“进行”一词可换到“协调”之前,同样也可省去。因为,它们都是高层表示。

你觉得这样的分析过程存在陷阱么?

问 24 续:

汉语理解的核心难点就这样被突破了,不免心有“余悸”。我相信从概念关联性约束和句类分析可以解决一些问题,但我同时也相信还存在仅仅依赖这两种武器不能解决的复杂情况,不过一时想不起实际的句例。

答:

我理解你的心情。不过,也不必挖空心思去想什么“鸡不吃了”之类的歧义句,以后碰到实际句例时再继续这一讨论。

关于关系句,还应就它的标准格式补充几句话。前面对转移句、作用句和效应句的要素都引入了内容项,但关系句未谈到 RC。实际上,两种标准格式的关系句都可以加上 RC:

RB1 + R + RB2 + (RC)

RB1 + RB2 + R + (RC)

在 RB2 和 RC 之间还可以插入“的”字。

关于哪些 1 概念可以作为语义块切分标志,哪些不能,在〈概述 14 章〉已给出了足够的说明,你最好自己总结一下。

问 25:

〈概述 14 章〉最后说“独立逻辑概念 j1 留待词汇级有关概念收集比较齐全以后,再着手综合研究”。当前的工作急需这方面的知识,不知先生可否现在就谈一下这个问题?

答:

对这个问题的思考还很不成熟。逻辑按其本来意义是研究思维规律的科学。在〈概述 14 章〉开头说过“拟在逻辑层次网络里包含……反映思维运作的基本概念,并把这类概念列为 j1”。

另一方面,为人类的思维活动已专门设置了层次网络 8,那么,是否有必要另行设置 j1 概念节点,这确实值得思考。所以,这里先谈一下思维网络的设计要点。

思维网络分为 5 个二级节点:

8-0 一般思维

8-1 认识与理解

8-2 探索与发现

8-3　设计与规划

8-4　评价与决策

从标题来看,思维网络二级节点的设置与其他网络相比,没有什么特异之处。从设计思想来说,8-1 及 8-2 表达思维(主观)对事物(客观)的反映;8-3 及 8-4 表达思维的能动性,即主观对客观的反作用。这样,从表达的角度大体上满足了完备性的要求。但是,一般网络的类与层次是两个可分离的变量(这个概念从数理方程借用而来,非常切合我们当前想说明的问题),而思维网络则是不可分离的。思维学对思维的内涵、方式及规律都有分类,例如抽象、形象、灵感及创造性的 4 级分类法,但层次性本身都是分类的主要依据之一。可见,思维概念的类与层次是不可分离的。相反,如果你回忆一下作用、效应、转移及关系的二级节点配置,你可以明确无误地感受到,这些网络的类与层次是可分离的,同层次的网络节点具有层次同一性,虽然相互之间可存在交式或链式关联。

我曾一直试图对思维网络作出类与层次的分离,然屡经碰壁,最后才认识到两者的不可分离性。这一现象的深层意义,则未予深究。

回到思维的二级节点,8-1 与 8-2 就属于不同层次,但两者又存在质的区别,不宜并为一个二级节点。就层次而言,8-2 深于 8-1(请注意,我对于层次的序分别采用高低和深浅两种说法,高与浅对应,低与深对应,与习惯用法有所不同),大体上 8-1 相应于判断,8-2 相应于推理。

抽象的谈论已经太长,下面转入具体的说明。

8-0 的基元概念是"想"。古汉语的"思",现代汉语的"思考",都可以作为 vg8-0-0 的汉语映射,许多 8-0-0 概念可由基元概念与 8-0-0 组合而成。如:

回顾　[vg8-0-0, lv9-0-4-0, g1-1-0-1]

展望　[vg8-0-0, lv9-0-4-0, g1-1-0-3]

预料　[vg8-0-0, lv9-0-4-0, g5-3]

猜测　[vg8-0-0, lv9-0-4-0, g3-2-2]

想像　[vg8-0-0, v3-6-1]

我们看到,以逻辑形式(vg8-0-0,lv9-0-4-0,gi-k-m)组成的 vg8-0-0 概念为数不少。这种逻辑表达是 3 号语义结构方程的实际形式(由二级规则给定)之一。下面将谈到,这种表达方式可以简化。"想像"是另一种组合形式,它相应于 1 号结构方程。可以考虑将它定义为 vg8-0-1,因为它是所谓形象思维的基元,但这一点当前不应作出定论。

"想"是 8-0 的 vg 型基元,g 型基元是"思维",而"概念"本身则可定义为 r8-0-0,"抽象"可定义为 vru8-0-0,其对偶句"具体"的符号下面再谈。这里,顺便说一个小插曲。我引入 r 类概念的刺激点是两个,一是关于作用效应对的想法,二就是为了"概念"本身的表达。

如果要对 8-0,8-1,8-2 给出某种界定的话,那么,可以说 8-0 是关于"一般事物"(g12-0-0)的思考,8-1 是关于"问题"的思考,8-2 是关于"规律"的思考。于是,我们要进一步回答,"问题"与"规律"如何用层次网络符号来表达?"问题"是一种隐事物或事物的隐成分,可映

射为 g12-0-0/g3-2-2。“规律”则有其两方面的含义,动的方面指事物过程的趋向与转化,静的方面指事物之间相互联系,因而,可映射为 [vg1-3,vg4-0]。由于这两个组合概念是 8-1 与 8-2 的思考对象,我们可以将它们定义为 g8-1-0 和 g8-2-0。实际上这种定义方式对 g2-3-0(信息)已采用过。如果你查对语言或哲学词典,你会看到这里关于“问题”和“规律”的定义与经典定义有所不同,但我不想对这种差异性发表看法,因为,层次网络理论着重于实用。

给出上述说明以后,你可能立即想到 8-1 的 vg 基元是“判断”,而 8-2 的 vg 基元是“推理”。然而正是这个似乎是水到渠成的推论使我困扰甚久。如果将“判断”定义为 vg8-1-0,那“比较”如何定义?一切判断和推理的基础是比较,“比较”才是高级思维的运作基元。请回想〈概述 14 章〉引言中的一句话:“反映思维运作的基本概念”。最后我是把“比较”的概念列入 j1。现在,你对〈概述〉14.1.3 节的下列几句话应有所体会了:“独立逻辑概念与思维层次网络(8 行)强关联,但思维层次网络的总体设计方案仍有待完善,所以,本节的讨论留待以后进行较为适当”。

遗憾的是,写下上面的话以后半年来,始终没有腾出时间对思维层次网络的总体设计再作深入思考。这里的问题在于:思维网络能否按照 l 网络的样板采用如下的双重结构形式:

(思维层)(基元层)

[(思维层),(基元层)]

这包括思维层本身的层数约定为 3。这个约定与〈概述 14 章〉的约定完全一致,因为那里的字母 l 相当于一个概念层次。

这个问题暂时也不作定论。如果采纳这个方案,则前面关于(回顾、展望、预料、猜测)的符号映射就可以简化为:

回顾　vg8-0-x-1-1-0-1

展望　vg8-0-x-1-1-0-3

预料　vg8-0-x-5-3

猜测　vg8-0-x-3-2-2

利用这种结构,还可以表达诸如下面的概念:

前提　g8-2-x-1-2-1

假设　vg8-2-x-1-2-1

检验　vg8-2-x-1-2-2

怀疑　[vg8-0-x-1-2,j1v1-2]

证实　[vr8-2-x-1-1-4,j1v1-1]

我之所以对思维网络是否采用双重符号结构犹豫不决,主要是因为感到,对思维概念的总体把握还比较欠缺。思维概念没有相应的词典,不像逻辑概念有虚词词典可供参考。不过,这种谨慎也许是多余的。我们可采用而不限于双重结构,因为我们还有组合结构符号和语义结构方程两项法宝,而且,复杂概念的表达可以不要求唯一的符号形式,语言本身正是如此。当然,程序如何处理符号的不唯一性或多样性是一个难题。下面举几个例子来说明这种不唯一

性和‘法宝’的妙用。

分析　vg8-1-0↓vg3-8-2

综合　vg8-1-0↓vg3-8-1

但也可以采用上述双重结构来表达：

分析　vg8-1-x-3-8-2

综合　vg8-1-x-3-8-1

再看‘法宝’的运用实例：

演绎　(vg8-2-0,/g12-7-1)↓(lv8-0-1-2-1|jug7-0-2-1,v3-9-1|jug7-0-2-2)

归纳　(vg8-2-0,/g12-7-1)↓(lv8-0-1-2-1|jug7-0-2-2,v3-9-1|jug7-0-2-1)

把这些符号映射为汉语就是“演绎是根据一般性求得特殊性的推理方法,归纳是根据特殊性求得一般性的推理方法”。

我希望这个例子有助于说明 如果没有引入结构符号↓和|,复杂概念的表达、概念关联性的揭示都是很困难的。反过来说,运用这些符号,任何概念表达的难题似乎都可以迎刃而解。在4个特殊结构符号中,只有→同语法有些联系,另外3个与语法无关。这一点并不奇怪,因为语法的着眼点本来就不是概念内涵,而是概念形式的关联性。当然,应该声明,这个说法对于语法学的开山鼻祖亚里斯多德并不适用,这位祖师爷在语言上的思想方法倒是与我国的传统语言学非常一致。

问26：

关于双重结构的想法,先生在讲课时谈过多次,在〈概述10章〉也有所说明,在〈概述14章〉给出了系统的约定。下面为了表达方便,不妨把第一层称作本体层,第二层仍用先生的命名—基元层。双重结构的关键问题是如何表达两层之间的相互关系,先生在〈概述14章〉并没有解决这个问题。在上面的示例中,“回顾”一组和‘前提’一组的两层关系就有所不同。回顾是关于过去的思考,而前提则是推理的依据。如果把回顾映射到思考的过去,前提映射为关于依据的推理,显然不对。

答：

你提出的问题极为重要。本体层与基元层的关系肯定要加以说明或约定。不过,关系离不开‘↑’,‘↓’,‘→’,‘|’,‘()’,‘[]’和‘/’这7种。我的想法是,如果把采用双重结构表达的关系限制在‘|’这一种,就不需要另加说明。

你刚才举的例子十分有趣,似乎引入了关系‘/’。我在〈问答22〉中说过“这个符号的含义最为模糊”,因此我认为它特别不宜于引入到双重结构。在我看来,“关于过去的思考”与‘思考过去’是等价的,“关于依据的推理”与“推理依据”也是等价的,它们的结构都是：

(本体层)(基元层)

你可能奇怪,“推理依据”明明是歧义的,它可以是动宾结构或定中结构,怎么这两种结构竟可以等价视之？问题在于‘前文’的约束,g要求定中结构,vg要求动宾结构,但这个‘中’和‘宾’

又统一在内容之中。

当然,我绝不反对在本体层与基元层之间引入其他的结构关系,因此,似乎应考虑预备一个关系说明层。这是一个工程问题,你们讨论商定。我的看法倾向于“有备无患”。

基于上述,实际上我不支持对“分析综合”这样的概念采用双重结构,采用符号“↓”更为简明。前面在示例中的两层之间加了“x”号,就隐含着关系说明层的意思。

问 26 续:

先生关于动宾结构与定中结构等效的说法,人理解起来都大费周折,何况计算机?所以,还是加上说明层为妥。

答:

理解受到语言习惯的制约。我倒不认为,人难以理解的东西,计算机也一定难以理解。层次网络符号所提供的信息与自然语言符号差异甚大,各语种之间也有差异。像我,已很习惯于用层次网络符号来思考,可惜年龄和身体状况已不允许我去充当计算机的教师了。但在这个英才辈出的年代,尽管“海潮”澎湃,总会出现“识智勇”兼备者去接过这根接力棒的。

下面继续思维网络与 j1 的讨论。

前面对 8-1-0,8-2-0 分别给出了定义。依据惯例,你应该已经想到,8-1-1,8-1-2,8-2-1,8-2-2 的定义就是前列的标题,即:

vg8-1-1	认识	vg8-2-1	探索
vg8-1-2	理解	vg8-2-2	创造

在这里,又一次看到类与层次的不可分离性,认识与理解是不同层次的判断,探索与创造是不同层次的推理。换一个角度来说,对于思维概念,扩展层次不是增加信息的有效途径,这同信号处理中正交化的概念是一脉相承的。因此,在思维网络内部,v、u、g、z、r 之间的对应也限制在二级,三级以后自成体系。

与 8-1 相对应的 u、r 概念有:

u8-1- 肤浅、深刻

r8-1- 答案、看法、意见

与 8-2 相对应的 u、r 概念有:

u8-2- 严谨、严密

r8-2- 学说、理论

表达思维能动性的两个节点 8-3,8-4,一方面与 8-1,8-2 强关联,另一方面又与人类智能活动(广义的,不限于 9 行)强关联,这是容易理解的。正是由于这个缘故,专用于这两个节点的词汇级基元概念主要是 r 型概念,如 r8-3 的策略、方案,r8-4 的结论。作为这两个节点标题的概念,设计 8-3-1,规划 8-3-2,评价 8-4-1,决策 8-4-2,都具有这两方面的关联特性,一是它们都属于 vgr 型概念,二是它们也代表人类的一项智能活动。在例句 6 中,对“计划”一词的讨论,就利用了这两项特点。

下面转入 jl 的讨论。

在〈概述 14 章〉讨论一般逻辑概念 l 时,是以 l 与概念基元(即作用效应链)的关联性作为 l 网络设计的依据的。同样,jl 网络的设计,应以它与思维概念的关联性为依据。

思维运作(活动)有其基本方式和基本效应(结果),对于这两者的表达,是 jl 网络的使命。

前已指出:思维运作的基本方式是比较,而思维的基本效应是肯定与否定。所以,jl 网络的一级节点可分配如下:

jl0　比较

jl1　肯定与否定

现在就来对这两个一级节点作矩阵分析。jl0 的二级节点可分为三类:

jl0-0　基本比较

jl0-1　集比较

jl0-2　标准比较

基本比较是两相比较,它显然是一切比较的基础。集比较是指一个集合内的相互比较。标准比较是指多对一的比较,当"多"不受限制时,"一"就具有"绝对"的意义。

以上是行分析,下面进行列分析。先看一些有启发性的例子。

1)

jlvg0-0　比较

jlvg0-0-8　参照

jlvr0-2-1　符合

jlvz0-0-1　接近

2)

jlru0-0-1　同

jlru0-0-2　异

jlru0-0-3　反

jlru0-1-1　一致

3)

jlg0-2　标准

4)

jlug0-1-1　共(性)

jlug0-1-2　个(性)

jlug0-1-0　相对(性)

jlug0-2-0　绝对(性)

5)

jlz0-0　差距

6)

[jlguu0-1,jz0-0]　最

jlu0-1-1/jg4-0—0　交

jv2-2-12-0-1|jlru0-0-1　(相)似,像

这里,需要对基本概念 j 的 0 分行作一点交代。按层次网络理论的定义,j0 应该是基本概念中的基本概念。什么是 jg 的 jg?我曾打算回避这个"讨厌"的问题,这就是以前定义 j0 为时间的原因。后发现不妥,改时间为 j1,但为了寻找 j0,颇费周折。现在似乎是"找到"了,那就是"序"。

你看,时间的先后,春夏秋冬,年月日时;空间的上下左右,东西南北;数的一二三四,百千万亿;量的多少;质的优劣,精粗;度的欠,够,过;性的好坏,主次,一般与特殊;不都是序的

表现么？作用的作用—承受—反应；过程的过去，现在，未来；过程的预备，开始，持续和结束；过程的源与流，趋向与转化，新旧交替；转移的起点—途径—终点；效应的影响—变化—完成不也是序的表现么？从某种意义上说，序即规律，理解就是对序的把握。所以，把序作为 jg0，应该不至于引起太大的争议。

集比较产生“最”的概念，但“最”是一个复合概念，应与 jz0-0 并合。

“交”的概念在〈概述 14 章〉曾说过，不得不放进 j1 中，这曾使我困扰甚久。但现在这个定义“共同的部分”倒是完全符合“交”的逻辑意义。

思维基本结果的表达比较简单，如下表所示：

第一层	1	判断
第二层	1，2	肯定，否定
第三层	1，2	存在性，非存在性
j1v1-1-1		有
j1v1-1-2		是
j1v1-2-1		无
j1v1-2-2		不（是）
j1v1-2		非，否

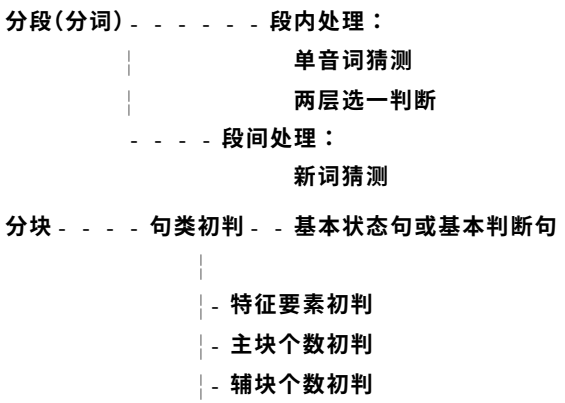
问 27：

关于分段、分词、分块和分类处理的相互关系，先生能否作一个综合性的说明？

答：

分段和分块是一切自然语言处理的起点。分块是分段的目的，分段是分块的手段。在下一个层次来说，分类是分块的目的，分块是分类的手段。再往下一个层次来说，理解是分类的目的，分类是理解的手段。

讨论任何问题都要有一定的背景，理解问题更是如此。所以，我想应该先谈一下理解处理的整体框架。这个框架如下图所示：



句类分析 - - - - - - - - - - 混合类考证

- - 单类考证 - -

子类设定

要素关联性检验(评估)

子类评分

子类初判,分块终判

上下文分析 - - - - 语境更新

子类二判

模糊知识更新

要点、主题分析 - - 要点更新

主题更新

体裁、风格分析

子类终判,分词终判

一个完整的汉语理解系统必须包括框架中的 5 个基本模块 :分段分词(两者也可合称为预处理模块) ;分块及句类初判 ;句类分析 ;上下文分析 ;要点及主题分析。每个基本模块的子模块构成则与文本的性质和具体的使命有关。上图所示代表文本模糊并以解模糊处理为核心的情况。今后将对各基本模块的使命与难点、知识与规则、判断方法及推理策略作必要的说明。本次问答着重讨论 1 号基本模块——分段(分词)。

使命 :为语义块切分作准备。分出单音词和非单音词,把所有可能的 lv 概念“点”给出加权标志。

难点 :(1) 三音双词现象,简称 3—2 现象,指相继的三个音节的前后两个音节都可以组成双字词的现象。在模糊文本时,这个现象非常普遍 。

(2) 假双或假多现象,指词库指示可组合成词的相继音节实际上应该是单字词或新词的情况。

(3) 假单现象,指词库指示不能组合成词而实际上应该是一个新词的情况。

知识 :字库、词库、音节感知库。字库中的字按频度分为 5 级,反映字的无条件频度知识。词库中的词分为 4 级,A、B 级为高频词,C 级为专业词,D 级为低频词。音节感知库提供 :音节独立使用的无条件及条件频度知识,该音节的独立运用义项。

规则 :“前下后上中上下”。对这条规则将在下面详细说明。

分段永远是理解处理的第一步,这我已经讲过多次了。在分段的时候,是仅选用词库中的 A、B 级词,还是全部选用,这显然与语料的体裁有关,也与系统的处理功能的配置有关。如果系统的新词辨认功能较强,选词的标准从严为好,反之从宽为好。可惜现在的词库由于空间的限制,仅分为 4 级,实际上只有两种选择。如果分为 5 级,可有三种选择,就比较理想了。

下面来说明上述的“前下后上中上下”规则，将简称上下规则。这里的“上下”指音段的上层和下层，“前中后”指单音节“词”的位置在音段前面（头）、中间或后面（尾）的情况。这条规则的意思是：

“假设一个音段里只有一个单音“词”，如果它在音段的头，则取音段的下层；如果它在音段的尾，则取音段的上层；如果它在音段的中间，则分别取它所在位置前后的上层和下层。”

音段是模糊文本处理时必须引入的概念，我在〈问答 3〉中已作了说明。由于上述的 3—2 现象，一个音段自然分为上下两层。让我们先来看三个例子：

	拼音	yao qiu wo men si xiang geng jie fang yi dian
(1)	正文	要求 我们 思想 更解放一点
	上层	要求…我们…(思想) …耿介 防疫
	下层	(解放)一点
	拼音	dui dang qian de gai ge he jian she ju you shi fen zhong
(2)		yao de zhi dao zuo yong
	正文	对 当前 的 改革和建设 具有十分重要的指导 作用
	上层	对…当前…的… 改革核减 …具有十分重要得志…作用
	下层	隔阂建设… 优势分钟要得(指导)…
	拼音	jing ji luo hou jiu yao shou zhi yu ren
(3)	正文	经济 落后 就要受制于 人
	上层	经济… 落后…就要(收支) …人
	下层	摇手(至于) …人

例子的“拼音”行代表模糊文本（音调模糊），音段的切分标志用空格或“…”表示。上层或下层文本中的模糊双音词用“（）”表示，并选出该模糊集中的第一个词作为代表。这三个例子给出了分层处理和上下规则的一般模样。下面对上下规则作进一步的说明。

这个规则是对于“现代汉语以双音词为主”和“单音词的语音分布和语义比较集中”这两个基本语音现象的具体运用。前者乃众所周知，后者则属于个人之管见，这里不妨多说几句。

语言的每一个词都有一定的意义。在语法学里，还区分所谓词汇意义和语法意义。这个区分对于汉语其实没有多大意义，比这个区分远为重要的是意义的独立性和非独立性，但语法学很少涉及这个问题。一方面是由于对这个问题的重要性认识不足，另一方面也由于语法学习惯于与语义、语用划清界限，而独立性与语用有密切的联系。语法学引入了“自由运用”的概念，把这个概念作为区分词和语素的基本依据。读者不妨温习一下语法学关于语素、词、短语和句子的定义：

	自由运用	意思	语言单位
语素	不能	不完整	最小
词	能	不完整	最小
短语	能	不完整	最大
句子	能	完整	最小

国内语言界根据这个定义,曾对一般汉字或某些具体汉字是语素还是词进行过一场不得要领的争论。这场争论完全是由于这个定义本身的缺陷所造成的。所谓‘自由运用’,只是独立性的一种极端情况。具体的汉字既有可自由运用的义项,又有不能自由运用的义项,因而实际上往往具有‘词’和‘语素’的双重特性。词和语素之分,源于屈折语的形式特征,没有语言结构层次上的本质区别。有些语法学家不理解这一点,因此,一碰到汉字这个西语里不存在的‘怪物’,就难免有些无所措手足之感了。

从汉字的本源来说,字就是词,其本源义项都是独立的。但随着字义的不断延伸,就出现了独立性较弱或完全不能独立运用的义项,即出现了语素的特征。例如‘年’,其独立义项是:(wjzz1-0,jv4-0,wjzz1-0-0)。非独立义项有:新年、年龄、一生中按年龄划分的阶段、每年、时期、年景、一年的最后一天或其傍晚,相应的词如:拜年、年轮、青年、年会、近年、丰年、年夜、年饭。后面这些意义的‘年’都是语素,只有第一个意义的‘年’才是词。

语素的本质是符号简化。在西语,这个简化的思想主要用于一些最常用的概念,并采用形态变化的手段。而汉语则采用组合的方式予以广泛地运用。不过,现代英语也日益普遍地用组合方式构造新词,例如 IC,DSP,LASER,IBM 等等,这里的每一个大写字母就是一个语素。

因此,语法学的词和语素应视为词的独立性的两种极端表现——完全独立和完全不独立,当然,还有不同程度半独立的中间情况。例如‘春’,其主要义项春天——(xjg1-1,j7-0-4-1)是半独立的,其他义项则是完全不独立的。从独立性的观点来看,关于‘春’字是词还是语素的争论是多余的,而且在语法的框架内必然是没有结果的。

对独立性的描述可采用简单的 4 级表示:A——完全独立;D——完全不独立;B,C——半独立,B 接近 A,C 接近 D。春天的‘春’就是 B 级。

上面我们说明了词的独立性的理论意义,下面来进一步说明这个概念对于分段—分词的实践意义。众所周知,汉语的每一个音节都有意义。一般说来,如果把该音节的全部汉字及其意义都考虑进去,则每个音节的意义集合的模糊度太大。但是,如果只取字义中的独立义项,模糊度就会大大减小。如果再进一步略去频度较低的字,那么,一个音节就会凸现出便于联想的有限数量甚至是唯一的意义。这些凸现出来的音节意义正是分段处理后最需要的信息。

分段处理后的段内处理,中心问题是猜测单音词的位置。位置猜对了,上下层的选择必定正确,不但可以大大减轻后面解模糊处理的负担,甚至可以说是胜券在握。音节感知库的根本使命就是提供猜测单音词位置的必要信息。

请注意,我在这里用的是‘猜测’这个词。“猜测”对于程序设计的负担我不来妄言,这里只谈一点猜测的窍门。

第一,关于单音词的个数及其奇偶性。单音节和双音节段无所谓上下层。双音节段如果无模糊,就是比较可靠的‘安全岛”。单音节段如果连续出现,往往是人名、地名或新词的迹象,至于这三者之间的选择,显然需要语境知识,这将在以后说明。这里先看一下奇段(指音节数为奇数的段,偶段类此)和偶段的区别,并假定段内无三音词。这时,奇段内至少有一个单音词,总个数必为奇数;而偶段内可以没有单音词,如果有,则至少有两个,总个数必为偶数。

第二,关于单音词位置的判断。第一个单音词位置必然在奇号位置上,如果还有第二个单音词,那它必然在偶号位置上。具体到 3 音节段,它必有而且只有一个单音词,其位置非首即尾。具体到 5、7 音节段,它必有一个单音词在首(1 号)、中(3 号或 5 号)或尾(5 号)的位置上。具体到偶段,如果有单音词,对 4 音节段必在 1 和 4 的位置上(这里不考虑假双的情况);对 6 音节段,必在 1 和 4、1 和 6 或 3 和 6 的位置上。

从上面的两点分析可知,单音词在音段内的可能位置是完全确定的。这个位置一旦确定以后,上下层的选择也就唯一确定。这就是为什么我反复说明分词处理要从分段、分层做起的原因。

谈到这里,可以对上述的上下规则作更为严格的表述了。这就是:

“对第一个单音词,左上右下;对第二个单音词,左下右上。以此类推。”

判断单音词在音段内的可能位置只是一个逻辑问题,不涉及语义、语法及统计知识,但要从可能位置作进一步的判断,就需要这些知识了。音节感知库就是这些知识的集中表达。现代汉语单音词的分布特征,用信号处理的语言来说,就是具有明显的线谱。这个线谱特征表现在语义和音节两方面:语义集中于逻辑概念、基本概念和常见物的命名;音节集中在 4% 5%、即 50 60 个音节位置上(这个数字是按照 95% 的占有率统计出来的)。语义集中度我没有作过统计,音节集中度是作过的,但所选题材偏于应用文。不过,如果限制在 3 音节以上的音段进行统计,我估计这个统计结果不会有太大的出入。值得特别指出的是,音节线谱中有 6 条特强的线,就是:de, le, zai (4), he (2), shi (4), bu (4),这六条线谱的占有率可达三分之一。我相信,有关音节线谱知识的运用会产生良好的效果。

分段—分词处理的困难方面在于假双、假单现象和多音词的介入。前一个困难将在以后论述,这里只谈一下多音词介入问题。对多音词首先要作区别对待。

第一,要区分 3 音词和 4 音以上(含 4 音)的词。后者绝大多数是短语或句子,而前者则无所不包,从语素到句子都有(语素 3 音词是硕士卡引入的,如‘越来越”,“百分之”,在双音词里也引入了“有所”、“无所”之类的语素)。在词库里,4 音词略多于 3 音词,但使用频度远低于后者。也就是说,4 音词的出现是小概率事件,更不用说 5 音以上的词了。因此,对 4 音以上的词,可考虑优先原则。

第二,要区分 3 音词中 2—1 组合和非 2—1 组合情况。所谓 2—1 组合,是指前两个音

节和第三个音节都是词的情况,例如“北京市”。凡 2—1 组合式 3 音词,都不宜优先认定。非 2—1 组合 3 音词则可考虑优先。

如果考虑多音词,那么,上述规则要作一些调整。最简单的办法是,将多音词从音段中扣除,其他一切照旧。把认定多音词的可能错误放到句类分析过程中进行。这是唯一可行的方案。通常采用的都是不可取的。

在本次问答的最后,还应说明一点。独立性的意义虽已如上述,但对于字和双字词仍有一些差别。对字,只能说某级独立性义项,不能说某级独立性字,因为字一般是多义的。简单地说某级字是指频度,这个概念要分清楚。对词(指非单音词),可以说某级独立性词,因为汉语的词一般是单义的,如果多义,则默认其频度最高义项的独立性,各义项的准确说明由相应结构方程的第四级内容规定。

问 28:

一般汉语理解处理文献里经常用“分词处理”的说法,而先生却用了“分块处理”这个新提法。关于语义块和分块的概念,请先生作一次专题谈话。

答:

分块处理是理解预处理的第二步,是汉语理解处理承前启后的关键一步。如果说词是语言的基本单位,那么,语义块就是理解的基本单位。人对语言的理解,无论输入方式是看还是听,都是以语义块为单位。从形式上来说,语义块大体上相当于语法的短语或词组。但我心中的语义块是与语句要素和句类分析相联系的,所以,觉得有必要用一个新词来称呼它。

句类分析基于这样一项基本假设:简单句就是对作用效应链某一环节或一项基本判断(指关于比较、肯定与否定、存在性三者的判断)的表述。由于作用效应链有 6 个环节,所以,基本句类只有 7 个。那么,句类分析究竟与常规的语法句型分析有什么不同呢?这个问题将在以后(注:在〈问答 31〉)作详细讨论,这里只说明一下要点。

对一个句子的分析,从词的角度很难给出约束性的完备表述,但从语法结构成分的角度则可以给出主、谓、宾、定、状、补的约束。这不仅是量的约束,而且是语法意义的约束。这显然是一个巨大的进步。但从理解的高度来看,这个进步远远不够,因为,理解的关键在于概念联想脉络的约束,也可以说是词汇意义的约束。这个要害问题是显而易见的,因此在 80 年代后期,出现了语料库学派,他们试图通过统计的方法寻求词汇之间的统计关联性。然而,这是一条避难就易的道路,是与人类思维方式背道而驰的道路,不能盲目追随。

词汇意义的约束过于复杂,因为,它涉及语用或习惯问题。我的想法是先探求概念节点之间的关联性,把概念节点的联系和词汇的联系分为两个不同范畴的问题来处理。前者作为活知识对待,用规则来表示,后者作为死知识对待,用数据库的方式来表示。

语句活知识的表达单位必须是概念节点综合而成的某种基本语言结构,这个结构的核心称为语句要素,整个结构称为语义块。从独立性的意义来说,语法的主、谓、宾语与要素相

对应,但定语无独立性,不能构成要素,只能作为语义块的辅助部分;状语和补语具有半独立性,有时是要素,有时不是。这就是说,主、谓、宾与定、状、补不是一个层次的东西。词与句子之间的结构层次,语法学叫做短语。有的短语是语义块,有的不是。这就是说,词—短语—句子是语法的结构层次,要素—语义块—句子是语义的结构层次。这是两种不同的层次观点,将导致两种不同的语句分析方法。几十年的实践已经表明,前一种观点和方法不能引向理解,理解的需要在呼唤新的层次观点和分析方法。我们必须响应这一呼唤。

我和王华在汉语知识表示讨论会(注:1992 年由关定华教授和陈一凡先生发起在声学所召开的会)上的报告"概念层次网络理论概述"实质上就是讨论如何响应这一呼唤。这里不来重复该文的基本论点,只对两种分析方法所运用的不同知识作一个简单的对比。

基于语法的语句分析主要运用词性知识,表现为两项基本规则:第一,不同词类的搭配规则,例如,形容词与名词、副词与动词、数词与量词的搭配等。第二,词类与句子成分的对应规则,例如,名词与主语及宾语、动词与谓词、形容词与定语、副词与状语等。我曾检验过这些知识的解模糊效果(对语音模糊文本),只有百分之十二左右的消解率,与理解处理的要求相去甚远。

基于语义结构层次的分析方法尚处于探索阶段,山克先生的先行性工作并未取得突破性的效果。原因在于山克先生急于建造一座大厦,而基础过于薄弱。山克先生的概念基元,只有转移一项达到了层次网络的二级水平。从这个起点直接进入篇章分析,必然是力不从心。

概念层次网络的方法吸取了先行者们急于求成的教训,以全部国标汉字为素材,抽象出通用语言概念的层次网络符号体系,层次性和网络性是这个符号体系的基本特征。这个符号体系是否与人类大脑中的知识体系有某种相似之处?我不知道。因为后者的知识结构及其生成、运作过程至今所知甚少。但是,无论如何,我们必须创立一个有别于自然语言而又能表达语言知识的符号体系。这个必要性,就如同必须用阿拉伯数字改造其他一切数字表示方法一样。大家知道,没有这项改造,就没有现代科学。我认为这个对比并不夸张,再多的话就不必说了。从这个宏伟目标来说,概念层次网络符号体系的建立和完善绝非一日之功,甚至不是一代人的努力可以完成的。但我们不能知难而退,不能继续在语法学的基础上走修修补补的道路。

层次网络理论的语句分析,主要是运用概念之间的关联性。奎廉的语义网络是第一次试图从概念关联性的角度去分析语句。但他只是借用了一阶谓词逻辑的思想作为表达的工具,并没有从总体上思考概念之间的最普遍、最基本的关联性。我认为,这项总体思考与概念符号的优化设计相结合,是使计算机步入自然语言理解殿堂(广义地说是实现计算机思维)的关键性基础研究。这显然是一条漫长而艰巨的历程,但 HNC 理论至少取得了下列进展。

第一,深化了语法学的词性搭配关联性,从这个"泛指"关系中提炼出可精确化的部分,使之变为"特指"关系,这就是我多次说过的"同行优先"准则。但这条准则的涵义决不只是

词性约束的简单深化,它也包含了最基本的主谓约束和述宾约束。同行的(u, z)与(v, g, r)构成最优先的偏正搭配,同行的 v 与(g, r)构成最优先的述宾搭配,同行的(g, r)与 v 构成最优先的主谓搭配。

第二,深化了语法学三大基本词类——名词、动词、形容词——的分类思想,把它改造成关于抽象概念五元组(v, g, u, z, r)的思想。五元组这个名词是从乔姆斯基的形式语言文法理论里借用来的,但我的五元组与乔先生的四元组没有什么关系。五元组基于这样一个基本的假定:五元组是表达任一抽象概念的五个基本侧面,甚至可以说,这五个基本侧面是描述抽象概念的充要条件,故称之为抽象概念的五元性。西语的词根和汉字及汉语双字词的词性"兼类"现象,正是这种五元性的表现。关于汉语实词的分类,曾有过一场高层次的争论。高名凯教授和某些外国汉学家坚持汉语实词不能分类,认为实词的"兼类"现象是逻辑意义而不是语法意义或词汇意义的表现。语法学家不同意这个看法,努力去寻找实词的语法形态,据说还真是找到了,命名为汉语实词的功能形式,简称"广义形态",以区别于西语的"狭义形态"。语言界的许多争论,类似于古代哲学界关于宇宙的争论,"兼类"现象之争就是一个典型的例子。词根与兼类都是抽象概念多元性的表现,词根就意味着兼类,反之亦然。西语并没有把词根方式全部规范化,也有兼类。汉语并非一律兼类,也采用词根方式,例如,加"化"字变为动词兼名词,加"然"字变为形容词,加"性"字变为名词等。本来从词根和兼类的现象很容易作出抽象概念多元性的概括,可惜基于西语形态的语法框架阻碍了这一水到渠成的发展,从而迟至今日,才看到了抽象概念五元组的真面目。

五元组的新贡献在于对效应 r 这一侧面的发现。每个抽象概念必有相应的效应,这正是作用效应链的思想。值 z 这个侧面也有新意,不过,这是从明斯基的框架理论借用来的,实际上是明斯基的贡献。这里,用一个例子说明五元组的表达特点。

作用	vg0-0
力量	g0-0
主动	u0-0
能量	z0-0
功能	r0-0
弱	zu0-0-0-1
强	zu0-0-0-2
打,击	v0-0-

这里的作用、力量、主动、能量、功能、强弱是从五元组的不同角度对同一个基元概念"作用"的不同表述。这个基元概念用数字串 0-0 表示,5 个不同的侧面分别冠以字母串 v、g、u、z、r 来表示。作用的基本属性是"主动",与之对应的"被动"是基元概念 0-1——作用的承受——的基本属性,层次网络符号为 u0-1。这类基元概念可称之为"词根"型概念,因为,只有这类概念才能构成词根。西语用形态变化来表示这种概念的多元性,汉语则认为这种多元性是其固有特征,不必画蛇添足。我个人赞成汉语的选择。就"作用"这个概念来说,无所谓名词

与动词之分 就‘主动’这个概念来说,无所谓形容词与副词之分。古汉语无‘地、的’之分,只有精炼之美,并无模糊之弊。

通过这个例子,你对抽象概念的多元性应该有了一点印象,但对于这个多元家族的新成员 r 可能还印象不深,所以,不妨再介绍几个 r 概念。

迹、轨迹、历史	r1-0、r1-0-9、r1-0-10	一般过程和运动、演变过程的 r
结果、灾难	r3-0、r3-2-2	一般效应和有害效应的 r
权力	r12-4-4-1	人类在社会关系中主宰性的 r
概念、想法	r8-0-0、r8-0	人类一般思维活动的 r
学说	r8-2-0	人类创造性思维活动的 r
策略、方案	r8-3-0-	人类设计活动的 r
胜利、失败	r11-3-0-1、r11-3-0-2	人类竞争活动的 r
友谊	r9-7-1-0-	人类交往活动的 r
文明	r9-0-0	人类智能活动的 r
财富、财产、知识	r10-0、r10-2、r10-3	人类一般专业活动及经济文化活动的 r

第三,从抽象概念中分离出基本概念 j 和逻辑概念 l 两大类,其层次网络符号的设计已基本完成。在 l 层次网络设计中运用了〈本体层〉〈基元层〉的组合结构思想。我认为,这项组合设计非常成功,并等效于从语言逻辑的角度对基元概念设计的完备性提供了一项旁证。从另一角度来说,l 网络和五元组的组合运用实际上是深化了基本语法知识,并使之语义化。

第四,将具体物的命名概念分为 w、p 两大类并进而细分为 jw、pj、wj、pw、gw 等多种组合形式,从而在语义块内部和相互之间提供了更为精细的关联信息。这些字母组合的意义将在〈问答 30〉里作详细的说明。

第五,依据作用效应链的思想,提出了语句的语义分类准则。在这个准则的约束下,每类语句各有自己的联想空间或联想脉络,各有自己的句型变换特征,甚至在语句要素的数量上都有显著的差别。从实用的意义上来看,这一点最令人鼓舞。

以上所述,就是语义块这个概念产生的背景。当中国计算语言学界纷纷抱怨方块字的种种缺点,特别是在引入自然语言处理技术的过程中又纷纷抱怨汉字无词的切分标志、汉语的理解远远难于西语的时候,我也曾感到困惑。但语义块的概念彻底明确以后就完全摆脱了这一困惑。原来语言真正需要的内在切分标志只是语义块而不是词或短语。而在这个关键点上,汉语是一点也不含糊的,其标志的精细程度决不亚于任何语言。这类标志大体上相当于汉语的所谓虚词。虚词是语言表达的逻辑需要,所以,我把它与基元概念和基本概念并列为 3 大类抽象概念之一,命名为语言逻辑概念 l。

但是,语言逻辑概念的作用并不限于充当语义块的切分标志,它有自身的独立意义。独立性有强有弱,按其强弱顺序分为 4 大类(详见〈概述 14 章〉)14 小类。下面把这 14 小类中与语义块切分直接有关的 13 小类以表格的形式简述于下:

小类名称	作　　用
l0-m	主要素标志，m 取值按 E、A、B、C 顺序。
l1-m	辅要素标志，m = 0　7，详见下文。
l2-m	主要素远搭配标志
l3-m	辅要素远搭配标志
l4-m	指示 E 的逻辑意义，有助于 E 的判定。m 未定。
l5-m	暂缺
l6-m	指示辅要素，多用于句中。m 未定。
l7-m	暂缺
l8-m	指示辅要素，多用于句首。m 未定。
l9-m	指示 A、B、C
l10-m	对 E 的强调，有助于 E 判定。m 未定。
l11-m	句子切分标志
jl1-m	对 E 的势态说明，也可独立充当 E。

辅要素的分类和排序是十分重要又比较复杂的问题。在这方面语法学有不少可资利用的成果。下面所示的类序是多次反复的结果，虽并不令人满意，但总体上不会有大漏洞。以后如果有所变动，也只是子类内部的局部调整，可利用程序自动修改。

l1-0	参考对象
l1-1	方式
l1-2	工具
l1-3	途径
l1-4	比照
l1-5	条件
l1-6	因
l1-7	果

l1-m 可对全部辅要素给出标志，为什么又要引入 l6、l8 重复 l1-m 的功能？部分原因是为了迁就自然语言的缺点，但更重要的考虑是，前者仅仅是一个指示符号，而后者则另有独立的逻辑意义，对中级和高级联想有不同的参考价值（参看〈问答 31〉和〈问答 36〉）。

有切分就有并合。语法学的前缀、后缀、助词、连词和副词的概念里有并合的共性，我们引入字母 q 和 h 代表这一共性。传统意义的前缀和后缀单用 q 或 h 表示，以逻辑意义为主的助词（如‘着、了、过’）和副词（如‘已、正、将’）用 lh 和 lq 表示，后面可附加五元组的字母。汉语里词性身份不明（词典里用‘在某类词之后’或‘之前’的说法）但对于构造新词十分活跃的语素（如‘到、出、成、化、性’）也用 lh 或 lq 表示，但五元组的附加方式分两种：插入在 l 和 h（或 q）之间的表示新词的词性，因此，这实际上是一种词性变换符号；句内连词（如‘和、或、及、并’）和汉语里使用频度最高的‘的’用 lhq 表示。这套符号的设计，除了充分揭示语义的共性（这是层次网络符号的灵魂）之外，就是为了对语义块的切分与并合尽可能给出明确的

标志。并合概念的数字串沿用逻辑概念的一般规则,但由于并合概念的本体层限于 4-m 或 6-m,而这两者的区分也仅有深层意义,所以,在简化表示或中间表示的情况,本体层甚至基元层都可以省略。

我所说的语义块切分准则——lv 准则,就是以上述理论考虑和符号设计为依据。我想,这条准则的 l 部分已不需要再说什么了。v 部分的问题则既简单又复杂。说它简单是因为 v 是天然的语义块切分标志,一句话就够了。说它复杂是因为,从浅层方面来说,它与语法学的中心动词、形态标志、词性等概念有关;从深层方面来说,与句类分析的概念有关。因此,对这个问题的深入讨论要等到句类分析的讨论以后,今天只说一点务虚的看法。

中心动词、形态标志、词性的作用都被语法界过分夸大了(当然,对于语言教学也许不算过分夸大),从而产生了一种似是而非的推论:即"汉语的理解大大难于西语"。语法框架导致的这个错误结论,其危害性不言自明。下面,我想用轻松的方式讲点故事,引发大家的思考,也算作今后正式讨论这个问题时提前抛出的引玉之砖吧。

大家知道,西语的一个句子必有一个中心动词,中心动词与非中心动词带有明确的形态标志。计算机根据这些标志,很容易找到中心动词,从而完成特征要素的辨认。西语在这方面确有一定的优势,但也应该看到:第一,由形态标志给定的中心动词不一定是语义的中心;第二,形态标志不可能十全十美,西语的词同样有多词性的现象,不过,不像汉语那么普遍罢了。汉语在这方面的劣势不宜夸大,也应该看到汉语的优势方面。

夸大汉语劣势的人喜欢举下面的例子:

开会讨论教育改革计划

5 个动词连用,从语法形式来说,简直是"不成体统"。为了比较,不妨看看这个例子的英译。

Have a meeting to discuss the plan of education reform

这样一对比确实把英语在形式分析方面的便利表达得非常生动。plan 和 reform 虽然有词性混淆,但伴随的 the 和 of 把这个混淆完全消除了。各词之间的语法关系标志得清清楚楚。英汉的优劣似乎是判若云泥。但是且慢!如果把汉语的表达方式改成:

开会讨论一下教改计划

开个会,把教改计划讨论一下

教育改革计划将开会讨论

你看,情况就有所不同了。只用第一种在汉语里较为少见的表达形式与西语比较是不公平的。汉语虽然没有形态变化,但仍然拥有完备的手段在必要时指示语法成分或词性,这里的"一下"、"个"、"把"和"将"就是这种手段。

屈折语的形态变化曾被 19 世纪的某些西方语言学家视为语言进化的标志。这个错误观点到 20 世纪初叶已为西方比较语言学家所彻底抛弃,然而却仍然在中国有一定市场,这一点很值得反思。中国有个训诂学,专门探究中国古代文献中的疑难问题。从语言理解的本质来说,它是自然语言理解的先驱。按照语法学的形式逻辑,训诂学与语法学的土洋结合,应有"如虎添翼"之效。然而,正是近代中国最有成就的训诂学家(章黄学派创始人之一、

有侠儒之称的黄侃)把‘马氏文通’戏称为‘狗屁不通’。不能把这个故事简单地看成是保守的训诂学家的排外心理表现。因为训诂学基本不需要语法学,在训诂学家看来,语法学的那点知识,不过是语言大义的 ABC。学术实践也确实表明,正是‘传统训诂学家用古老的弓箭射中了别人用现代武器未能击中的目标’(北师大俞敏教授语)。这种现象并不奇怪,因为训诂学的疑难是音义的疑难,不是语法的疑难,更没有形态的疑难。自然语言理解不应该忘记这个历史教训。

汉语还有一个比‘无形态变化’更骇人听闻的现象,就是书面古汉语没有任何标点符号。所以,训诂学的疑难中有断句的疑难。断句歧义现象(就是语义块切分问题)有不少著名的有趣故事,有的纯粹是文字游戏,有的则表现了汉语文学的意境美。这里不妨各举一例:

下雨天留客天留人不留
下雨天留客,天留人不留
下雨天,留客天,留人不?留!

黄河远上白云间,一片孤城万仞山。
羌笛何须怨杨柳,春风不渡玉门关。
黄河远上,白云一片,孤城万仞山。
羌笛何须怨?杨柳春风,不渡玉门关。

第二个例子的两种断句方式都是很好的诗。前者是唐代边塞诗人王之涣的名作之一。后者则是清代大文豪纪晓岚的趣作,他奉命为乾隆皇帝题写扇面,他选了上面的王诗,但无意中掉了一个‘间’字,皇帝责怪下来,纪先生急中生智,用第二种断句方式朗读一遍,结果乾隆皇帝转怒为喜,免除了一场不敬之罪。

这个例子说明,从最高层次的理解即文学语言欣赏的角度来说,连标点符号都是多余的东西,何况形态标志和词性之类的外在标志?艺术作品不需要标点符号,音乐美术是这样,文学语言也是这样。这些看起来是题外话,但在语法框架依然统治汉语理解的今天,这一最高层次理解的本质——只依赖于内容,不依靠外在的包装——难道不是很有一点‘振聋发聩’的启示意义么?

我的故事今天就说到这里。

问 29:

最近我正在按先生提出的句类分析思路进行程序设计的思考,觉得难点在于省略语义块标记的非标准格式,像下面的句子:

1. 诸葛亮是政治家的典范,政治家的典范是诸葛亮。
2. 西风漫卷红旗,红旗漫卷西风,红旗西风漫卷。
3. 雨打芭蕉,芭蕉雨打。

4. 我去过上海,上海我去过,我上海去过。

先生如何看待这些句子的处理?

答:

标准格式的说法也许在前面加上限定词‘书面语’更合适一些,因为书面语基本遵循(对遵循一词应作为广义理解,改变标准格式的顺序时,按约定加上语义块指示符当然也是遵循)标准格式,而口语的情况则大不相同。如果说口语的庄严体和正式体还大体遵守,那么,普通体,非正式体及随便体就可以说基本不遵守。这些语体中常见的重复和省略,更使得标准格式的提法毫无意义。但是,不论是书面语还是口语,都必须遵循格式。

现代汉语的书面语比古汉语更遵守标准格式。古汉语把打破标准格式作为创立新颖风格的手段之一,把‘枕流漱石’将错就错作为‘枕石漱流’的替换表达就是著名的例子。“红旗漫卷西风”不能说与李清照的诗句‘帘卷西风,人比黄花瘦’没有关系。

句类分析的第一步就要用到标准格式的知识。格式的标准顺序可以打破,因此不能把标准格式当做死板的规则来使用。但格式的本质,即格式里的特定角色是不变的,否则,语言就不能被理解。对这一不变性和可变性的把握,要做到像人类思维那样游刃有余,概念层次网络理论目前可提供的知识也许不够,但随着联想脉络的逐步完善,这一目标是可以逐步逼近的。

现在将上面的论述作一个概括性的叙述:每一句类各有若干特定的角色(语义块),语言理解就是辨识这些角色。在标准格式里,角色由位置唯一确定,因此,偏离标准格式时就必须加以标记。但是,当词汇本身具有明确的角色特征时,可以省去角色标志。语义块位置的可变性及其角色的不变性是句子的基本特性,“是”字句也不例外。

自乔姆斯基以来,语言深层的概念已为人们所熟知,但语言深层结构如何表达呢?这一直悬而未决。句类和格式的概念回答了这个根本问题。

语义块切分的 lv 准则就是以格式概念为依据的。

但是,在例 2 到例 4 中,我们看到了 l 准则或标准格式失灵的情况,如何解释这一现象呢?答案是词汇角色的确定性(注:应为语义角色),或曰 E 与 A,B 的强关联性。例 3 中的“我,上海”必然是分别充当 TA 和 TB,这个知识已包含在自身转移句的定义或节点 2-3-11 的定义中。例 2 中的‘红旗,西风”,例 3 中的‘雨,芭蕉”,其角色的确定性不像例 2 的‘我,上海’那么明显,但西风和雨属于效应物 rvw,而红旗分别属于 pw 和 jw6-1,这里的类别知识已蕴涵着角色信息。

标准格式的概念,意思是指各角色的自然排列顺序,显然,它与语种有关,现在制定的标准以汉语为蓝本。不同语言的翻译就有标准格式的转换问题。标准格式只是一种表达的习惯,不是理解的本质,本质是一种格式里有几个不同类型的角色。格式是角色的排序,或语义块的排序,不是语法学的词序概念。语法词序概念过于宽泛,包括语义块之间的排序及语义块内部的排序。我们将把词序的概念内涵压缩,专指语义块内部的排序。语言逻辑符号 l0 到 l3 专门用于格式的指示,语义块内部的组合符号,即‘词序’的指示符号则用结构符号

(主要是“ / ”)及类别符号 q,h 表示。

我曾经多次说过特别喜爱“角度,角色,类别,层次,格式,脉络”这六个词,它们概况了语言的精髓。就如同“礼”、“无为”、“空”分别概况了儒学,庄子和佛学的精髓一样。我曾反复阐述这六个词的内涵,然而总觉得意犹未尽。这种困境,有点类似于训诂学家对“礼,无为,空”的诠释之不可穷期。句类的转换是表达角度的转换。亚里斯多德把句子统一视为命题,从句类转换来看是有一定道理的,因为作用效应链的六种基本句类都可以转换成“是”字句。这一“六到一”的转换是句类转换的第一项基本特征,也是“是”字句的独一无二的特征。惟有这个“六到一”的转换是无条件的,而其他的转换则都是有条件的。亚里斯多德当然没有认识到这一点,惟其如此,“命题”说的提出更表现了他的天才。不过,还应该进一步说明,虽然“六到一”的转换是无条件的,但实际语言运用这一转换的情况毕竟是少数。命题说是高度的“合”,而语言理解需要“合,分”兼顾。基于这一点,我曾在某一〈问答〉中批评过语言命题说。

下面就来着重谈一下“是”字句,它是基本判断句的子类之一,是最常用的句类。其句类格式是

DB + jD + DC

它不存在非标准格式。“是”字之前一定是 DB,“是”字之后一定是 DC。

从语义块的形式来说,DB 应优先于具体概念,DC 应优先于抽象概念。但应该指出,“是”字句的 DB 可隐含 C,而 DC 也可以隐含 B,这是一般规律。“是”字句的独特之点在于,隐含的 C 和 B 可以转化为主构成,这是“是”字句独特的句类知识。因此,诸葛亮既可充当 DB,又可充当 DC,就不奇怪了。

基本判断句有一个独一无二的特征就是特征要素 jD 与对象要素 DB 和内容要素 DC 的关联性几乎为零,而其他句类不是这样,其关联性可强可弱,但绝不等于零。

“是”字句是以三级节点 j11-1-k,k = 1, 2 为特征要素的三级句类,它有命题说的历史背景,有上述一系列独一无二的特征,所以,对这个节点冠以类别符号 j1 是“当之无愧”的。那么,这一子句类的 E 语义块是否也有与众不同之处?这里想顺便说一下 E 语义块的特色,即其自身构成的联想脉络。E 语义块以特征要素为核心,这不在话下,但其“绿叶”部分的构成与其他语义块大不相同。“绿叶”优先于同行概念及交式关联行的概念,这是一般原则。但非 E 语义块的“绿叶”优先部分并不集中,联想脉络的把握并不容易。惟有 E 语义块集中在三个方面,一是情态 j11-2,j11-3;二是时态及语态 qv,hv,vq,vh;三是 l10-k,把握比较容易。层次网络逻辑符号的设计,完全打破语法学的虚词分类标准另立体系,就是为了便于各类语义块的辨识。jD 的确与一般 E 语义块有所不同,一是它的“绿叶”在前不在后,无 hv,vh;二是前面的“绿叶”比较稀少。

下面谈两个“是”字句的特殊表现,属于语言学的“小儿科”,不过,你作为程序设计者却不可等闲视之。所谓两特殊表现,一是“是”字句的一种特殊表达形式,二是“是”字句的简化。“是”字句的一个特殊形式表现为“是……的”句,DC 语义块以“的”字结束。如:

鲁迅的骨头是最硬的。

曹雪芹的晚景是很凄凉的。

苏联的解体不是偶然的。

台独分子制造一中一台分裂祖国的叛逆活动是注定要失败的。

改革开放的历史进程是不可逆转的。

霸权主义是中国政府和人民坚决反对的。

我们的事业是正义的,正义的事业是一定要胜利的。

我希望通过这些例句表明两个语言现象,一是书面语的"是……的"句属于一种郑重的、强调性的陈述方式;二是势态的表达适合于采用"是……的"句。

"是"字句的简化:英语的"Yes","No",汉语的"是","不,不是,否"("否"是 j11-1-2 最郑重的表达方式,毛泽东在"中国农村的社会主义高潮"一文中曾用过这种方式。)是"是"字句的最简化形式,简化到只剩下特征要素 jD。

一般的简化方式是承袭上文,省去 DB 或 DC 语义块,但前者更常用。例如:

诸葛亮是杰出的政治家,但不是军事天才。

拿破仑不仅是天才的军事家,也是天才的政治家和外交家。

俄国不再是超级大国,但依然是军事上的超级强国。

五台山是著名的佛教圣地,普陀山也是。

武汉的夏天很热,南京也是。

承袭上文省去语义块是汉语十分发达的语法手段,西语则常采用"代"的方式,以"代"为"省","代"就是"省"。(当然上面的例句西语也用"省"的方式。)不同句类的"省"各有特色,例如作用句常省去 A 语义块。"省"或"代"是理解的一大课题,超出了本次问答的范畴。

"是"字句的简要说明大体上可以结束了。最后,还想补充一点,就是汉语的"是"字句表达要比西语丰富多彩。因为汉语为 j11-1-k 都设置了多个语言符号,j11-1-1 的表达就有"为(阳平),即"等字及"……者,……也"的方式。在名称的表达上,还有"也叫,亦称"等替代方式。这就不多说了。

问 30:

先生用小写英文字母串和数字串表达自然语言抽象概念体系的类别和层次特征,用大写字母串表示语义块的句类和类别特征。这两类符号体系凝聚了先生的精深思考。第二套符号体系比较简明,第一套符号则非常错综复杂。这次,请先生就该符号体系的总体要点、字母串的连用规则、具体概念向抽象概念的挂靠以及非挂靠类具体概念的表示作一个综合说明。

答:

首先我要说,任何一个抽象概念都具有类别性、层次性和网络性三个方面的特征,不只

是类别和层次两方面。数字串不仅用于表示层次特征,也用于表示网络特征,层次符号的高中低层之分即由此而来。高层主要用于表现层次性,中层用于表示概念的对比、对偶和包含性,低层则主要用于表现网络性。所以,我把这个理论叫做“概念层次网络理论”。

那么,如何从最高层次去概括概念类别特征?这显然是一个带有一定哲学意味的难题。因此,我不敢说现在的理论概括是接近完善的,但可以说,它成功地经受了汉语的检验。

概念的类别首先表现在所谓实物与抽象之分,这是众所周知的。我的新贡献在于把抽象概念分为基元、基本和语言逻辑 3 大类,并对它们用(v,g,u,z,r)五元组表述。五元组这个名词是从乔姆斯基的四元组概念借用来的,v,g,u 来于语法学,z 来于明斯基,只有 r 是全新的概念。五元组是一个整体,r 不可缺少,尤其是对于基元概念,没有 r 就不可能获得对一个概念的完整理解。我在这里用了不够谦虚的“新贡献”一词,就是由于对 r 的发现。至于基本概念 j 和语言逻辑概念 l 的引入,把实物概念又分为一般物 w 和人 p 两类,并将一般物的物性又独立出来记为 x 等等,都不过是一些小技巧罢了。

三类抽象概念的“地位”有所不同,基本概念与基元概念大体并列,逻辑概念次之。基本概念包含的内容是古代哲学探索的基本对象。虽然这个对象的主要部分已独立成为现代自然科学的基本分支,现代哲学实际上只需要继续探索我们约定的 j7 和 j8,但从概念的总体来看,古代哲学所探索的基本对象仍然是最高层次的一类概念。

对三类抽象概念不同“地位”的认识,应视为所有知识中最基本的知识。因为基本概念 j 的“地位”最高,才用它形成独立的概念类别 jl——基本逻辑概念。而且,基元概念和语言逻辑概念都可以向它挂靠。由于基元概念的“地位”高于语言逻辑概念,所以后者也可以向前者挂靠。这样,从三种抽象概念的类别基元 j,φ,l 又形成了 jl,lj,lφ 三种复合类别基元。所谓挂靠就是对本体层(前者)的“义”用挂靠层(后者)的“义”予以更充分的说明。

挂靠的思想显然也适用对具体概念的表述,就是说,用抽象概念的“义”去阐发具体概念的“义”。当然仅仅依靠挂靠,具体概念的“义”可能不完整,甚至很不完整,只是一个很粗糙的近似。但关键在于这样就容易引发具体概念与相关抽象概念之间的联想,也就是我“挂在嘴边”的说法,人工地制造出同行优先的符号表示。这种粗糙近似的效果,你可以从盲人的语言感知能力并不亚于或不明显亚于常人得到启示。这个现象表明,自然语言理解处理对具体概念的把握可以充分利用“不求甚解”的“花招”。我看到一些语义分类系统(主要是日本人的)把具体概念搞得很细,简直成了生物学教科书,而对抽象概念则大而化之,这太本末倒置了。

至于字母串的连用规则,我认为很简单,可概括成下列 3 条规则:

1. 概念类别有基元与复合基元之分,基元只有一个字母,复合基元有两个字母。

2. 概念类别基元一定在五元组基元之前。

3. 五元组基元连用主要就是表现所谓词性兼类现象。当然也有特殊约定,如自身连用 zz 表示量词等。

j1 已有网络节点表,这里只对另外 4 类具体概念作简要介绍。

(1)jw 类,目前给出了下列定义:

jw	宇宙
jw0	火
jw1	光
jw2	电
jw3	声
jw4	微观物质
jw5	宏观物质
jw6	生命体

这些宇宙构成的基本物质显然应自成体系,不宜采用挂靠的方式。当前的安排只能是一种尝试,需要逐步完善。下面说明由这些定义构成复合概念的简单思路。

由 jw 和 j4-0—0(部分)、j4-0—0-0(成员或元素)(注:"—"表示包含性概念)之并可构成星系、星球、恒星、行星、卫星等概念。但这样做,后面 3 个概念的表达过于复杂。实际上,恒星和行星分别采用 w0-0 和 w0-1 的定义比较合适,而卫星则采用 w0-1/w4-4-2 的定义,由此,太阳、地球和月球分别定义为 ww0-0、ww0-1 和 ww0-1/ww4-4-2,人造卫星自然就是 ww0-1/pw4-4-2。

由 jw0 和 jw1 可引入两项基本物性概念——温度和色彩,它们分别是 jx0 和 jx1。另外两项基本物性——体积和重量则分别定义为 xj2-0 和 xj5-1-8。应该说明的是,jx 与 xj 都是把 w 置换成 x,以表达相应 w 的物性。

把"火"定义为基本作用物 jw0 还有另外一点想法,就是对于古代哲学的纪念。"光"定义为 jw1 是无可争辩的,但把"声"定义为 jw3 多少也带有一点纪念的性质。

(2)wφ 类示例:

w1-2-1	原料
w1-2-3	材料
w2-2-10	路
w4-1-1	化合物
w5-1	山

(3)pj 类示例:

pj1	时代
pj2—	国家
pj4-0—	集体
pj5-2	民族
pj7-1-1	男人
pj7-1-2	女人

在这六个基本定义中，pj2— 和 pj40-0— 是包含性概念。例如，

pj2—0　　省

pj2—0-0　　县

pj2—0-0-0　　乡

国家行政单位分为 4 级，多于 4 级者用包含性符号“ — ”表示，例如中国的专区可定义为 pj2—0—。

(4) wj 类示例：

wj1—，wj1—0，wj1-0—，wj1-0—0，wj1-0—0-0，wj1-0—0-0-0

wj1-1，wj1-1-12-4-1，wj1-1-12-4-2，wj1-1-12-4-3，wj1-1-12-4-4，

wj2—，wj2—0，wj2—0-0，

第一组相应于世纪与年代、年、月、日、时，第二组相应于季节和春夏秋冬，第三组相应于地域、地区、地点。

具体概念除基元符号 w 和 p 之外，还定义了下列复合基元：

pw　人造物　　其中 pwj2— 定义为城市

gw　信息产品

rw　效应物 如：

rw1-0-9　　流

rw2-2-11　　河流

rw3-0　　一般效应物，如云、雾、霞等

最后，考虑到语言的特殊性，定义了下列特殊符号：

jgw10-3 * 0　　言语

jgw10-3 * 1　　语言

jgw10-3 * 2　　语音

jgw10-3 * 3　　文字

jgw10-3 * 4　　图符

“ * ”表示挂靠结束，它后面的数字串是自定义的层次符号。

问 31：

我们的工作即将进入以句类分析为主的阶段，所以，希望先生提前谈一下句类分析问题。从具体工作来说，首先是句类的划分标准或依据问题。先生是按照作用效应链的思想来划分句类的，但实际的概念很少仅依附于作用效应链的一个环节，往往是多个环节的组合，这是第一个问题。其次，许多概念节点之间存在交式关联性。第三，6 行以后表达人类活动的概念都是复合的，如何分解为作用效应链的个别环节？最后，我想提一个特殊问题，就是究竟是否有必要划分作用句和效应句？先生曾一再强调作用与效应不可分离，分为两个句类也许从理论上有一定的意义，但从程序设计来看，很可能是弊大于利。

答：

从层次网络理论的三个构成部分来说，我对于句类分析部分的把握性最小。如果说五元组和概念矩阵的思想已较为严格地经历了“假设、求证”的循环，语义结构方程的想法大体上也经历了这个循环，那么，应该说句类分析的想法基本上还处在理论上的推想阶段。前两个部分的“假设、求证”循环过程可以主要依靠词典，但句类分析的这个循环过程则必须主要依靠语料库。现在，我们要把建设语料库的工作和句类分析的“假设、求证”工作结合起来。胡适之先生在提倡“假设、求证”方法的时候，分别加了“大胆”与“小心”的方法性说明。一个人做任何事情都应该注意这两方面（做学问更不必说），可是，人的素质往往是此强彼弱，我这个人的性格就是“大胆”有余，“小心”不足。因此，这里不能不说明一下，在没有对句类分析作初步求证之前讨论你提出的问题，过于“大胆”的缺陷是在所难免的。

自然语言理解的困难，或者准确一点说，计算机理解自然语言的困难，主要来于自然语言的词而不是语法。词的多义性、近义性和语用性是理解处理的最大难点。但是，要突破自然语言理解的难关，你不能老是回避这个难点。当然，不能幻想一口吃成一个胖子，要细心寻找比较容易着手、又能通向理解彼岸的“入口”。我认为，这个“入口”就是句类分析。

如果把自然语言的语句按词变换成层次网络符号，一般来说，句类的确定是一个比较容易的问题。句类分析是一个新概念，不能沿用语法分析的老观念，把它简单地理解为主要是寻找中心动词或特征要素的问题。句类分析是在一个句子或一个语义块的范围内进行上下文概念联想的过程。这个概念联想过程的主要使命是：判断哪些概念是可以优先组合的，这些优先组合的概念能否构成一个完整的意思。所谓词的多义性问题或句子的歧义问题正是进行这一项判断时需要解决的难点。例如，汉语的“去”字，有独立意义又能独立使用的义项有 4 个：v3-8-2，v0-0↑v3-1-2，v2-2-11，v2-2-11-2。在下列组合——“去伪存真”、“去皮吃”、“去上海”、“扬长而去”中，依次取上列 4 义项。从词性或语法的角度，这个“4 选 1”问题是无法解决的，它实际上超出了语法学的范畴。但是，转换成层次网络符号以后，这个问题并不难解决。先看第一个组合，这里的“存、伪、真”都有多个独立义项，因此，义项的选择不再是简单的 4 选 1 问题，而是 n 的 4 次方问题，似乎非常复杂。但是，如果注意到“存、伪、真”的义项中只有 v3-8-1，j8-1-1，j8-1-2 可相互构成同行匹配，那么，显然就可以通过一系列比较简单的基本判断“就地”（指不必寻求更大范围的语境知识的帮助）解决问题。第二个组合的“皮”，其独立义项只有一个——jw6-1—，它优先与作用型概念匹配。“去”中只有一个作用型概念，因此，这是一个简单的 4 选 1 问题。后面的两个组合也是如此，因为，pwj2（城市—上海）与 v2-2-11（自身转移）优先匹配；至于最后的“去”应唯一选择 v2-2-11-2（自身转移的离开），则是一个有趣的巧合，因为，“扬长”的符号恰好是 u2-2-11-2。

我希望通过上面的例子说明：句类分析的第一步就是进行概念的初级联想，其具体内容就是判断哪些概念可以优先组合，判断的准则极为简明：概念矩阵的同行优先和交链式关联优先。而这个判断过程本身就是一个强有力的解模糊过程。这里需要着重指出一点：例子

的模糊虽然是多义性的,但上述原则是普遍适用的,因为,所谓 5 重模糊,无论在本质上或形式上都可以作为多义性模糊来处理。当然,不同的模糊各有不同的解模糊手段,例如,语音识别模块输出的多音模糊、汉语的音字转换模糊都能通过与一个适当的汉语词库进行匹配而予以有效的消除。但残余模糊是不可避免的,对这个“残余”,实际上只能当做(实际上就是)多义性模糊来处理。因此,句类分析的本质使命之一就是在一个句子或一个语义块的语境范围内进行解多义模糊的处理。整个理解系统的‘3-7-3’构成也要从这个角度去认识,前面的‘3’就是为了把前 3 重模糊简化为解多义模糊来处理,或者说是为中间的‘7’进行解多义模糊处理‘打扫战场’,而后面的‘3’是在更大的语境范围作进一步的解多义模糊处理。这里有必要顺便说明一点:预处理在减少多音模糊和音词转换模糊的同时,也会增加其他的“模糊”,例如词库匹配过程中出现的‘假词’(包括双字词和多字词)现象。“假词”的出现,是错误问题,因此,对它的辨识就不是解模糊,而是纠错问题。但是,这两个问题的界限并不是绝对的,我们可以把‘假词’问题转化为解模糊问题来处理,这一点,将作为一个专题另行讨论。

在上述概念初级联想的过程中,自然也就完成了句类的设定。在假定句类以后,就可以进行语句合理性的分析,这个问题以后来谈。现在回到你提出的第一个问题,就是当一个复杂概念中出现不同基元概念的层次网络符号时,是否会形成句类映射的不确定性。这个问题我想从技术和理论两方面都谈一下。从技术方面来说,这只是一个约定或定义问题。这涉及两类定义,即组合结构符号的定义和复合基元概念的定义。

组合结构符号可分为 4 类:作用效应类、对象内容类、逻辑类和语法类。这个次序大体上是按照它们对句类分析的作用大小来排的。第一类的作用最大,如果在符号序列中出现了这类结构符号(当然必须在整个结构的外层),则句类由它唯一确定。符号↑一定相应于作用句,符号↓一定相应于效应句,不论这符号左右的内容是什么。例如‘教导’和‘危及’这两个词,映射到层次网络符号是:

教导　　v9-2-3↑v8-1-2

危及　　g5-3↑v3-2-2

这里的基元概念并没有作用基元,但组合概念是作用型概念。能够理解这些基元概念的‘同志’(包括计算机),就不难理解这两个复合概念。

对象内容类结构对句类映射的作用不像作用效应类那么简明,甚至可以说,这时决定句类的因素不是结构符号,而是其左右两侧的层次网络符号。但是,这个结构符号对句类分析仍有重要参考价值。对象结构是将 E、B 合并,内容结构当左侧为 v 时是将 E、C 合并,因此,在进行语句要素 B、C 的判定时,这类结构符号显然是最明确不过的指示。

谈到这里,为了下面叙述的方便,我想把组合结构符号的类名、符号约定、结构方程编号、同语法学名称的大体对应关系综合成下示的表格。

组合结构名称	符号	结构方程编号	语法名称
作用效应	↑,↓	4,5	后补
对象内容	→,|	6,7	述宾
逻辑	(,;),(,|,)	1,3	前－联合、后－无
语法	/,‖	2,9	前－偏正、后－主谓
“假”组合		0	联合

从这个对照表可以看到,每一个组合结构类都有两个子类。读者也许对结构方程的编号感到奇怪,为什么不严格按照组合类排序?这涉及组合的另一种特性——对偶性。语法学由于植根于印欧语系,所以完全忽视了这一重要特性。如果借用语法学的术语,那就是除了“后补、述宾、偏正、主谓”结构之外,还有相应的反结构——“前补、宾述、正偏、谓主”。但是,逻辑结构的子类——(,)或(;)以及“假”组合类不存在对偶性。结构方程的编号以对偶性的有无为参考,0 1 号无对偶性,2 号以后(含 2 号)有对偶性。这里顺便说明一下“假”组合的含义。它主要是现代汉语双音化所造成的一种特殊的语言现象,实际上不存在组合作用,词的意义完全由两字之一或其中的任一个所决定。不过这个字义多数情况是不独立的,必须与另一字搭配以后才能独立使用。从语言深层来看是画蛇添足,但从表层来说则起着减少模糊和调节韵律的作用。这种“假”组合包括汉语原有的所谓连绵字现象。下面回到原来的话题。

逻辑结构中的第一个子类对句类分析几乎不提供附加的信息。这里应该说明的是:当一个复合概念是若干概念之“并”时,句类分析时以第一个概念为主。例如,“同意”和“反对”这样两个最基本的反应概念,在层次网络符号里仍然是复合概念。“同意”——(v9-0-2,j1v1-1),“反对”——(v9-0-2,j1v1-2),在作句类分析时,只需要考虑 v0-2,从而判定它是反应句。逻辑结构的第二个子类具有指示辅要素的作用。关于辅要素的问题我在〈概述 14 章〉和〈问答 28〉里都有所涉及,但远不够全面和深入,将在这一个问答的稍后一点详谈。

语法类结构是语句中最简单的结构,儿童学习语言一定是从这种结构开始。说句笑话,如果说某些动物也有言语的话,我相信,其结构超不出语法类的范围。不过,要注意,我们说的语法类也包括反结构在内(读者可能还很不习惯反结构,那么请记住“年轻”和“下雨”这两个词,它们就分别是偏正和主谓之反)。语法类中的主谓和对象内容类包含了特征要素 E 及要素 A、B、C 之一,因此,如果一个词按这些结构组成,它就具有了语句的基本形态。这类词的独立性大多数属于 A 级,在口语里往往可独立构成句子。不过,我不希望“最简单结构”的说法引起“内涵也简单”的误会。实际上,偏正结构的内涵是非常丰富的,我曾寄厚望于语法学的研究成果,可惜,即使是在这个局部领域,语法学也没有达到理解处理所需要的系统性,还需要作很大的努力,才能形成 2 号结构方程 2、3、4 级表示的规则性知识。

由组合结构符号派生出来的语义结构方程共有 4 级表示。上面只涉及第一级和第三级(对偶性),还有两级的定义问题以后再谈,下面转到复合基元的定义问题。

这里说的复合基元概念仅包括基元概念矩阵的 6 13 行。这 8 行概念是 0 5 行概念的扩展集,用于表述人类的活动。扩展方式有两种:简单升级方式和综合方式。前者的数字串设计采用(本体层)+(基元层)的挂靠结构,后者则不与基元层挂靠。我们用升级方式表述三个层次——本能层次、智能层次和社会层次的人类活动,分别安排在基元概念矩阵的 6、9、12 行。从理论上来说,人类活动的三个层次是十分明显的。自然语言对这三个层次未予严格区分,这是自然语言符号体系的重大缺点之一,但在词汇级仍有一定程度的反映。这项知识非常宝贵,是确定大语境(指段落级或篇章级的语境,语义块或一个语句范围内的语境称之为小语境)的重要依据。这三类概念的本体层只定义了一层。从容纳性来考虑,定义两层较为适当,以便留有余地(注:6 行后来加了一层),对挂靠方式或其他不可预测的因素提供一个作补充说明的空间。但也可采用扩展基元层的方式以提高容纳性。我觉得这种方式的理解效率要高一些,所以决定选用后者。这样,这三行的 k(以后为便于叙述,将用 ikm 分别表示数字串的前三层)就不限于 0 5,而可以扩展为 0 13。例如,对 6 8 给出了统一定义(参看节点表),分别命名为"劳作与服务"、"交往与娱乐"、"幻想与信仰(宗教)",这实际上是综合方式。所谓综合方式,也可称之为模糊组合方式,是某些基元概念的一个模糊集合。这里"模糊"的含义是:它大体上相当于逻辑组合类的并,但也兼有其他组合类的特征。例如"劳作",它主要是 v0-0,v3-0-1,v3-1-1 的并,又兼有↑和↓的组合含义。"服务",主要是 v3-0, v4-2-1,v4-3-1,v4-4-4 的并,又有↑和|的组合含义。"交往",主要是 v2-4-9,(v3-1-1;v3-4-1;v3-5-1)|g4-0 的并,又与(v2-2-11,v3-9-1)有"并选"的组合关系。从综合方式的定义可以看到,其数字串的设计不可能采用挂靠结构。

按综合方式设计的复合基元概念有人类的专业活动、追求活动和观念活动,分别安排在基元概念矩阵的 10、11、13 行。人类的心理活动和思维活动也按综合方式设计,分别安排在 7、8 两行。7 行是人的心理形态,由它构成的句类主要是反应句。但应该提请注意:7-3 以后的节点用于表达语用型的语法知识,在层次网络符号中仅作为概念的注解之用(注:后来将此类语法知识独立为 f 类概念)。创造性是人类区别于动物的基本特征,而创造性来于人类的高级思维活动,因此将思维活动排在人类的所有创造性活动之首。复合基元概念的排序大体上依据人类活动的进化过程。当然,这个排序也反映了我个人的一些想法:人类的活动正处在 10、11、12 的巅峰,但处在 13 的低谷,而 13 才是人类最高层次的活动。因此,我深信:一个类似于 15 世纪的文化复兴运动必将在某个时候来临。但下一次不是向柏拉图回归,而是向柏拉图的老师苏格拉底回归,向孔子和庄子回归。在物质现代化的狂热中,人类终于萌生了对物质环境恶化的警惕,因此,精神环境的退化和堕落是绝不可能永远被人类忽视的。

关于表述人类活动的复合基元概念的句类映射,从上述说明可概括出下列两条基本规则:(1)以简单升级方式扩展的概念按挂靠的基元层作句类映射,这包括节点 6-k, 9-k,12-k, k = 0 5。(2)以综合方式扩展的概念应按其导源的基元概念作句类映射。由于导源的概念多数不只一个,因此,这类概念多数导致混合句类,如下面的清单所示:

7-0,7-1	反应句
7-2	状态句
8-0,8-1,8-2,8-4	狭义判断句
8-3	效应句
10-k	作用效应句
11-0	效应句
11-1	作用效应句
11-2	效应过程句
11-3,11-4	作用关系句
13-k	效应状态句

这个清单纯粹是基于理论设计的推论,没有通过“假设、求证”的循环。语言现象的复杂性要求我们必须有规则外的应变措施,这种措施将安置在语义结构方程的多级表示里。

说到这里,大体上已回答了关于句类映射的技术性问题。下面转到理论方面的讨论。你提到了作用句与效应句界限的模糊性,实际上,这种模糊性同样存在于所有有交式关联性的概念节点之间,包括类间模糊和类内模糊。类间模糊的典型例子有:运动过程(1-0-9)句与自身转移(2-2-11)句,过程的代谢(1-4)句与效应的生灭(3-1)句,效应的予取(3-10)句、分合(3-9)句与关系的得失(4-6)句、离合(4-1)句,过程的趋转(1-3)句与一般效应(3-0-9)句及消长效应(3-4)句,因果判断(8-1 的子集之一)句与过程的源流(1-2)句等等。类内模糊的典型例子有:效应的消长(3-4)句与积累消耗(3-11)句,关系的依存排斥(4-2)句与支持反对(4-3)句。我一下子举了这么多的例子,而且还用了与节点表的名称略有不同的句类命名,感到十分不智。因为,这会给读者造成一个句类分析十分困难的误解。句类之间的模糊是概念的固有属性,但句类分析并不追求无模糊的分类,它允许“模棱两可”。问题的关键就在这里。当然,这需要进一步说明:为什么可以这样做?如何实施这种做法?

首先,应该明确,分类不是目的,而只是一种手段。人理解自然语言,不会作什么句类映射(当然,也不作语法分析),但毫无疑问,进行概念之间的联想肯定是人类理解过程的第一步。前面我们提到了同行优先和交链式关联优先的联想方式(这是层次网络符号的独特贡献),这当然非常重要,因为,虽然它只是最低级的联想,但却是进入中级联想的起点。什么是中级联想?就是我在前面说的——判断“这些优先组合的概念能否构成一个完整的意思”。“完整的意思”这个提法借用了语法学的术语,不过,语法学并没有对这个提法给出明确的定义。这是明智的做法,因为这个定义很不好下,读者反正可以心领神会,这就够了。但我们的读者不是人而是计算机,它没有心领神会的本事,因此,我们不得不为“完整的意思”粗略地勾划一个“思考”的步骤和范畴。它应该包括下列 7 个方面:

(1) 联想句类(即句类设定)。

(2) 联想语义块的个数,它是句类的函数。

(3) 联想语义块内部搭配的协调性和语义块之间要素搭配的协调性。

(4) 联想基本逻辑概念,这是引向深层理解的重要线索之一。

(5) 联想辅要素,它们是句类的‘泛函’。

(6) 联想基本概念,不同句类与基本概念有不同的特定联系。

(7) 联想要素的排序,这不仅涉及语法、语用和语言艺术,也涉及表达的侧重点,因而也是引向深层理解的线索之一。

这 7 个方面可作为‘广义完整性’的定义,而前 3 个方面可作为‘狭义完整性’的定义。语法学心目中的‘意思完整’,大体上与‘狭义’相当。从解模糊来说,关键在第三步。因此,能进行‘狭义’范畴的‘思考’已经相当不错了,但从理解的角度来说,还相去甚远。这 7 个方面可统称为中级联想空间,每一项联想可称之为一个联想脉络。我认为,人的大脑中存在类似的联想空间和联想脉络。这个联想空间内的操作过程显然不是固定的串行式或并行式,而是一种‘灵感’式或‘映射’式。上述 7 个方面的联想内容充分表现了这一特点。

从上面的说明可以看到,后面的 6 项联想除了第 4 项以外,都是以第 1 项为前提或参考的,所以我们说,句类设定是中级联想的起点或基础。实际上,第 4 项联想也不是与句类毫无关系,比如,状态句通常就特别需要这一联想。

下面,将句类与各联想脉络的关联性用表格的形式给出一个轮廓性的描述。

句类	要素个数	辅要素	基本概念
作用句	3	Ms,Wy,In	
作用效应句	4	同上	j4,j5,j6
效应句	2 3		同上
过程句	2 3	Re	j1
转移句	3 4	Wy,In	j2
关系句	2 3	Wy	
状态句	2 3		j1

这个表格可借用当前流行的术语,称之为联想菜单。这个菜单是句类或联想脉络的泛函,是句类知识的具体体现。不同句类联想脉络的差别,就表现为这个菜单的差别。我在上面说,人的理解过程不会作什么句类映射,但人的大脑中肯定有一个完善的联想菜单,并在思维过程中充分利用这个菜单的泛函性,即句类知识。

联想菜单中的具体项目并不是硬性规定,只是优先项目,应采用上面说的‘灵感’启动方式进行操作。所以,一个句类分析专家系统的运行机制必须遵循下列 3 条原则:各联想脉络独立运作;联想的具体项目依据小语境自动调整;联想的深度可深可浅。我不敢妄谈实施这些原则的难易问题,但我相信,按照这些原则沿着上述 7 条联想脉络进行‘思考’是通向高级联想的必由之路。人们早就应该清醒地看到这一点,可是,语法学的统治地位在客观上是一堵妨碍视野的墙。因为这堵墙的存在,我们至少失去了 20 年的时间,实在不应该再在旧框架里继续浪费时间了。

我大体上回答了你提出的前 3 个问题。在后半段,务虚的色彩多了一点,现在打住。第

4 个问题属于比较具体的问题，放在下一问答里连同务实的问题一起讨论。

问 32：

上一次问答我提出了四个相关联的问题，先生已谈了前三个。在总体思路上我感到十分清晰，关于语句‘完整的意思’的七项内容使我倍受启发。先生谦称尚未对句类分析进行假设的求证，但我知道先生正在同杜燕玲一起从事这项繁琐的工作。我有一个看法，只要所有的动词都能按先生的设想给出句类信息，这个求证实际上是不必要的。所以我建议先生不要在这方面花费太多的精力，不知先生是否同意这一看法。好，现在请先生继续上一次问答的谈话。

答：

无论是概念层次网络理论的体系结构本身或句类分析的实践，也许最容易引起争议的问题之一是作用和效应的划分。这首先涉及到效应节点的配置问题。效应网络的 11 个二级节点（不计 3-0 节点）与作用节点关联性的强弱有显著的不同。3-1、3-4、3-5、3-6、3-7 的关联性很强，3-3、3-10 的关联性最弱，其他的节点介于两者之间。那么，为什么不按照与作用节点关联性的强弱来排序？为什么不可以把关联性很强的效应节点作为底层作用节点？我可以举出一系列的理由说明当前配置的合理性，但是，如果换成另一种配置，我也不能完全否定它的合理性。概念二级节点的合理配置或最优配置问题十分复杂，目前还不具备作深入探讨的条件。但这个问题极为重要，因为，它关系到按二级节点建立的联想脉络的有效性。表面看来，这只是一个智力差异问题（人的智力差异，实际上就是联想脉络有效性的差异），当前谈不上造就电脑天才，能够做到让计算机思考起来就不错了。但是，我不希望费了很大的力气仍然是搞成一个电脑白痴或接近一个白痴，实际上这就是以往 40 年的教训。因此，我对于二级节点的配置问题一直是惴惴不安，但这个问题的最终答案只能在概念层次网络理论的实践过程中逐步解决。

上面我谈到了二级节点的配置会影响联想脉络的有效性，而没有提到一级节点。这就是说，我认为一级节点的配置已经是合理的甚至是最优的。具体到当前的话题，就是说必须把作用和效应分成两个一级节点。这里，我打算从语义结构和语句结构两方面对这个问题进行阐述。

在语义结构方面，我引入了作用型和效应型两种语义结构，分别命名为结构方程 4 和 5，并引入相应的符号 ↑ 和 ↓。与语法学 5 种组合结构（联合、偏正、后补、述宾、主谓）的经典提法相比，我引入的作用、效应、逻辑、内容（后两者分别相应于结构方程 3 和 7）4 种结构是全新的概念，是反映语言深层的概念，而汉语语法学家提出的‘后补’之类的概念，则是典型的语言表层的概念。下面用具体词例来说明引入这些新概念的必要性。

汉字的‘击、打’是典型的高层作用型概念，由它可以派生出许多复合概念。先看下列三组：

击溃、击毙、打倒、打垮、击（打）落、击（打）沉、击（打）败、阻击、拦击、截击

击毁、击(打)中、击(打)退

伏击、狙击、袭击

第一组是作用型组合结构,第二组是效应型组合结构,(注:这里说的两种结构实际上是指它能否直接转换成效应句,前者不能而后者可以。目前的知识库对此有明确表示。)第三组是逻辑型组合结构。先来分析第一组。它的第一行和第二行分别相应于正反两种作用结构,这里的"击或打"表示作用,第一行的第二个字"溃毙倒垮落沉败"和第二行的第一个字"阻拦截"表示在该作用下产生的效应。这是作用结构的正反两种类型,如果按照语法学的分类,第一组的第一、第二两行分别叫做后补和偏正,岂不是给人一种隔靴搔痒的感觉么?汉语里的这种正反对偶结构(我对汉语语法学未注意到这一点感到不可理解)蕴涵着极为重要的语用和类别知识。正作用结构的 v 特性极为突出,不能与"进行"连用;相反,反作用结构的 vg 特性十分突出,可以与"进行"连用。应该指出,上述类别及语用属性并不是我的推测,而是定义。由此可以看到,虽然语用知识一般只能在词汇级予以表达,但并不等于说,语用知识不能在更高的层次给出规律性的表述。当然,定义仍然存在合理性问题,但对上述定义,这一点是可以保证的。有趣的是,这个保证正好来于分别引入了作用和效应两个概念,因为,你可以把不符合上述定义的作用型结构放到效应型里去。

在语句层次,我对作用和效应的知识负载分工寄予更大的希望。为了便于下面的阐述,这里有必要先解释一下作用对象和效应对象、作用内容和效应内容的概念。而为了说明这些概念,又必须先弄清对象和内容的概念。

大家知道,基本的语句要素是 E、A、B、C,后两者分别命名为对象和内容,对象和内容又按作用效应链各分为 6 类。就对象来说,过程、关系、状态的对象比较简单,可以用"承载者"或"体现者"予以表述;但作用、效应、转移的对象则需要慎重地予以定义。转移的对象定义为转移"物"的接收者:转移"物",它是转移的"承载者"或"体现者",反而被定义成转移的内容。这样定义的主要根据是为了赋予对象和内容下列重要特性:对象不能扩展为另一个语句,而内容可扩展为另一个语句。从另一方面来说,语句的对象总是被影响者。一般说来,转移的直接被影响者是接收者或接受者,而不是转移"物",对"物"的影响仅仅是位置或形式有所变化而已。因此,关于转移对象和内容的上述定义是更好地理顺了知识基本类别的关系,并不违反常识。

上面我们以转移为例,说明了对于对象和内容的奇特定义方式。这个定义方式确实有些奇特,然而是关键性的。它先脱离这两个词的常规意义,仅从语义块的可扩展性给出最抽象的定义,然后参照它们的常规意义赋予两者以函数形式的范定。

回到作用和效应,它们的对象可定义为被影响者或接受者。按照这个定义,作用和效应的对象可以是任何事物,从具体的人和物到人的任何活动以至人的认识和观念。这完全符合常识,也无可非议。但理解要求建立联想脉络,我们必须把无可非议的"任何事物"的"任何"二字加以限制,否则就不能前进。施行这项限制的基点就是把作用和效应划分开来,区分作用对象和效应对象。有了基点,下一步的问题就是制定区分的标准。这个标准应该说

比较容易选择,这就是:(1)具体与抽象;(2)整体与局部。这样,就能形成下面的定义:

作用对象:具体及整体的事物

效应对象:抽象或局部的事物

这个定义的要害是"及"、"或"两字,在形式上给出了两类事物或两类对象的无模糊界限。有了两类对象,对复杂的 B 语义块就有了表述的手段。复杂的 B 语义块通常包括多项要素,但其基本骨架一定是由作用对象、效应对象和效应内容构成。这一点没有任何奥妙,因为,对于对象的充分说明不外乎具体与抽象、整体与局部这两大方面的两个侧面,而它们都已包含在上面的定义里。这就是说,复杂 B 语义块的构件清单已经明朗了,剩下的问题是它们如何排序。由于抽象从属于具体,局部从属于整体,如果给这个从属关系的双方规定一个顺序,这个问题也就解决了。例如中国人的思维习惯是从属方在后,即整体在前,局部在后;具体在前,抽象在后;表现在语句中就是作用对象在前,效应对象在后。这就是汉语 B 语义块组合结构的基本规则。说句多余的话,我们之所以能得到这个规则,就因为我们引入了作用和效应的概念。

说到这里,必须对关于两类对象"形式上无模糊"的提法给以说明。所谓"形式上无",意味着"实际上有",因为上述定义中的"事物"二字是广义的,因而所谓具体或抽象事物的界限是模糊的。虽然我们可以人为地规定一个无模糊的界限,但这个做法的利弊现在很难判断,因此不如先保留这个模糊,以留有余地。

现在回过头来看上列前两组词例,你就不难理解为什么它们分别属于作用型和效应型组合结构了。但是,如果你仅仅是套用关于作用对象和效应对象的定义,就感到万事大吉,那就大错特错了。首先,你必须进一步思考关于作用与效应不可分的观点。诚然,当我们用作用型词汇构造句子的时候,只需要给出作用对象,句子就是完整的,这就是"作用句为三要素句"这一基本规则的理论根据。但是作用对象之所以"溃毙倒垮落沉败",离不开具体的效应对象,作用是通过对效应对象的具体作用而产生最终效应的,这就是隐知识。从联想脉络来说,必须有这个隐知识的支脉,尽管在很长的时间里,这项支脉知识很可能是空集。

其次,你必须认识到:上述两类对象的定义应视为对复合对象的分类准则。当对象"复合"时,两类对象的区分才有确切的意义,并在一个句子的范围内不会出现模糊。当对象"单一"时,在许多情况下区分作用对象和效应对象是没有意义的,而且在一个句子的范围内会出现模糊。从字面的意义来说,把作用对象和效应对象的定义反过来亦无不可。你完全可以争辩说,作用对象是局部和具体的,而效应对象是整体和抽象的。以"击毙"为例,枪弹的作用对象只是人体的一部分,例如脑袋、心脏等,而最终的效应却是人的整体死了。但从另一方面来说,这个"击"所针对的对象是人的整体,而不是局部。两种说法都有道理,但后一种说法更符合作用效应链的总体设计思想。不过,我并不想争个"水落石出",这样做显然是不明智和徒劳的。我宁可说,上述定义纯粹是一个人为的约定。

上面"滚滚而来"的一连串抽象论述,读者也许很难适应,因此,在这里插入一个具体的示例。现代中国领导人喜爱的说法之一是:一手抓××,一手抓△△。这里的××和△△通

常是指党和政府的工作,但是如果你用‘整体’和‘局部’二词分别代入亦无不可。在这种情况下,你总不能说前者是作用句,后者是效应句。不要以为这是特殊情况,实际上人类高级智能活动(基元概念的 9 13 行)的对象大多数情况是带有抽象‘成分’的具体事物,这些事物的‘总体’或‘局部’性是相对的,并没有绝对的界线。因此,对人类高级智能活动区分作用句或效应句往往没有什么实际意义。说到这里,我们可以引用‘山穷水尽疑无路,柳暗花明又一村’这两句诗了。这个‘村’并没有什么奥妙,不过就是‘作用效应句’罢了。这就是说,表述人类高级智能活动的句子大多数是作用效应句。作用效应句中既有作用对象,又有效应对象,当两者在一个句子中同时出现时,模糊问题几乎就不存在了。上述关于两类对象的定义对作用效应句极为有效,不妨戏称为‘灵丹妙药’。(注:这里说的作用句和效应句是指仅有作用对象或效应对象的作用句,而作用效应句是指同时含有作用对象和效应对象甚至效应内容的作用句。现在的作用效应句专指作用与效应的复合句类。)

对作用效应句,接下来的问题是:什么情况两种对象是必须存在而不能省略的,什么情况两对象之一是可以省略的?这个信息由结构方程 4 的第二级予以表达。具体表达方式在〈问答 34〉中叙述。

但是,我们不能回避单纯的作用句或效应句。我曾在一个材料中提到:“惊涛拍岸”是典型的作用句,而‘卷起千堆雪’是典型的效应句。张全对此提出了不同看法,认为后者也是作用句。张全的看法是有道理的,因为‘千堆雪’完全符合作用对象的定义。说它是典型的效应句,是由于‘卷起’这个多义词在这里的层次网络符号是 v0-0 ↓ v3-1-1,但是,为什么不可以把这里的结构符号↓改成↑呢?这似乎是一种两可情况,其实不是。对这个问题的深入探讨,就必须从表达的角度来考察作用和效应的划分,这就是我们即将论述的第三点。

第三,你还必须认识到:作用与效应虽然不可分离,但却是作用效应链的两极,这一点与磁性的两极现象极为相似。作用是源极,效应是末极,这一点无可争辩。从表达来说,这更是泾渭分明的两极。站在作用极进行表达,作用者是不能含糊的,在语句中是理所当然的主语,而对象是理所当然的宾语。站在效应极进行表达,作用者是可以含糊的,语句中理所当然的主语是对象而不是作用者,而宾语可有可无。在这里,我使用了语法学的主语和宾语的概念,但我注入了新的观点,即认为只有站在作用效应链的源极角度,宾语的概念才有意义。这个观点来于汉语效应句的独特表达方式,但我认为,汉语的这种表达方式更符合语言深层的简明性,可作为深层表达的规范。这里不能不举一些例子来说明汉语效应句的表达特点。西语由于恪守主—谓—宾的语法规范,当作用者处于含糊状态、又需要站在效应的角度进行表达时,往往引入不定代词充当主语,这种做法不能不说是画蛇添足。例如,汉语说‘下雨了’,英语说‘It is raining’,这个 it 只有语法意义。汉语说‘房间打扫了’,英语说‘The room has been cleaned’,这里 has been 也只有语法意义。从这两个例子可以看到,当站在效应极进行表达时,汉语与英语相比,有两项省略,一项简化。省略的是形式主语和被动表示,简化的是对时态仅作粗线条的描写。这里不来讨论这种省略和简化的利弊,而要着重指出,省略的说法本身就是用西语的语法框架来看待汉语,是应该受到质疑的。到底是汉语省略,还是西语

多此一举？汉语说“玻璃窗打碎了”,“自行车修好了”,西语都要变成被动式,难道这是必须的么？大家知道,在语义深层要求得主被动的统一形式。这个统一,是将被动统一到主动。显然,当作用者含糊不清,不需要予以强调时,被动式纯粹是多余的形式。因此,在这种情况下,与其说汉语简化,不如说西语多余。

汉语严格区分作用极和效应极两种表达方式,这是汉语的根本特点之一。前一种表达方式导致标准格式的三要素作用句,后一种表达方式导致标准格式的两要素效应句:YB + Y 和 Y + YB。后一种格式与作用句的简化形式 X + B 很容易混淆。你在上一问中说到作用句和效应句的区分会给句类分析程序带来困难,我想主要就是指这一点。这个混淆确实十分严重。问题似乎不仅在于形式,而且在于 Y 的定义本身,因为,Y 容许↑和↓两种组合结构,上面例子里的“打碎”和“修好”就是↑结构。这样,效应句的概念似乎陷入了困境。然而,这只是一种假象,因为,从理论上说,应该存在↑和↓的复合结构,这一点也不奇怪,本来作用与效应是不可分的。符号↑的意义仅在于突出作用,即站在作用极进行表达;符号↓的意义仅在于突出效应,即站在效应极进行表达。这里的“混淆”不过是两者的中间情况。实际上,符号↑突出作用的意义只是结构方程 4 的一级意义。结构方程之所以分为 4 级,就是为了表达语义组合中的各种复杂情况,包括一部分语用知识。上面谈到作用效应句是符号↑的一个特殊情况,将标记在结构方程 4 的二级表示里,对这里的“模糊”情况也将作同样处理。

上面着重阐述了作用对象和效应对象、作用表达极和效应表达极的概念。这些概念对于建立联想脉络、跨入语言深层极为重要。这些概念的细节还有待完善。我希望,读者不仅要深入考察细节,更要深入思考这些概念的基本前提:对象是作用效应链的函数。下面将对内容这个概念作进一步的阐述。内容同样是作用效应链的函数,不仅如此,它还是基本概念和逻辑概念的函数。

当概念层次网络理论正式公之于世的时候,某些粗心的语法学家可能会说:黄某人不过是把一些自然科学的概念引进到语言学里来,并别出心裁地制造了一堆似是而非的概念,他的 E、A、B、C 语句要素概念不过是谓—主—宾—补概念的翻版罢了。我倒是希望,本文的读者不妨随时联想一下这一假想的议论,以加深对语句要素和句类概念的理解。

上面曾给出了关于内容与对象相比较而表现出来的一项根本特性,即内容可扩展为另一个语句,而对象不能。这一特性是从语言深层去理解汉语兼语句和西语宾语从句的关键。

汉语的兼语句曾被视为汉语难以理解的怪物,其所以怪就是因为它违反了主—谓—宾—补的规范。可是,这个规范只是西语的表层规范,没有任何理由要求其他语言(包括汉语)也服从这个规范。语句要素的概念是在语言深层建立规范语句的基本概念。所谓语言深层的规范语句就是指 7 种基本句类的标准格式。

对基本句类标准格式的全面阐述是一本专著的任务,有资格的作者必须精通多种语言,我是不具备这一资格的。但对句类标准格式的基本原则的说明,我这个始作俑者则不能辞其责。实际上,在句类的语料尚未充分收集的情况下,过去我对句类标准格式的说明已经太冒进了,所以,这里只结合内容概念的说明探讨一些兼语句的例子。

张先生怕李小姐发脾气	X2B + X2 + X2A + XC
张先生怕李小姐的脾气	X2B + X2 + (X2A + XC)
张先生怕李小姐怕得要命	X2B + X2 + X2A + X2C
李小姐一发脾气,张先生就怕得要命	X2B + X2 + (X2A + XC) + X2C
张先生选定李小姐当公关部主任	A + X + YB + YC
李小姐当公关部主任是张先生选定的	
李小姐被张先生选为公关部主任	
李小姐警告过张先生不得草率从事	T3A + T3 + TB + T3C
李小姐对张先生说过不得草率从事	
不得草率从事的警告,李小姐事先就对张先生说过	

这里给出了三组例句,每一组的第一个句子都是典型的兼语句。第一组是反应句,第二组是作用句,第三组是信息转移句。句中的‘发脾气’、‘怕得要命’、‘当公关部主任’、‘不得草率从事’都属于内容。由特征要素‘怕’、‘选定或选’、‘警告或说’可唯一地决定句类。这三个句类的要素个数是不确定的,而这个不确定性的具体情况又与句类有关。就信息转移 2-3 和效应 3-8 来说,内容是必须的,否则句子的意义就不完整,但对象可有可无。就反应 6-0-2 来说,对象是必须的,但内容可有可无。这项关于对象和内容的知识是句类分析最重要的知识,是二级概念节点最重要的属性,也是概念节点关联性知识库最基本的项目之一。回到上面的具体例句,如果从句子中砍掉内容,根据上述知识,你就可以判断,第一组仍然是完整的句子,但第二组和第三组是不允许的,虽然主—谓—宾俱全,意义并不完整。当然,在口语或标题句里,像‘张先生选定了李小姐’或‘李小姐警告张先生’之类砍去了内容的句子非常普遍,但你必须明白,这个被砍掉的内容必然在上下文里有所交代。

二级概念节点所蕴涵的关于对象和内容的知识虽然极为重要,但究竟是高层知识,实际上往往只能当做启发性知识来使用,远不能满足句类分析的需要,因为具体的词汇常常是跨节点的。因此,在结构方程中,还必须对这项知识给出更精确的描述,这是结构方程 0 9 的 4 级表示的主要任务。

到现在为止,我们只说明了内容的根本特性并指出内容和对象一样,是作用效应链的函数,还没有给出内容的定义。这个定义需要区分状态句和非状态句。非状态句的对象和内容可定义为:对象是 v 概念的承受者,内容是对 v 概念的进一步说明。状态句的对象和内容可分别定义为表达的主体和对表达主体的进一步说明。状态句和非状态句的鉴别是句类分析的第一道关口,这个问题将在以后阐述。这里似乎暴露了作用效应链说法存在某种缺陷,为什么这个链条的状态环节要特殊对待?这是由于状态的表达可以没有 v 概念,过程表达也有这种情况,不过,不像状态那么突出。所以,上面的定义如果改成按表达是否需要 v 概念来说明,也许更好一些。然而,这并不重要。重要的是,上述关于内容可扩展为另一语句的特性是普遍适用的。

关于对象和内容在高层方面的要点,上面大体上都说到了。至于它们在低层方面的知

识,例如,C 的细分,A、B、C 的并合,C 扩展语句的特点及扩展的条件等等,将在其他适当的问答里详谈。这里只结合第一组例句说一下 C 的细分。首先,应该说明,反应有主动和被动之分。主动反应预定用组合符号 0-2↑0-0 来表示,主动反应的内容与作用内容是一回事。通常所说的反应内容是指被动反应内容,用 X2C 表示,作用内容用 XC 表示。在第一组例句中有意安排了两种内容,并用()表示 C 与 A 的并合,这些由读者自己去体会。

关于作用与效应、作用对象与效应对象、对象与内容、作用句与效应句,这次就谈到这里。也许没有完全解答你的问题,我期待着我们的讨论能向纵深方向发展。

最后,你关于句类求证的说法,我从内心是同意的,因为我相信知性的力量。但我在求证过程中还有其他的探索目标。谢谢你的关心。

1994 年 7 月 21 日于庐山

问 33:

先生在谈到句类分析时,总是对状态句另眼相看。在〈问答 27〉中,先生对第二个模块,就把基本状态句和基本判断句作为句类初判的一个特殊分支来处理。为什么要这样作?

答:

有这么一个英汉翻译系统,参观者出了一个最简单的题目"The earth is round",它翻译成"地球圆"。意思是对的,但作为一个汉语的句子,未免太别扭了。这表明该系统对状态句和基本判断句的表达缺乏最基本的知识。这不能责怪系统设计者,因为语法学根本不涉及这些概念,当然谈不上状态句和基本判断句的知识。

在讨论状态句之前,不能不说明一下状态网络二级节点的设计。关于状态的二级节点配置,最容易想到的是形态、结构、层次和等级。但是,状态与哲学上的存在是密切相关的,而潜在的存在性及存在的潜在性是两个非常重要的哲学概念。在汉语里,这两个概念叫做"势"。汉语对"势"的表达很有特色,中国传统哲学对"势"的理解也比西方传统哲学深刻。柳宗元在论述封建制(这不是现代的封建概念,而是指分封制)的产生时,只用了一个字:"势"(原话是两字句"势也")。你看,在一千多年前,这个认识是多么深刻。由于这个心情上的因素,我把形态、动态和势态放在前面,把结构、层次和等级放在后面。

5-0	一般状态
5-1	形态
5-2	动态
5-3	势态
5-4	结构
5-5	层次
5-6	等级

状态句的根本特点是存在无特征要素的两要素句 SB + SC。其他句类的省略形式可以

出现两要素句,但省略的只能是 A、B、C 当中的一个或两个,而不能省略特征要素 E。因此,状态句与其他句类在理论上不会发生混淆。但问题在于,一些简单的命题既可以用两要素状态句来表达,也可以用基本判断句来表达。这不是句类的混淆,而是表达方式的两可。对于这种两可情况,西语一律用基本判断句形式,汉语则依情况的不同而分别选用状态句或基本判断句。上述译文的失当,就是误用了状态句代替基本判断句。

那么,如何掌握状态句和基本判断句的使用准则?这不仅是语义表达及语法规范问题,更涉及语言艺术问题,因此,制定严密的准则几乎是不可能的。但下述具有一定弹性的准则仍可供参考。(1)对于 SC 仅包含单一的 u 或 z 的情况,如果需要强调 SC,则采用基本判断句。反之,如果不需要强调 SC,则采用状态句。(2)如果 SB,SC 重复出现,或 SC 包含多项 u 或 z,则通常采用状态句。(3)现代汉语很少采用 2-1 结构的三字状态句作陈述句。

读者或许以为,我是专门为了"地球圆"的不当翻译而引入第三条准则的。为避免这一误解,这里不能不讲一点关于汉语的韵律性知识。古汉语的韵律讲究音节数和押韵。对这一语言艺术现象,不能用"士大夫文学死板规定"的恶谥把它一棍子打死。诗经被称为四言诗,它主要采用四字句,但诗经是中国古代的民歌大全。中国诗歌艺术的顶峰是唐诗。现代中国作家如果有一点唐诗的底子,其作品的名称就不至于俗不可耐,说明唐诗仍有强大的影响力。但唐诗是严格的五言诗和七言诗,即五字句和七字句。因此,可以说汉语的韵律优先四五七字句。这里应顺便说明一下所谓多字词的说法,实际上是很不科学的。五字以上的"词"都是句,四字词绝大多数是句,三字词的大多数是名称,符合句子标准的都很少。这是一个很有趣的迹象,表明三四之间有某种阈值的味道。以韵律美著称的诸葛亮"前出师表"可提供进一步的佐证,该文的四字句高达 43%,三字句仅占 4%,其他五到十二字句的比例分别为:14%,18%,15%,9%,4%,1%, 1%,1%。我猜想(正在庐山休息,无资料可查),如果对中国古文做类似的统计,其最终结果不会改变上列数据的基本规律,即中国古文的句子长度集中在 3 9,从三到四有一个跳变,九以后缓趋于零。现代汉语扩展了这条分布曲线,但并没有改变它的基本特征,三到四的跳变依然存在。为什么有这个跳变?这个现象值得研究,其主要原因似乎是三字句有其特殊的使命或功能。中国以前有名的儿童读物"三字经",都是三字句(当然其中有些是把六字句拆成两个三字句),这说明简单的三字句适合于儿童。70 年代流行的"深挖洞,广积粮,不称霸"显示了三字句的另一项功能,就是它适合于作标语口号。标语口号意味着强烈的感情。某些宋词词牌有较多的三字句,如"满江红"和"钗头凤"(著名的词分别有岳飞的"怒发冲冠"和陆游的"红酥手")。可以说,用三字句表达感情是汉语的独特风格之一,这在现代民歌里也比较突出。最后一点是三字句适合于口语和对话。总括上述各点,似乎可以说,三字句对语境的要求比三字以上的语句要严格得多。我认为,这是一项重要的知识。

关于不带特征要素的状态句与基本判断句的区别,暂时就谈到这里。基本判断句有三个子类:比较、是否和有无。最容易与状态句混淆的是以 u、z 为内容的是否判断句。至于像"鲁迅就是周树人";"百分之九十以上的中国人是汉族"这样的是否判断句,并不会与状态句

发生混淆。有无判断句也容易与状态句混淆,这个问题下文再谈。这里先谈一下状态句与效应句的区别。

我曾用下面的句例说明作用句、效应句和状态句的相互转化。

(1) 张三打扫了房间,这房间张三打扫过了,这房间张三扫好了,

张三把这房间打扫得很干净。

要打扫一下这个房间,要把这房间打扫一下,这房间要打扫一下。

让张三把这房间打扫干净,要把这房间打扫得干干净净。

(2) 这房间扫过了,这房间扫好了,

这房间扫得很干净,这房间扫得干净极了。

(3) 这房间很干净,这房间干净极了,

这房间干净得几乎一尘不染,

这房间干净得让客人惊叹不已,

这房间干净得可以制造集成电路了。

第一组是作用句(其第二、第四行是作用效应句),第二组是效应句,第三组是状态句。上一〈问答〉里所说的汉语两极表达,在第一、第二组的句例里有生动的体现。两极表达是针对作用与效应而言的,就整个作用效应链来说,是六极表达。现在我们来看一下效应极和状态极的区别。第二组的第一行是 YB + Y 形式的效应句,第二行是 YB + Y + YC 形式的效应句,第三组的第一行是 SB + SC 形式的状态句,这些都是标准形式,容易把握。但第三组的第二行却不那么容易把握,它是简单句类还是复合句类?如果是复合句类,它是效应状态句还是状态效应句?每一复合句类的语义块符号又如何选定?下面就来回答这些问题。

复合句类大体上相应于汉语的另一怪物——连动句,但从作用效应的观点来看,这一点也不怪。按照作用效应链的六个环节,理论上复合句类有 $6 \times 5 = 30$ 种,但从理论思考方式上来看,这个复合分类可以大大简化。简化的原则是将六极简化为"作用"和"效应"两极,这里的"作用"包括作用和效应;"效应"则包括过程、转移、关系和状态。这样,复合句类就并为两大类:作用效应句和效应作用句。实际上,我们以前定义的作用效应句就是在这个简化意义下使用的,因为,作用效应句的效应内容 YC 可以是任一句类。反过来,效应作用句就不能这么处理,因为,广义的"作用"和"效应"不是简单的对应,而是一对多的对应,是源与流的关系。因此,这里的效应作用句按其深层意义应分为:过程效应句、转移效应句、关系效应句和状态效应句四类。那么,过程、转移、关系、状态之间就不存在复合么?读者回顾一下作用效应链的观点就会理解,这种复合可以归并到作用效应句,不必另列。

在作了以上的理论说明以后,我想,前面提出的关于复合句类的问题已不言自明了。不过,对第三组第二行的句子类别似乎仍然有点疑问。它是简单状态句还是状态效应句?这取决于你对"干净得几乎一尘不染"这一语言结构的判断。如果你把"一尘不染"看成一个词,则整个句子是简单状态句;如果你把"一尘不染"看成一个句子,则是状态效应句。实际上"一尘不染"是典型的效应句 3-10-1,因此,这是一个状态效应句。如果你对这个结论感到

不习惯,不妨对照一下最后的两个例句,就不难理解了。

语法著作中通常都要讲到"把"字句和"被"字句,其实,更能反应汉语特色的是"得"字句。我觉得"得"字单用是效应句的语法标志或逻辑标志。这一点需要小心求证,希望读者一起来做这项求证工作。

在上面的例句中,我有意用同样的词汇构造了三类句子,以体现它们之间的相互转换,这种转换是不同表达基点(极)的必然现象。汉语不仅如上一〈问答〉所说,十分注意对作用与效应两极的区分,也同样十分注意其他各极的区分,这里所讨论的状态极和基本判断极,又是一例。汉语与西语在语法现象上的种种区别,必须从这个总的根源上去理解。这样,才能站在更高的角度去观察复杂的自然语言,避免削足适履的错误——用西语的语法框架硬套汉语的语言现象。

上面只讨论了最简单的状态句标准格式 SB + SC。下面对这一格式的内涵及其一般形式作进一步的说明。

大家知道,抽象概念都具有五元组的特性,因而都具有构造语句的条件。可是,句类概念中并不包含"基本概念句",那么,仅由基本概念构成的句子归入哪一类?答案是状态句。这个答案是非常自然的,因为,只有状态的二级节点 5-0-0 与基本概念挂靠。所以,由基本概念构成的句子一定是一般状态句。(注:这是典型的演绎失误,"一定"二字应改为"绝大多数"。)

状态句的一般格式是 SB + S + SC,S 是状态句的特征要素或特征语义块。若 S 属于 5-1 或 5-4,句类分析通常不会出现模糊,这就是说,形态句和结构句通常没有模糊。但一般状态句则十分容易与基本判断句和关系句 4-6 混淆。下面再对这个问题作一些讨论。

这项混淆主要来于汉语中的"是"、"有"两字特别是"有"字的意义过于宽泛,这是我个人对汉语的唯一遗憾。西语对"所有"、"具有"、"存在"三个不同的概念配置了不同的词汇,也较为注意它们之间的区别。汉语在这方面的表现就要差得多,例如:

(1) 张三有很多股票。　　桌子有四条腿。

(2) 张三有很多朋友。　　这桌子有一条坏腿。这桌子有一条腿坏了。

(3) 张三有作案嫌疑。　　桌子上有两只苍蝇。

第一组例句是关系句,其中的"有"是所有。第二组是状态句,其中的"有"是具有。第三组是基本判断句,其中的"有"是存在。多义词"有"的解模糊问题有一定的难度,因为它涉及较多的语言低层知识,但并不是说,高层知识对此毫无作用,恰恰相反,对"有"的理解,也应该从高层知识入手。下面就来对"有"字句作一点高层分析。

"有"字句的标准格式是:DB +"有"+ DC。"有"的具体意义通常可以由 B、C 唯一确定,即仅依赖于近程语境,而不依赖于远程语境。这一点对判定"有"的歧义十分有利。B 与 C 的情况有所不同:C 必须存在,但可以省略;B 则不仅可以省略,而且可以不存在。基于这一情况,可以制定相应的判断准则。但在叙述这些准则之前,应该对"有"的三项意义的距离(也许很有必要引入"义距"这一新词)略加说明。"所有"和"具有"的义距很小,"所有"域与

“具有”域实际上有重叠的模糊区,但两者与“存在”的义距较大,不存在重叠的模糊区,这就是英语的 there is 和 has 一般不能混用的原因。当然,细分起来,“存在”与“具有”的义距要小于“所有”。英语的 has 本来只有“所有”和“具有”的意义,但由于“存在”与“具有”的义距较小,英语有时也赋予 has“存在”的意义。下面建议的准则将利用这些义距知识,同时利用关于概念抽象性及具体性的高层知识。

(1) 如果 B 不是省略,而是不存在,则为“存在”。

(2) 如果 B 由 l1-5 所范定,则为“存在”。

(3) 如果 B 属于抽象,则优先于“存在”。

(4) 如果 B 属于具体,又不含 l1-5,则优先于非“存在”。

(5) 如果 B 和 C 都属于具体,而且 C 是 B 的一部分,则优先于“所有”。

(6) 如果 B 和 C 分属于具体和抽象,则优先于“具有”。

提出上列准则的目的主要在于示范,演示运用高层知识的一般方法。这些准则的置信度和完备性都有待论证;它们的软件化过程,还有大量深入细致的工作要做。

汉语的状态句表达有其弱点,所以,我过去对状态句的句类分析保持低调。写到这里,我感到,汉语状态句句类分析的困难比过去设想的要小一些。不知你有同感否?

1994 年 7 月 26 日于庐山

问 34:

先生将语义结构方程作为词语知识表示的总框架,并认为它对于汉语尤为适用。语义结构方程的软件实现已进入设计阶段,因此,请先生就语义结构方程作一次专题谈话。

答:

作为概念层次网络理论三大组成部分之一的语义结构方程,我以前在不同场合谈过多次,但每次都只涉及部分内容,这一次希望进行全面、系统的论述。但由于语义结构方程的内涵仍在发展之中,在全面性上,有些只能点到为止;在系统性上,有些仅以提出问题、说明要点为目标。这里,先给出全部问题的清单。

1. 语义结构方程的多级表示。

2. 类别符号的组合运用。

3. 三字词的结构表示。

4. 四字词的结构表示。

5. 多(5 以上)字词的结构表示。

其中的问题 2 在〈问答 30〉作了专门讨论,所以,下面只讨论另外四个问题,分为两个小题目。

1. 关于语义结构方程的多级表示

在王华的硕士论文里,语义结构方程多级表示的“多”约定为 3。前两级为结构的层次

性表示,第三级为正反性(对偶性)表示。后来加了一级层次性表示,于是就成了现在的不协调安排:1、2、4 级为层次性表示,3 级仍为对偶性表示。当然,把次序和名称调整一下都很容易,不过,保留现状亦无不可。

王华的论文给出了 1 级表示的定义。后来,在内容上加了 9 号方程,在次序上也作了调整。现在的安排如下:(注:本文里的'X,Y'代表两项概念,与句类符号无关。)

0 号	X + Y = X ;Y ;(X ;Y)	假组合
1 号	X + Y = (X,Y) ;(X ;Y) ;(XY)	并、选、交
2 号	X + Y = X/Y ;Y/X	属性说明
3 号	X + Y = (X,l,Y) ;(Y,l,X)	逻辑
4 号	X + Y = X↑Y ;Y↑X	作用
5 号	X + Y = X↓Y ;Y↓X	效应
6 号	X + Y = X→Y ;Y→X	对象
7 号	X + Y = X|Y ;Y|X	内容
8 号	X + Y = ∑(Xm + Ym)	多组合
9 号	X + Y = X‖Y ;Y‖X	主谓

与语法学的组合理论相比较,这里的逻辑、作用、效应、对象、内容是全新的概念。对这些概念,已在前面的问答里作了详细的阐述。这里需要补充说明两点。第一,上列排序仍然不够协调,应将 8 号调整到 0 号位置,其他依次后推,这样,3 号以后都具有对偶性。8 号方程从整体来说,无所谓对偶性,只好虚取默认值 0。第二,8 号方程在形式上要求前后两字有同样数量的多个义项,但实际上组合后的某一义项可能只决定于两字之一的义项,这就需要对另一字的义项补零。另外,(Xm + Ym)的写法表示各义项的组合方式未定,因此 Xm 或 Ym 一般带组合符号。Xm 带,表示正向组合;Ym 带,表示反向组合;两者都不带,表示取默认组合方式(Xm,Ym)。从定义可知,由 8 号方程组合的词一定是多义词,但多义词不一定来于 8 号方程,因为,其他方程中的 X 或 Y 也可以有多个义项。

层次表示的特点之一是:低层特性是高层的函数。因此,结构方程的 2 级表示是 1 级表示的函数;4 级表示又是 1、2 级表示的函数。软件设计必须适应层次表示的这一特点,但并不是说,这样就无巧可偷。特别是 4 级表示,如果不赋予它某种独立性,显然过于繁琐。这就是下面讨论的主题。

2 级表示对 1 级表示的依赖性是难以割弃的,基本上只能老老实实。但各号方程的情况有所不同,按情况的接近程度,可分为(0、1)(2、3)(4、5)(6、7)四组,8、9 自成一体。我们从(4、5)(6、7)两组谈起。(4、5)组只涉及动词,而且一定是及物动词。我在前面的〈问答〉里讲过,及物性是一个表层的粗浅概念,从理解的角度来说,重要的是"及"什么物?"及"几重物?这当然是形象化的提问。严格的陈述是:在作用对象和效应对象、在对象和内容的选取方面,它是只需要其中之一,还是两者都需要,或两可?在对象和内容的具体化方面,它是优先于人、物或事,或两可,或三可?这里的问题涉及语句完整性和语句合理性的判断,涉及联

想脉络的主干。如果你已熟悉概念层次网络理论关于对象与内容、关于作用对象与效应对象的概念(详见〈问答 32〉),你就不难对这些问题的极端重要性有所体会了。

这里总共是三个问题。对前两个问题假设以最简单的方式来回答,每个问题需要 3 个答案,第三个问题需要 7 个答案。前两个问题是相互制约的,两者的答案总数不是 6 而是 7。正好是两个 7,分别安排在 2、4 级表示里。具体约定如下:

2 级表示	格式	
0		作用及效应对象都需要,效应对象不可省略
1		只需要作用对象
2		只需要效应对象
3	2-1	只需要内容
4		同 0,同时需要内容
5		同 1,同上
6		同 2,同上
7		默认值,含义不定(实际上是确定的,见下文)
4 级表示		
0		三可
1		优先于人
2		优先于物
3		优先于事
4	4-1	优先于人和物
5		优先于人和事
6		优先于物和事
7		优先于某一或某些中、低层节点

通过上列约定,实现了层次性知识的局部分离。细心的读者能够想到,4 号方程的 2 级表示里应无 2、3,而 5 号方程的 2 级表示应无 1,否则就是错误的数据。或者,将来利用这一点作其他的表示。4 号方程的 2 级表示若为 0,则必须是作用效应句,否则该句就不完整。这样的推论可以列举很多,不一一枚举。读者还应该想到,这里的局部层次性分离是不完整的。当 2 级表示为 0 时,4 级表示是指对象还是指内容?这一点是模糊的,我们默认取对象。最后一点,这里 4 级表示里的"事"包括基元及基本概念。

(6、7)组的两字组合(下面将简称为 6 号词和 7 号词),多数情况不是通常意义下的词,而是短语。也可以说,6 号词和 7 号词具有短语和词的双重特性。作为短语,意味着 6 号词已包含对象,用它构造句子时,待补充的只是内容。同理,7 号词意味着已包含内容,用它构造句子时,待补充的只是对象。这些是非常重要的信息。随之而来的问题是:对一个具体的 6、7 号词是否需要补充?答案有:是,否,两可。作为通常意义下的词,6 号词一定是表示"事",也就是说一定是含抽象意义的名词。7 号词如果是名词,同 6 号词一样,也必然是

“事”,但它也可能是 v 或 vu,这时,它也就不具有短语特性。表面看来,6、7 号词的上述双重性可以分别用结构方程的 2、4 级进行表示,似乎一切都顺利而自然,实际情况当然并不是这么简单。这里应该插进来谈一下结构方程 2、4 级表示的基本使命,也就是在结构方程中引入多级表示的设计目的,主要是下列两项:第一是给出语句完整性的知识;第二是给出类别水平的关联性知识(这里的类别有其特定的意义,已在〈问答 31〉中阐述)。第二项仅仅是初级联想的高层知识,但先迈开这么一小步比较现实,对解模糊也会有显著的效果。那么,词汇级关联性知识的理想表达方式是怎样的呢?是给出有关的中层概念节点。显然,不是每一词汇都能给出相应的节点,但是,可以通过低层节点设计的逐步优化,使可采用理想方式的词汇逐步增多。这当然只是一个设想,但其前景十分诱人,不能不在这里提一下。

当多级表示的‘多’具体化到 2、4 两级之后,自然的分工似乎是由它们分别承担上述两项设计目标。上面对(4、5)组 2、4 级表示的具体约定就是根据这一分工原则。但是,两项设计目标的有关知识并不是彼此独立的。由于 2、4 级表示空间的限制(各为 3 位),许多情况不得不另作约定。例如,当(4、5)组的 2 级表示为 0 或 4,即作用对象和效应对象都需要时,就需要约定 4 级表示属于作用对象。我并不认为扩大多级表示的空间以免除这类约定是适当的对策。面对这一类的模糊问题,遵循语言天然顺序的原则不是更好的办法么?先对象,后内容;先作用,后效应;先 v,后非 v。这就是 4 级表示的优先原则。

回到(6、7)组的 2、4 级表示。显然,是否需要补充对象或内容的问题属于语句完整性的范畴,应纳入 2 级表示。为此,引入下列约定,并命名为格式 2-2:

4　不强求补充,即上述答案的‘两可’。

5　不需要补充,即上述答案的‘否’。

6　需要补充,即上述答案的‘是’。

数字 0　3 仍沿用格式 2-1,用于表示 7 号词不具有短语特征的情况。读者必然会问:4、5 号词一定是动词,而且一定是作用效应型的动词,2-1 格式是以这两点为前提而制定的,怎么能够直接移用于 7 号词?这个问题当前只能给出部分答案:当 7 号词含 v 时,不论该 v 是否为作用效应型,上述约定都适用,因为 v 至少需要内容。如果 7 号词非 v 类,上述约定就需要另作解释,因为非 v 类词蕴涵的语句完整性知识较少。非 v 类词的行间关联性知识如何利用结构方程的 2、4 级表示进行表达,是一个有待考虑的问题,暂不作定论。目前,仅在二级表示中取含义不定的默认值 7,以示区别,这时,4 级表示的约定另行定义,见下。这样,2 级表示的模糊问题至少在形式上已完全消除。当 6、7 号词多义时,通常选取频度较高的一个义项作 2、4 级表达,这当然是一条普遍原则。

4 级表示虽然在理论上密切依赖于 2 级表示,但上面的说明已暗示 4 级表示可简化为两种规格:v 类规格及非 v 类规格。上面给出的(4、5)组 4 级表示规格实际上是 v 类规格的统一形式。对于非 v 类,可采用下列统一规格,并命名为格式 4-2:

0　优先于充当非特征主要素

1　优先于充当 A

2　优先于充当 B

3　优先于充当 C

4　优先于充当辅要素

5　优先于充当语义块的非核心成分

以上通过对(4、5)(6、7)组结构方程 2、4 级表示的具体介绍,阐述了构造结构方程多级表示的一般方法。其他各组的情况要复杂一些,因为它们都各有自己的二级组合规则。这些规则必须纳入结构方程的多级表示里,于是就出现了以两级表示表达三类知识的矛盾。解决这个矛盾最直接的办法是扩大多级表示空间,但词义库规范结构已成定局,暂不考虑。扩展词义库的表达能力,不能只依靠扩大单元空间的方法,这个问题以后再作通盘考虑。

下面先讨论(0、1)组和 8、9 组的情况。这些组的二级组合规则可简化为三种答案:记为甲、乙、丙;语句完整性知识简化为两种答案:(B,C)、(B;C),意思是:需要对象和内容、需要对象或内容。并进一步约定:(B,C)答案只与甲答案同时出现,这只是一个对二级组合规则的调节技巧问题。于是,可为上述各组制定下面的 2 级表示编码方案,并命名为格式 2-3:

0　甲,(B;C)

1　乙,(B;C)

2　丙,(B;C)

3　甲,(B,C)

4　甲,非 v 类

5　乙,非 v 类

6　丙,非 v 类

7　默认值,意义不定

其 3 级表示则可沿用前面的两种统一格式。关于甲、乙、丙的具体含义,用下面的表格作一个交代。

	0 号方程	1 号方程	8 号方程	9 号方程
甲	(X;Y)	(X,Y)		
乙	X	(X;Y)	(暂缺)	(暂缺)
丙	Y	(XY)		

最后,谈一下(2、3)组的 2、4 级表示。这一组结构方程的二级表示比较复杂。3 组的二级表示已定,恰好占满了 2 级表示的全部空间。2 组的二级未定,但估计其复杂度也会占满 2 级空间。因此,需要考虑的问题只是:是放弃语句完备性知识,还是放弃类别性知识,我们决定选择前者。3 级表示仍采用上述两种统一格式。当具体的词有多个义项时,采用哪一种格式,按上述天然顺序原则处理。

2. 关于多字词的语义表示

多字词语义库即将进入设计和建库阶段。希望通过本节的说明,能推进这项工作尽快上马。

从组合结构来说,多字词的字间组合不外乎语义结构方程所规定的 10 大类,但多个组合结构之间还有并合的先后顺序问题,或括号问题,或组合方式问题。下面就先来讨论三、四字词的组合方式。

三字词有四种组合方式:

(1) 1-2 方式,即第二第三字先组合,然后与第一字组合。

(2) 2-1 方式,即第一第二字先组合,然后与第三字组合。

(3) 1-1-1 方式,即组合无先后,字间组合结构只能是"并、选、交"之一,即 1 号结构方程。

(4) 3-0 方式,或称虚组合,其语义适合于直接表示。

四字词的组合方式采用两层次表示为宜,第一层次有四种组合方式:

(1) 一个双字词与两个单字词组合,有三种二级组合方式:

1-2-1

1-1-2

2-1-1

(2) 两个双字词相组合,只有一种组合方式 2-2。

(3) 一个单字词与一个三字词组合,有两种二级组合方式:

1-3

3-1

(4) 无先后顺序,有两种二级组合方式:

1-1—1-1 组合结构只能是 1 号结构方程

4-0 其语义适合直接表示

除了组合方式以外,还有多字词在语句中的作用或地位问题,也就是前面所阐述的 2、4 级表示的基本使命问题,或简称为语句完整性知识及类别知识,而语句完整性知识也可称为句类知识,两者可合称外联知识。这里,不能不插进来对"语义"这个词说几句话。我在本文中所说的语义,或语义库的语义,实际上超出了通常意义下的语义,是广义的语义,因为它包括上述外联知识。外联知识和一个词通常意义下的字面知识是有区别的。所谓"字面"知识是指一个词的内部知识;内部知识具有低层性和深层性,表达和把握都比较困难。而外联知识则具有高层性和表层性,比较容易表达和把握。这两类知识对句类分析和解模糊的作用很难说孰轻孰重,也许在初级联想阶段外联知识的作用更大。这就是当前确定的词义库建库步骤的主要依据。在这一段插话的最后,我想说:也许对上述广义的语义用一个新词"语意"来代替更好一些,不过,这种情况太多了,不胜应付,内容、对象、作用、效应等等,在本文中使用时,常有其特定含义,不可能在每一处都加以说明,只好请读者留意了。这一困境使我想起季宏在对这类新词作英译时采取的一个办法,就是把第一个字母大写,可惜汉语没有这个便利。

现在来讨论三、四字词的外联知识。我似乎在什么地方说过,所谓四字词,大部分是语

句而不是词。三字词也有语句的情况,不过较为少见。一个多字词,是否具有语句的特性,显然是一项极为重要的知识。但多字词语句,有完整性和独立性两方面的属性,这一点又与词相似。例如,“岂有此理”和“依然故我”都可以独立成句,但都有所欠缺,不够完整,而且两者的完整性还有很大的差别,因为,“岂有此理”通常只对“彼”而不对“我”,而“依然故我”则不分“彼”和“我”。

作为语句,当然所属句类的知识是基本的。这项知识的表达,建议直接采用前三级层次符号,但表达空间用两个字节,余下的半个字节用来标志该语句的“陈述、疑问、祈使、感叹”性。所属句类知识从某种意义说是冗余的,因为它可以从内部知识推断出来,但这项冗余是必要的,它不仅有助于提高多字词处理软件的效率,而且可作为句类分析软件的检验工具。从语言知识表示的策略来说,冗余常常是必要的,这里就是一个明显的例子。不难想像,大脑知识结构的奥妙之一也许就在于它的冗余机制。

除上述冗余知识之外,多字词无论作为语句或词来处理,其外联知识的基本构成是一样的,仅在细节上略有差别。所谓外联知识,实际上就是上一节详细介绍过的 2、4 级表示的知识,只是说法上有所变化。上一节的说法是:需要什么?这里的说法是:欠缺什么?如此而已。这里换一个说法是希望引发读者对外联知识的深入思考。外联知识的概念感觉上似乎很清楚,可是写出来的东西却像“夹生饭”。在庐山这样美好的环境里都未能写出自觉满意的定义,只好放弃了。庐山之写到此结束。感谢季宏和杜燕玲赶在我离京之前为我装配了带硕士卡的便携式计算机,使我在这里度过了许多个化难耐为充实的凌晨,并以喜悦的心情赏望从五老峰飘来的曙光。

在转入对多字词内部知识的讨论之前,对外联知识再作一点补充说明。上一节给出的五种格式外联知识受到表示空间的限制,但多字词的外联知识表达可以不受这种限制,而重新建立格式。这件事,由张全来考虑。

多字词的内部知识主要是下列三项:

(1) 组合方式

(2) 每个字的特定字义

(3) 组合单元内部的结构符号

像双字词一样,字义用义项号表示,结构符号用结构方程的序号表示。这里应强调指出:组合方式起着单元之间结构符号的作用,因此,不需要另行给出单元间的结构符号,而且上述冗余知识还给出了更明确的结果。其次,结构方程序号只需要给出一级,因其二级信息已体现在组合方式和字义里。

多字词的语义表达,也像双字词一样,不需要全部用结构方程的方式给出,有的可以或必须用直接表示的方式。“可以”的例子是名称,“必须”的例子如“越来越”,其层次网络符号是 uu5-3-1-3-9。组合方式中的 3-0 和 4-0,将全部用直接表示。结构方程方式和直接表示方式的选择涉及多方面的问题,原则上不必多谈,在建库的实践过程中酌情处理。直接表示的“地点”,不要限定在词义库,也可以是字义库。

多字词语义库的主要问题,上面大体上都谈到了。不过,还有一系列的问题有待讨论。下面,把想到的问题都列举出来,但有些只能点到为止。

(1)关于将多字词分为“词”和“语句”的提法,显然会引起争论。这只是一个工程定义。这里的“词”包括词、短语、语义块;这里的“语句”按西语语法框架一定是有缺省的语句,但至少包含两个语义块。对多字“词”的分析相应于语义块分析;对多字“语句”的分析,相应于句类分析。软件设计必须考虑到这一点。是否将多字“词”再分为词、短语、语义块三小类?不必,这纯粹是劳而无功。是否在多字“词”与“语句”之间就不存在模糊?当然存在,但这里不存在两可处理的问题。一方面总可以人为地作出选择,另一方面在“词”和“语句”的独立性知识中包含着两者的模糊信息。独立性强(0 级)的“词”具有语句特征,反之,独立性弱的“语句”又具有“词”的特性。

(2)成语要不要特殊处理?当然要。

成语的特殊性在于它的字面意义可能不提供什么信息,例如“叶公好龙”,重要的是它的比喻意义及引申意义。章含之为周恩来当翻译时,对“越俎代庖”不知所措,引起周恩来的莫大感慨(注:因为章含之是国学大家章士钊先生的女公子)。岂能对机器提出过高要求?当然,花点功夫可以做到让机器对成语的理解强于章含之,但不是当务之急。当前的目标只能是下列两项“基础建设”:

第一,按层次符号 7-11(注:这是“语法”概念的原表示方式,后将数字 7 改为字母 f,将“语法”概念定义为 f 类概念。7-11 的意义是“修辞”)的定义,给出语句类型。这个类型要与句类分析的类别区分开来。这里的语句类型实际上就是成语类型,7-11 的原定义域为 0 到 6,加数字 7 代表非成语。原定义的 6,可改为“典故”,或加 8 表示亦可。

第二,成语一定是“语句”,上述的冗余句类知识表达了成语的部分非字面意义。这一项知识采用三级层次符号,多余的表示空间很有利用价值。也可以考虑现在就采用四级(两字节)表示,最后一级暂时虚设,供以后知识表示升级使用。

(3)词根或非单字语素知识的利用。从技术上看,这个问题最为重要又比较复杂。这个问题先给读者一点感性认识比较合适,所以不妨举两组例子。

第一组　　大有人在,大有作为,大有可为,大有文章

第二组　　过甚其词,闪烁其词,隐约其词,支吾其词

这里的“大有”和“其词”我们称之为词根或语素,这个称谓并不科学,但这是语言学的通病,只能迁就。词典决不会把这类语言单元列为词,因为它们似乎不符合词的自由运用定义。但是,自由或不自由的说法过于模糊,所以我最不喜欢这个概念,而代之以独立性(当然也可用自由度)的概念。像“夜未央”里的“央”义项那样绝对不自由的情况是罕见的,大多数情况是半自由,这就是我们对独立性采用四级表示的依据。这里的“大有”和“其词”有相当的自由度。例如对做过错事、说过错话的好人,你可以说“可叹其人,可悲其词”;对敢于直言的人,你可以说“勇哉其人,壮哉其词”。可见“其词”是具备“词”的资格的。我们曾努力收集语言学的“等外”词,但远不完全。张全的新词库可自动将“大有”变为词,但“其词”还做不到。

我在这里想说的第一点就是,希望能在多字词语义库的建库过程中一起解决这个问题。这个问题的重要性在于意义的表达方式。你不能每次碰到这一类的语言单元就得从零开始,人的大脑肯定不是这样。应该把“其词”之类的语言单元当做词一样处理。

这就是说,对多字词中按上述组合方式划分出来的非单字单元,不但要充分利用现有双字词语义库的已有知识,还要补充建立上述的知识。因此,对非单字单元,有两种知识表示方式。一种是上面说的单元内部结构符号,另一种是指示该单元有现成语义知识可查。对后一种方式还要说明两点。第一,它并不限于双字词;第二,要标明双字词的义项号,因为相应的双字词可能是多义的。

(4)远搭配知识的利用

远搭配知识对四字词具有特殊意义。其中的 7-8-1-3 和 7-8-2-4(注:这是“语法”概念搭配的两种表示方式。)两种格式在四字词中极为常见。读者可能对这两种搭配格式很不熟悉,看下面的例子就会明白。“大吹大擂,大慈大悲,大模大样,大是大非,大手大脚,大摇大摆”是 1-3 格式的特殊情况;“大材小用,大醇小疵,大惊小怪,大同小异”是 1-3 格式的一般情况;“惩一儆百,杀一儆百,以一儆百”是 2-4 格式的一般情况。这种格式知识对于这类四字词的理解显然是关键性的,在多字词知识库中必须增加这项知识。

(5)多字词知识库的数据结构

我倾向于搞规范化。不可预测的扩展可放到字义库中,留一个扩展标志即可。当然,最终决策由张全酌定。

(6)五字以上词的知识库

上面的说明并没有限制词的数量,对五字以上词实际上已不需要另作说明。但需要提一下一个要点和一个特点。

一个要点是:组合方式要准备采用多重结构。例如“先天下之忧而忧”,第一步应分成 1-4-2 方式,第二步将“4”分成 2-2 方式,或直接写成 1-(2-2)-2 方式。组合方式的标定大体上相当于分词、分块处理。

一个特点是:五字以上词要么是语句,要么是名称,对后者一定采用直接表示方式。它主要是外国的人名、地名。

(7)唐诗和宋词知识库

举世无双的中国古代文学宝库可怜只剩下几本小说,一点唐诗、宋词和四字词的残迹了。我曾为这一时代悲剧写过“天地悠悠应泪下,古今忧患费思量”的诗句来纪念我的父亲。在深夜庐山默念“前出师表”,统计“句长”分布的时候,忆及父亲当年讲授此文音韵之美的情景,不禁热泪盈眶。我写这些话是想表明我的一个梦想:教机器写诗,写中国古诗。

古诗在某种意义上是对语言艺术进行了高度形式化处理,它虽然不完全是机器所要求的形式化,但似乎指出了形式化的根本途径,即上面所说的组合方式。在方式的约束下,对各语言单元又给出了长度、韵、类别性、句间对偶性、特征要素位置等的具体知识,机器不是正好在这些知识的运用方面,拥有巨大的潜力么?

作为这一梦想的“基础设施”,建议先作下列考虑:

第一,把一首古诗当做一个多字词进行索引和存储。

第二,建立句长格式知识库。

第三,在每一格式下,建立特征要素位置知识库。

第四,建立韵律知识库。

第五,建立句间对偶性知识库。

与我们的已建和在建的知识库相比,这些知识库是小菜一碟。

当前紧迫的事情甚多,以上的话很不合时宜,但它关系到词库和词义库的升级设计。这些话可能说晚了,果如此,不知能补救否?

本〈问答〉头绪较多,按常规应写一个小结,但这个小结由词义库设计者来做更为适当,我就不来“越俎代庖”了。

1994 年 7 月 28 日　8 月 12 日　庐山—武汉—北京

问 35:

最近,我一直在思考一个非常重要的问题,就是程序如何去提取和运用层次网络符号中所蕴涵的信息和知识。我的建议是:把这些原来以高度压缩方式表达的信息和知识还原成某种非压缩形式,并以函数或表格的形式建成知识库。这个知识库的功能不同于一般语义知识库,它提供知识、规则和推理三方面的信息。这样,理解程序运用层次符号的基本操作方式就是向该知识库发出询问,然后取得答案。不知先生是否支持这个想法。

答:

我基本支持你的想法。不过,首先要把同行优先用上。在确定句类和语义块的基础上,这件事很容易办。你设想的知识库涉及到一系列如何表示知识的根本问题。我只能就这些问题的理论方面作一些说明,以促进对这个问题的深入思考。为了叙述的方便,将把设想中的库简称为 U 库。(注:这个问题是我当时的研究生申凌提出来的。U 库之名来于此。在双拼中声母 sh 通常用 u 表示。)

1. 高层节点的链式关联知识

首先应该指出,概念层次网络理论赋予网络节点的意义与“语义”是有区别的。前者是对后者的净化描述。“节点意义”是完全抽象的,是对语言实际概念的分化、净化和深化,也就是去掉“个性”,保存“共性”。由于经过了“三化”处理,概念节点之间才出现了较为清晰的关联性。例如:

3-4		j4
3-5	5-0	j5、j0
3-6	1-0	j6

这是一组典型的链式关联。具体含义是:效应节点 3-4(扩展与减缩)与基本概念的“量与范

围‘强关联’3-5(建设与破坏)与状态和基本概念的‘质与序’强关联 3-6(推动与抑制)与过程和基本概念的度强关联。这种关联性是‘单向’的,效应节点是‘源’,非效应节点是‘果’。例如,(3-6-1)的高层概念‘推动’或‘推进’,可以与‘步伐’、‘进程’、‘进度’、‘速度’、‘趋向’等过程型概念优先搭配。这种搭配称之为‘内容’型搭配,以区别于‘对象’型搭配。这就是说,上述关联性指的是‘内容’关联性,不涉及‘对象’关联性。对象和内容是对语法学的宾语概念的语义分类,对象又按作用和效应分为两类。作用对象限于具体概念 w 和 p,效应对象限于抽象概念 g、r 或 w、p 的局部(详见〈问答 32〉)。在语义块的组合结构中,汉语的习惯排序方式是:(作用对象)(效应对象)(效应内容)。例如,‘推进中国经济改革的步伐’这句话就是按照上述标准顺序排列的。这里,‘中国’是作用对象,‘经济改革’是效应对象,‘步伐’是效应内容。这句话由两个语义块构成,第一个是特征要素语义块‘推进’,第二个是复合语义块‘中国经济改革的步伐’。这个复合语义块有下列 6 种简化形式(为了说明的方便,假定‘经济改革’是不再分解的词):

1 中国　　2 经济改革　　3 步伐

4 中国经济改革　　5 中国的步伐　　6 经济改革的步伐

值得注意的是,这 6 种简化语义块只有 2、4、6 可以与特征要素‘推进’构成合理的语句。这个现象与语种无关,传统的和现代的语法理论都不能解释这种现象,因为这是语义而不是语法现象。但用句类分析是容易解释的,因为这是作用效应句,这个句类的效应对象 YB 是不能省略的。

通过上面的说明,现在我们可以给出一个高层概念节点必备知识的清单:

(1) 与该节点有链式关联性的节点表

(2) 标明该关联性的类型(对象、内容或其他)及相关系数

(3) 该节点生成语句时不可省略的要素

这个清单的细节还有待深入,但总体轮廓是比较清楚的。对于高层节点,问题似乎比较简单,但是,低层节点和复合节点是否可以沿用这个清单?下面就来探讨一下这个问题。

2. 低层及复合节点的关联知识

上面我们谈到了‘节点意义’是经过‘去掉个性突出共性’处理以后的语义。但是,理想的‘去掉和突出’处理只适用于高层节点。对于中层及低层节点,必须有选择地突出某些语义个性,这是中层及低层节点设计的基本原则。其次,节点的组合就带进了个性,故组合结构符号被称之为语义的而不是概念的结构方程。不言而喻,个性的有选择表达是一个不能回避的难题。

人们在谈到语言时常说的习惯语‘语言太复杂了’,实际上就是指语义个性的变幻莫测。个性表现在许多方面,其中最为重要又比较容易把握的是个性在关联性上的表现,或称之为搭配个性。字义和词义的独立性指数,语义结构方程的 2、4 两级表示都是搭配个性表达的手段。字义的独立性指数实际上就是共性与个性的比例尺度。词义的独立性指数虽然引入了另外的意义,但比例尺度的信息仍然是隐含或显含的。这项知识是启发性的,它并不指明

具体的个性,但可以引导程序从字义库或词义库中找到搭配个性的部分或大部分具体表现。独立性和语义结构方程所表达的个性以义项或词为对象,完全抛开了个性的共性。而上述高层表示方式又完全抛开了个性。显然,介于两者之间的某种兼顾方式是必须的。

低层节点的设计尤其要考虑到这一兼顾方式的实际需要。要做到这一点,当然要以较为全面的资料为依据。从这个意义上说,第一期词义库可作为此项设计的有力工具。此项设计的理论原则并不复杂,就是将语义空间作正交分解。前面说的对象、内容分类就是对语义空间的一级正交分解。低层节点的设计实际上是这一分解过程的继续。让我们用两个例子对这一点作较为具体的说明。

例 1,效应节点 3-6。前面我们提到了,这个节点的高层没有优先的对象子空间,只有优先的内容子空间。但低层情况则又当别论。例如,人的精神状态 7-2-2 可以是 3-6 的对象,于是,以 7-2-2 为效应对象的 3-6 就构成了 3-6 的一个子集。有充分的理由将这个子集变为 3-6 的一个低层节点。它对应的语言概念包括'鼓励、激励、奖励、表扬、表彰、鞭策、鼓舞、振奋、勉励、批评、劝戒、警告、正告'等等。如果不设置相应的低层节点,这些概念就要写成复合形式(v9-2-3-9-2↑v3-6-i)→7-2-2;如果为之设置一个相应的 4 级节点,则上式就可写成 v3-6-i-k 的简化形式,而把它的严格表示式存放在 U 库中。这里,应顺便说明,上面的词例并不是都属于 9 行,'表彰'和'正告'就属于 12 行,在 U 库中可对此不加区分,这就是近似性。

例 2,劳作服务节点 6-6。在这个高层节点中,有两个意义特殊的子集,就是'农业生产劳动'和'烹饪'。前者以 jw6-1-4(农作物)为作用对象,后者以 pw6-2-1(食品)为作用对象,是作用对象中一个特殊的子集——生成物。按具体对象划分出来的这两个 6-6 子集,对 6-6 来说,它突出了'农业生产劳动'和'烹饪'的基本个性;而对于两子集的具体语言概念来说,它突出了两者的基本共性。这就是概念层次网络符号对自然语言符号的逐步近似法。这个逐步逼近过程毫无疑问应吸收信号处理中按本征值大小作正交分解的思路。我定义的主要素 A、B、C 和辅要素的手段(Ms)、工具(In)、途径(Wy)、条件(Cn)、因(Pr)、果(Rt)就是对语义空间作本征分解的本征值。就低层节点而言,U 库的使命之一就是对这些本征值和本征空间(相应的节点集合)进行说明。

现在我们已清楚看到,前述高层节点的链式关联知识清单不过是一般关联知识清单的一个特例。一个完整的清单就是对主要素 A、B、C 及各辅要素给出相应的节点表。我写下这些话只是想粗略描绘一下 U 库的终极规模。但 U 库的设计和建立过程一定要吸取字义库和词义库的教训,决不能一上来就贪大求全,而应该从当前工作的紧急需要出发,分期分批,逐步推进,上述终极目标只作为 U 库升级兼容性设计的依据。当前急需的是关于语句意义完整性、语句要素之间及语义块内部各组合成分之间是否协调的知识,这两类知识可满足解模糊处理及形成预期功能的基本需要。第二类知识就是我们常说的行内关联及行间关联知识。第一类知识主要是关于主要素的知识,其中,最重要的是下列 5 项:

(1) 哪项要素是必须的。这里,特别要注意作用对象和效应对象的区别。

(2) 哪项要素是可以省略的,并区分可语义省略和可语法省略两种情况。

(3) 哪项要素是不能独立存在的。

(4) 指明 C 要素必须、可能或不能扩展为语句的不同情况。

(5) 指明 A、B、C 要素的类别符号(层次符号属第二类知识)。

从上面的说明可以看到,第一类知识可称之为中级联想知识,第二类知识可称之为初级联想知识。我在〈问答 31〉中所说的 7 个联想脉络显然密切依赖于这些知识。

U 库的行间关联知识部分就说这么些"大而化之"的话,目的在于引发思考。显然,这项工作与低层概念节点的设计密切相关,因此,协同配合的水平将直接关系到这项工作的进程甚至成败,这是应该强调的。

3. 关于交式关联知识的表示

交式关联的表达表面上看起来比较简单,把存在交式关联的节点汇集成表就可以了。在初级联想的水平上运用这项知识去解模糊也比较简单。在语义块内部或主要素之间,如果说同行搭配是第一级优先,那么交式关联节点之间的搭配就是二级优先。这两个优先原则比任何统计结果都更为可靠。从语法的角度来看,这项搭配是词性的搭配,但我们把语法的"无条件"搭配变成了"条件"搭配。这个"条件"在交式关联的节点之间是有区别的。这个区别,主要表现在不是所有的 g,u,z,r 都可以共用,而主要是 u 的共用。对这个现象的精确刻画似乎必须放到词汇一级来处理。U 库的设计可暂时回避这个问题。

这就是说,概念节点之间的交式关联与实际语言概念的交式关联是有区别的,这个区别来于前面说的"三化"处理和它所带来的近似性。这个区别和不同语种之间的这种交式关联区别,是一个非常有意义的课题。语言对民族思维方式和民族心理的影响,或者说,思维方式和心理的民族性,肯定可以从这里找到富有启发的线索和规律。这些当然是后话(现在顾不上)。但在进行 U 库设计时,要留下民族性说明的空间,因为,我们不可能只针对理想化的概念节点建立交式关联知识库。建议将理想化的概念节点定义为 0 号"民族"。

上述二级搭配优先性,主要是 u 的共用性,是交式关联性在初级联想中的主要应用。但交式关联性更重要的应用是在中级联想阶段,句类分析时的混合句类判断和两可处理的选择都要利用交式关联知识。混合句类中,以作用效应句、作用关系句、效应关系句、转移(交换)关系句最为常见。与 0-0 节点(包括它的子集 0-2 和 0-4)交式关联的效应节点有:3-1,3-2,3-4,3-5,3-6,3-7。与 0-0 节点交式关联的关系节点有:4-2,4-3,4-4,4-7。由于这些节点与 0-0 交式关联,由它们所表达的语言概念就可以生成作用效应句或作用关系句。具体说,这些效应节点可生成 4 要素句,增加作用对象 B;这些关系节点也可以生成 4 要素句,增加作用内容 RC。由此可见,对交式关联节点,要给出是否可能增加要素及其类别的说明。这些说明是上述一般关联知识清单的一部分。

到此为止,我们提出了三类关联知识清单:链式关联知识清单、交式关联知识清单、一般关联知识清单。而前两者无论从内容和形式来说,都是后者的一部分,因此,似乎没有独立存在的必要。问题在于关联知识的层次性或个性,很难设想按不同层次的所有节点建立关联知识清单。高、低两个层次的划分是自然的选择。

4. U 库设计的若干技术问题

(1)关于高、低层次的定义。所谓高、低层次的概念只适用于基元概念和基本概念。逻辑概念采用(本体层 + 基元层)的结构,而本体层只有两层,无所谓高、低之分。对基元概念和基本概念,所谓高层,通常指层次符号的前两层,但下述两个情况是指前三层。第一是挂靠型复合基元概念 6、9、12,第二是非挂靠型复合基元。

(2)关于跨行知识的表示。所谓跨行知识是指若干行的共性知识。这些共性知识实际上已通过交、链式关联知识清单给予了详尽的说明,但这种说明方式是分散的,有"只见树木,不见森林"的弊病。因此,似乎有必要上升到行间层次进行总体性说明。有关人类活动的共性知识就是典型的例子。

人类活动语句的主语(虽然句类分析不采用主语的概念,但在这里主语的概念是必须的,这正好说明主语反映跨行的共性知识)当然是人 p。这项知识虽然极其粗糙,而且有模糊(因为人类的某些本能活动与动物的某些活动在语言概念上是不加区分的),但终究是一项有用的知识。

人类最重要的活动是专业活动 10 和社会性活动 12。两者强交式关联,共同以 p12 为优先作用对象,以 g9-0-0 为优先效应对象。这项共性知识,显然有必要上升到行间的总体描述。

在基本概念中,"间隔"和"变换"的概念都是跨行的,对这一类知识的总体描述是非常必要的。

(3)基本概念的特殊性。所谓"同行优先"和"交链式关联"的概念是针对基元概念引入的,虽然在形式上也适用于基本概念,但缺乏本质的意义。因为,基本概念并不是作用效应链的一环,而只是这一链式运作的条件。对基本概念来说,同行与跨行,在优先性上并无区别,交式与链式的差别也极为模糊,这与基元概念是不同的。因此,对基本概念没有必要分别建立交链式关联知识清单,用一个统一的关联知识清单就可以了。这个清单仍有必要分为高、低两个层次,但高层的定义与基元概念不同,不是指二级节点,而是一级节点。这就是说,对每一个基本概念,建立一个总清单,再选择某些节点,建立突出个性的清单。例如:

j1	1-0, 12-11, 5-2	
j2	2-0, 12-11, 12-9	
j3		z
j4	3-0, 5-0,4-0	
j5	3-0, 5-0	
j6	3-0, 5-3	vu
j7	12-10	g,x
j8	9-0-0, 13-0	g,x
j8-2		p
wj2-0-0	2-0-4	(TB1,TB2)

在这个清单里,我们希望反映基本概念的两种优先关联性:一是与概念节点;二是与概念类别。关于概念类别关联性,这里的 j3 与 z,即表示所谓数词与量词的优先搭配,j6 与 vu,即表示所谓程度副词的特性。这两点是语法学的常识。类似的概念类别关联性还有 j7,j8 与 g,x。基本概念是汉语中构成新词的活跃成分之一,原因就在于它具有很强的概念类别关联性。

从清单可以看到,虽然基本概念与任何基元概念都有关联,但优先性还是明显的。其中,时间与过程,空间与转移,量质度与效应,j7 与手段,j8 与事和行为,更是强相关。

清单的最后两行,是个性清单的示例。

在这一问答里,我试图说明 U 库的基本使命,哪些知识适合于纳入 U 库,哪些知识不适合于纳入 U 库,但意犹未尽,权作讨论稿吧。

最后我想说一点句类知识。

(1)不同句类要素的个数有所不同。例如,过程句、状态句、效应句和关系句可以是 2 要素句,但作用句和转移句则必须至少是 3 要素句。虽然在形式上各句类都可以缺省要素,甚至可以简化成单要素句,但从理解来说,你必须恢复缺省,因此,这个关于要素个数的约束是有意义的。进而言之,从要素或语义块的角度,不仅量的约束可以更为精化,而且要素的概念优先性也是明朗的。

(2)要素的表述重点有所不同。例如,作用句着重表述作用对象 XB,而效应句着重表述效应内容 YC。XB 一定属于具体概念 w 或 p,YC 一定属于抽象概念 g 或 r。这里,我故意用了"一定",而不用"优先"。

我的句类想法来于作用效应链这一基本假设。基本句类就是对作用效应链某一环节或一项基本判断的表述(基本判断是指关于比较、肯定或否定、存在性三者的判断)。混合句类就是作用效应链两个以上环节的表述。在这个假设下,句子这个语言的海洋就不再显得那么浩瀚无垠,而是一池清水了。这个说法,很有点胡言乱语的味道。不过,这是我在有了这个想法以后,再看各种句子的一种感受,就情不自禁地冒天下之大不韪把它写下来了。

第 五 部 分

语义学日记选录

1994.10.4

译‘上’。

“上”源于空间的上下,可作为等级、层次、品质、次序、趋向的对偶性分类标准,其中,对等级、层次、品质、次序用上中下作表述。趋向似乎无中,其实不是,是自然语言采用了其他的表述手段,例如,对趋向的上下有向上、向下、上涨上升和下落下跌等说法,但对其中间状态却采用‘固步自封’和‘平稳’等说法。

“上中下”用于三分对偶的表述。对偶有三分与二分之别,对此目前未加区别。汉语里的‘左右’常用于二分,例如左手和右手,左眼和右眼等等。

“上帝”一词的映射,需要作一点说明。宗教活动定义为 12-8-2,基督教的上帝和伊斯兰教的真主,显然不能映射为 p12-8-2(因为他们是神),而映射为(gp12-8-2, g5-5-13-0-1),而从事宗教活动的人则映射为 p12-8-2。gp 的意思是概念化的人,或虚构的人。小说人物也用此定义,当然挂靠层是 10-3-1(文学)。

1994.10.6

译‘身,深,神,审’。

“身”的映射符号是:jw6-3—,即身体。与它有关的词有:身材,身段,身长,身高,身板,身躯,身穿。

“身”有一个极重要的扩展义,就是自身,映射符号是 g4-0-0-5。身教、身受的‘身’就是此义。

“身”的上述两义项有不少四字词。

“深”字是汉字的精髓表现。“深度”的词典释义有三条,实际上只有一条,就是:jz4-2-12-2-2,把握了这一条,三条就融汇贯通了。许多概念都是这样,从基本或基元概念的角度去理解就比较简单。建立联想脉络,必须掌握这个基本要点。

“神”的映射符号是:gp12-8-0,从事宗教活动的人则记为 p12-8-2。那么,如何区分和尚、尼姑、神甫、牧师、阿訇、毛拉、道士、道姑的个性。这需引入挂靠层结束的标志符,将用‘*’表示。“*”之后的数字串是该具体概念的再分类表示。如 p12-8-2*k,约定 k=1 4 相应于佛教、基督教、伊斯兰教和道教。则上列各种宗教人士就不难通过组合方式加以表示了。这是具体概念精细表示的一般方法。

宗教是人的精神世界,所以,“神”字又表示人的精神,神采、神态、神色、神气等由此而来,并进而有神交、神医、神童、神人之说。

“审”的映射符号是:v12-2-1-9。审的目的是为了决策,两者链式关联,故有审批、审定之说。

1994.10.7

译“生,声,升”。

“生”的第一义项应该是 v3-1-1。3-1-1 的特点是,不必区分对象和内容。《现代汉语词典》将此义项列为“生 1”的 8,9 两项:8 项的释义是产生,发生;9 项的释义是“使煤柴等燃烧”,举例生火,生炉子。实际上,9 项应改为“使……产生”,释义者忽略了“生财”和“生事”这两个词就是这个意义。这就是说,“生”具有及物(作用)和非及物(效应)的双重性,这是效应概念的一般特征。

由“生”字以 7 号语义结构方程构成的词具有 0 级独立性,如果充当 E 要素,则通常为两要素效应句,如“生变,生病,生财,生火,生事,生效”都是如此。但是,“生气”、“生疑”如果充当 E 要素,则为反应句。

“声”的第一义项为 jw3,0 级独立性。另一义项为 r13-0,独立性应低于 2。由它构成的“名声,声威,声望,声誉”等词的意义虽以“声”为主导,但 r13-0 义项的“声”都应视为语素而不是词。

1994.10.9

译“绳;省”。

“绳”第一义项的一级近似可映射为 pw0-4。“绳索,绳子,麻绳,线绳”皆属此义。“绳”的另一义项为 v12-3-6-2,如“绳之以法”。〈问答 12〉中说过 3-6-2 与 0-4 强交式关联,“绳”字体现了这一点。第二义项独立性 1 级为妥,而第一义项为 0 级。

“省”字有三个 0 级义项,一是国家行政单位 pj2—0,二是节省 v3-11-2-1,与浪费 v3-11-2-2 对偶,三是(v3-1-2,jv7-7-1)。第三义项同时又有 r 特性,但语用上限于 jgw10-3 类概念。

1994.10.10

译“失”。

“shi”的每个声调都有许多 0 级独立性的字,在汉语的全部音节中可谓绝无仅有。

一声有“失,诗,师”。“诗,师”的意义比较单一,“失”的意义则大有学问。我们在效应节点里安排了 3-10,表示“获得与付出”。付出就是“失”,不过有主动的意义;社会意义的付出是 12-3-10-2,如奉献,牺牲,主动的色彩更为鲜明。“得”也有主动程度的差异,争取就是主动的“得”,应映射为 v9-3-10-1。由此可见,在 3-10 的前面加 9 或 12,表示加强了主动性,这是一般规律,用语法的术语来说,就是增强了及物性。

在关系节点里,我们安排了 4-6,表示“拥有和失去”,这就是说,得失的概念需要从效应和关系两方面去加以表述。从效应方面得到的东西,也就从关系方面拥有它;从效应方面付出的东西,也就从关系方面失去了它,所以 3-10 与 4-6 强交式关联。类似的情况有 3-8 与 4-5 等。

1994.10.11

译“事”。

“shi”的去声常用字最多,其中最重要的是“是,事”两字。

“事”就是人类活动,主要指 9 行以后的活动,但也包括生理活动的 6 行活动。词典里给了“事”字 6 条解释,都是关于人类活动的。

第一是“事情”,映射符号是 g9-3-0,“事后,事理,事例,事前,事先”等词汇的“事”是这个义项。

第二个义项是 g10-0 或 g10-0-0,如“事机,事权,事务,事项,事业,事由”等词汇里的“事”。这里最常用的是“事务,事业”两词,“事务”大体相应于 g10-0-0,词典里的两项解释可并为这一个义项;“事业”的映射符号是 gr10-0,但“事业”还有另一义项 xp12-10-0-0,它与“企业”p12-10-2 相对应。(注:这里的 p12 是一个特殊约定的类别符号,代表各种社会组织。在形式上它与扩展基元概念表示人类社会活动的本体层符号 12 相混淆,但实际上不会发生,这里的技术细节就不来说明了。但终究造成程序多了一级判断,对人的直觉也不方便。所以,后来改用小写字母 e 代替这里的 12,用 pe 代表社会组织。)

以前曾将“事”映射为 g9-0-0,显然不如映射为 g9-3-0 妥贴。

1994.10.13 30

清理常用词。

清理中发现,已填词义库相当混乱,殊为可虑。最大的混乱是对 0 和 8 的滥用。自明确对偶、对比、包含三类概念的特殊表示方式,从而明确高中低三层次的自然表示方式以后,8 用于表示一般,相当于原节点的 0,而 0 表示对偶性概念 1 与 2 之“对立统一”,或包含性概念之扩展符号。主体基元概念的第三层都存在 8。但新节点只标出不是用“一般”命名的 8 号节点,凡用“一般”命名者则一律省去。非“一般”8 号节点的设置,当然反映了我个人的一些观点,未必恰当。另外,在“一般”中,一是一般性的“一般”,二是带有基本性的“一般”,这类概念仅出现在主体基元概念的 0 分行。以前对基本一般性运用过多,清理以后仅对 5-0-0-保留了“基本一般性”的含义。

8,0 的滥用错误还未清除彻底,只能求助于未来的查库程序,作第二次清理。

1994.10.28

整理 jx1。

在自然物中,我们定义了一组基本物 jw。每一组基本物有相应的物性,用符号 jx 表示,这样的符号设计显然有利于联想和理解。在这些物性中,最重要的是温度 jx0 和色彩 jx1。冷热是温度的值,符号分别为 jxz0-0-12-2-1,jxz0-0-12-2-2;红橙黄绿青蓝紫是色彩的值,符号为 jxz1-0-12-i-k,明暗是光的值,符号分别为 jwz1-0-1,jwz1-0-2,这里的第二个 0 实际上是冗余数,其作用仅在于引出第三层来,因为约定第三层才进入中层表示。

汉语对色彩的表达非常丰富,例如“红”又分潮红,赤红,绯红,粉红,火红,金红,品红,肉红,深红,水红,桃红,剔红,通红,鲜红,杏红,血红,殷红,嫣红,银红,枣红,朱红,紫红等等。这些“红”又是“红”的值,在简化表示里这些细节是无法表达的。第一期词义库中不但不区分这些红,甚至不区分红橙黄绿,仅记为 jxz1-0。不过,以前填过的词没有掌握好这个简化标准,相当混乱,今天整理了一下,但还不彻底。

1994.10.31

译“收”。

“shou”的阴平只有一个字“收”,其本义是 v2-1,其扩展义有 v5-2-1-1-2,如“收兵,收工,收盘,收摊,收场”里的“收”。“收摊,收场”也有 v1-1-2 的含义。

这里值得一记的是“收复,收回”。这两个词的主角是“复和回”,其映射符号分别是:vg12-2-2-0 和 vg9-2-0-0。以前将“回”定义为 2-0-3,简直是鬼迷心窍。0 是 1,2 之“合”,4 是 5,6 之“合”;“回”是来去之“合”,非常贴切。这里有一个来去的次序问题,“回”不管这个次序,唐诗“少小离家老大回”的回,是先去后来,但“回个话,回一封信”里的“回”,则是先来后去;“回”恰好是 2-0-0。汉语有“回来,回去”的说法,其中的来去并不是冗余成分。

1994.11.1

译“手守首”。

“shou”的上声只有这三个字。三者的独立性都很强,而且不相上下,这在汉语的 1200 多个音节中是独一无二的现象。

汉语“手”的概念非常丰富,词典里给出了六个义项,实际不止。手是与人类同步进化的产物,因此,赋予手、人、人类活动以同一性意义是最自然不过的联想。词典只给出第一项同一性,未给出第二项同一性。手作为人类活动的语言符号,可以相当精确地映射为 gv9-0-0,由此义构造出来的词汇在“手”的各义项中仅次于“擅长某种技能的人或做某种事的人”。现将与人类活动有关的词列举如下:碍手,罢手,插手,缠手,扯手,得手,动手,棘手,接手,辣手,入手,撒手,甩手,洗手,下手,歇手,沾手,住手,着手。

1994.11.2

译“受”;“书叔输抒”。

“shu”的阴平常用字较多,但无突出者,频度知识的信息量很少。有趣的是,其语义信息却比较集中,上列四字的“书输抒”都与信息的转移 2-3 有关,而且“书抒”仅与 2-3-8 有关。这里再说一下基元概念第三层数字 8 的意义。8 相当于一般,相当于未引入对偶性和对比性特殊表示以前的 0。现在第三层的 0 与 4 一样,具有特殊含义:0 表示 1,2 之并;4 表示 5,6 之并。任何概念只要有 1,2,就必有 0;有 5,6,就必有 4,这是概念的基本规律。但自然语言不一定有相应的词汇符号,或是这种语言有,而另一种语言没有。如果自然语言没有,往

往是表示该项并的概念不常用,不必引入专用的词汇。包含性概念也用 0 表示扩展(被包含),但它后面跟标志符‘ - ”,不会引起混淆。

这里的‘书抒’的信息转移意义都属于 v2-3-8-2。“输”的本义属于 v2-0-10,但‘输入,输出’分别属于 v2-0-1,v2-0-2;“输血,输液”两词属 7 号结构方程,但类别符号为 vc,独立性为 1,表明该词需要补充对象。

“叔”的映射符号是 p4-0-9一 * 1-1,亲属为 p4-0-9,后面的 1-1 是本体层,相应的意义是:辈分-性别。辈分的表示方式是 :0,同辈 ;1,上辈 ;2,下辈。性别的表示方式是 :1,男性 ;2,女性。这里略去了两项信息,一是辈分的值,二是亲疏的值。辈分信息也采用了另一种含值的表示方式 :0 3 为同辈,1 为兄,2 为弟 ;4 7 为上辈,4 为一般上辈,5 为上一辈,6 为上两辈,7 为上三辈 ;8 11 为下辈,8 为一般下辈,9 为下一辈,10 为下两辈,11 为下三辈。12,13 暂缺。现在的词义库里两种表示方式并存,非常混乱。查库时必须以 p4-0-9 为标题查阅一次。

1994.11.4

译‘刷 :耍”;“衰 ;甩 ;率,帅”;“双 ;爽”。

“双方”是 g4-0-0-4 的精确表达,双方是此方 g4-0-0-5 和彼方 g4-0-0-6 之并。从语义结构来说,双方属 2 号方程的偏正结构,但在层次网络符号里,同行偏正的双音词采用直接表示更为简明,如这里的‘双方,此方,彼方”。“双”字的第一义项是 g4-0-0-4。“双边”一词可直接采用这一义项,但不必采用 0 号方程,用直接表示即可。采用方程表示有三个目的:一是语义知识库的规范化,二是信息压缩,三是为了便于提供更多的信息。这是从总体来说的,从个别词汇来说,是否采用方程表示,主要看第三点。像‘双方’等这样的词汇,直接表示的信息已非常充分,又不存在表示空间的限制,当然就不必采用方程表示了。

1994.11.6

译‘水 ;税,睡”。

“水气土”是我们定义的三大状态物之一,映射符号是 jw5-i-8,“水气土”分别相应于 i = 2,1,3,这里的 8 原定义为 0。在汉语里,以水气土为词根,构成了很多双音词。这些词的语义绝大多数适合于结构方程表达,但考虑到语义库投入使用过程的实际情况,大部分采用近似直接表示,仅少数采用方程表示。在‘水’中,“水电,水患,水碱,水利,水力,水压,水灾”诸词满足方程 3-6,但实际上采用方程 5-1,这是一个特殊约定的效应方程,关于它的含义,不能不回顾一下〈问答 34〉里的一段话:“4 号方程的 2 级表示无 2、3,而 5 号方程的 2 级表示无 1,否则就是错误的数据,或者,将来利用这一点作其他表示”。现在,就来利用这一点了。方程 3-6,3-7 表示因果律。因果是作用效应的表现之一,所以,结构方程 3-6,3-7 可用 4-2 或 5-1 替代,这是不言而喻的。作用、效应、对象、内容四个结构方程的信息最为丰富,满足这些方程的双音词应尽量采用方程方式。其他则多用近似直接表示。这是第一期词义库的建

库原则之一。

1994.11.7.星期一

译‘顺,瞬’;‘说 :颁’;‘挖 :瓦’。

“顺利,顺当,顺溜”等词都采用直接表示,映射符号为 u3-0-10-1,其对偶性概念‘困难’则映射为 u3-0-10-2。也许在‘顺利’等词前面加类别符号 r 更好一些,可与‘容易’有所区别,但这是枝节性的区别。根本区别在于,‘容易’只是效应概念 3-0-10 的属性,而‘顺利’同时还是过程概念 1-0-0-9 的属性,这一区别直接表示就无能为力了。

“瞬时”是一个典型的对比性 j1-2 概念。“瞬时,短时间,长时间”可分别映射为 :j1-2-12-3-k,k = 1,2,3。j1-2 按定义是不考虑起点或终点的,但有些 j1-2 概念需要考虑,例如‘悠久’是指从过去到现在的时间间隔,而且有很久很久的意思,其映射符号应为 :(j1-2-12-4-4,j1-1-1),直接表示就简化为 j1-2-12-4-4。

“说”字的词汇多数用方程表示,例如,‘说服’用方程 4-5-0-5,‘说穿,说破’用方程 5-6-0-5。4-5 表明,‘说服’是作用效应型概念,必须跟有作用对象和内容,句子才算完整。5-6 表明,‘说穿,说破’是复合效应型概念,跟随的语义块必须包含对象和内容。最后的 5 表明,它涉及的对象和内容优先于人和事(参看〈问答 34〉)。4-5 的对象和内容必须是两个独立的语义块,而 5-6 的对象和内容则可并为一个语义块,这是两者的根本区别,显然,这个信息对于作用—效应句的句类分析,具有头等重要性。

1994.11.8.星期二

译‘外’;‘弯 :完’。

“外”是基本概念 j4-2-2,与‘内’j4-2-1 构成对偶性概念。汉语充分利用了内外这一对概念的基本性,按照它们的联想脉络派生出众多的词汇。从基元概念来说,关系 4-0 有内外之分 ;结构 5-4 及结构体 w5-4 有内外之分 ;转移的输入 2-0-1 是从外到内 ;转移的输出 2-0-2 是从内到外。从基本概念来说,空间 j2 有内外之分,进而扩展到地域和地区的内外之分 ;质与类 j5-0 有内外之分 ;内容 j5-0-1 与形式 j5-0-8 是与内外相对应的。以上所列,是概念联想脉络的主干之一。这样的联想主脉络大约一共有十几条,这里给两个示例:

(1)10-i,12-,p12-10-i ,pj2 ,p10-i, j7-,12-9,12-10

这个清单可命名为人类专业活动清单,所谓新闻,主要涉及这个清单。从人类的万千活动中分出一项专业性活动,谈不上什么学问,但对于理解处理十分重要。我倾向于把它列为第一号联想脉络。

这一联想主脉络的特点是 :二级联想脉络的界限最为分明,这就是 i = 1 8 的各项专业活动。10-i,p12-10-i,p10-i 的搭配优先性几乎是绝对的,这项知识对于解模糊及纠错处理极为宝贵。〈问答 1〉里的句例‘刘嘉玲向上海中级人民法院起诉汕头雅丽斯实业公司’,如果以汉语拼音输入,模糊度很大,但只要运用这一联想脉络,问题都可迎刃而解。

表达人类专业知识以作用效应句为主,这很自然,因为专业活动的目的必然是为了取得某种效应。

词库里现在大约装进了 200 多位人物的名字,包括美国人评选的"影响历史进程"的 100 位历史名人。这些人名的绝大多数都可以挂靠 10-i,只有两类人例外,一是探险家,二是宗教活动家,分别挂靠 9-7-4-4 和 12-8-2。顺便一说,在 100 位人物中遴选了三位探险家,哥伦布,皮萨罗,科尔特斯,后两位拉丁美洲的殖民者似乎不够进入"100"之列,现未装入词库,但加了一位中国人徐霞客。

(2)13-,7-1 ,7-2,9-0-2,9-3-3,j8-,12-10

这个清单可命名为人类行为清单。这里应对"行为"和"活动"两词作一点说明,两者都是指人类的智能型行动,但我希望从两个根本不同的角度加以区分,这就是功利和道德。"活动"联系于功利,而"行为"联系于道德。本清单以 13- 为首,有突出行为表述的含意。现代西方哲学有一种观点认为,哲学只剩下一个问题需要继续探索,这就是道德,我比较同意这个看法。从语言的表达和理解来说,"行为"有它的特殊性,语种的个性表现也最为鲜明。就汉语来说,有许多"行为"概念是西语里不存在的,例如"恕,中庸,孝"这些儒家的基本概念,西语就没有。当然,这些概念在其发源地已接近消亡,这一民族文化现象让后人去评说。从现代汉语来说,设置 13 行确实已无必要,但为了过去和将来,这一行是不可缺少的。

1994.11.9.星期三

译"完;万"。

"结束"和"完成"是两个密切相关的概念,从过程的角度看是"结束"1-1-2,从效应的角度看是"完成"3-0-10。"完"两义兼有,词典里一共给出了 5 个义项,这两个义项是其中之二,另外 3 个义项的编号是:1.完整;2.消耗尽;5.交纳(赋税)。从独立性来说,义项 1 最弱,根本不能独立使用。"完整"是状态的基本属性,联系到基本概念量与范围 j4-0,较为精确的表达是:u5-0-0-4-0-1。"完美,完善"也是状态的基本属性,它们不仅联系到量与范围,还联系到质 j5-1,其精确表示是:u5-0-0-4-0-1 + j5-1-1。词典仅提到"完整",显然是不全面的。这类的语义结构本质上是 1 号方程,但用 0 号方程表示更简洁。目前词义库中采用了近似直接表示 u5-0-0-4-0 和 u5-0-0-5-1,未予改动。

在口语中"完"字的当然第一义项是 v1-1-2,第二义项是(v1-1-2,v3-0-10),第三义项是 hv|(v1-1-2;zu3-11-2-13-0-1),第四义项是(v1-4-2,v7-7-1)—完蛋之意。但在书面语中,第三义项用得更多。这四个义项皆可独立使用。至于上述"完整,完美,完善"等义项以及交纳赋税的义项都是不独立的。

1994.11.14

译"晚;万"。

"晚"的近似本义是夜,夜的精确含义是 wj1-0—0-0-2。晚是指夜的前一半,其精确表示

应为 wj1-0—0-0-2-1,词汇“晚上”大体上是这个意思。构词词典的释义一般与《现代汉语词典》相同,但对“晚”字加一条“指日落的时候”,是对的。

1994.11.21

续译“晚”。

“晚”的第一义项已如上述,其另一义项,“迟”的意思,可独立使用,如:为时已晚,太晚了。这个意思的映射表示用 vu1-0-0-9-2,相应的对偶概念 vu1-0-0-9-1 是“早,提前”。“早,晚”作为动词是不及物的,而“提前,推迟”是及物动词。及物性的表示对“词”和“字”采用了不同的方式。词主要用独立性的 0,1 来区别:0 级不及物,1 级及物。字则用 gv 和 v 来表示:gv 表示不及物,v 表示及物。采用不同的表示方式当然有技术方面的原因,同时也有含义上的区别。词汇级的及物性是传统定义,但字级意义的不及物则有“符合”和“违反”传统的双重性。不及物的“字”可按及物方式构词,如“晚婚,晚点”就是例证。汉语的这种辩证表现不胜枚举,黑格尔曾对德语词汇的辩证表现大表惊叹,但德语与汉语相比,不过是小巫见大巫而已。

1994.11.22.星期二

译“王往”。

“王”是一个复合概念,但它有两重意义。一是封建国家的最高统治者,二是最高的爵位。这里,“封建,国家,统治者,最高”都有节点表示:封建 gu10-1-0-2,国家 pj2,统治者 p12-4-4-1,最高 gu5-6-13-0-1。爵位本身就是复合概念,映射符号是 g12-5-6/g7-3-5。

“往”有两个独立义项。一是 v2-2-11-2,这个义项词典分为两项解释:去;向(某处)去。在第一义项中,举的例子是“来往,往来”。第二义项词典释义为“过去的”,只看作形容词,不管它的动词属性,这有点过分与古汉语“划清界限”了。我们将这一义项映射为 jvu1-1-1,并记独立性为 0,否则,毛泽东的著名诗句“俱往矣,数风流人物,还看今朝”,你就无法理解了。由此义构成的词汇“往常,往年,往日,往昔”一律用直接表示 wj1-1-1,以求简明,仅“往事”用 2 号方程表示。

“网罗”一词用 0 号方程 0-0-0-1,与类别符号 v、独立性 1 配合可知,它是及物动词,要求以人为对象。实际上上述全部信息已包含在 0-0-k-1 中,其他是冗余成分。

“往”的去声意义同介词的“向”,相当于英语的“toward to”,但介词“向”的映射符号值得作详细说明。“向”作为介词是一个高层 l0-k 概念,包括是对广义“方向”的指示。对这一点,以前认识不足。我在〈问答 14〉中将“向”的逻辑意义写成:

(l0-2-2-0;l0-5-2-0;l0-4-1-3;(l0,j2-1))

这个表达不仅繁琐,而且有严重的漏报错误,不如写成(l0-2;l0-3),以避免漏报的错误。当然,它又隐含了虚警的错误。通过繁琐的表达可以消除这个错误,但这个简洁表示必须采用(作为字义库的一级表示),因为它抓住了“向”是 B 语义块或 C 语义块切分标志这一最本

质、最重要的意义。这里举一些句例加深读者对以上说明的理解。

……向敌人发起进攻

……向 × ×(学习,致敬,运送,传递,说明,宣布,索取,提供,挑战)

……向△△(转化,演变,发展)

向左转,向前看,向东去

第一个例子表明原表示的漏报错误。第二、第三组例子表明精确化表示的优点,其中,× ×表示对象,△△表示内容。最后一个例子表明现高层表示的不足。这些例子也对句类标准格式的自然顺序 E + B + C 提供了证据。当一个句子打破这个格式时,就需要引入逻辑符号 l0-k 对顺序变动后的语义块进行标记。

上面讨论了"向"的逻辑意义,显示了它的复杂性,而"向"的基本和基元意义也同样复杂,值得深思。汉语用一个"向"字表达那么多的概念,不是偶然的。空间的方向 j2-1-8,过程的趋向 1-3-9,转移的定向和途径(称之为"途向"亦无不可)2-0-k(k 分别取值 9,10),人的志向 11-0-0 都显含或隐含一个"向"字,是汉语"字义基元化"的生动体现,表现了这些概念之间存在着很强的交式关联性。我在〈概述〉中讨论过程与转移的区别时曾说过一句话:"过程是状态序列的时间表现,转移是状态序列的空间表现"。而所谓追求,就包含人类活动的序列,没有序列,就谈不上追求。由此可见,"序"这个概念是所有上述概念产生交式关联的纽带。我们将"序"列为基本概念之首,并安排在 0 分行就是基于这些概念深层关联性的考虑。这里顺便谈一下酝酿甚久但未敢率定的 j0 二级节点设计问题。

上面谈到了"向"有具体空间之"向",有过程、转移、人类活动等抽象之"向"。物理和数学的向量,联系于空间之"向"。人们说时间不能倒流,意味着时间的单向性,然而有"向"。可见"向"普遍存在,为什么?因为向是序的基本属性,有序必有向。序还有什么基本属性呢?距离。"距离"这个概念同"向"一样,来于空间,但超出了空间,具有"基元之基元,基本之基本"的特征。值 z 的差就是距离。当然 jz0-0 之差也是距离,概念 j1z 0-0-0,即相似性的值,也是距离,不过习惯上叫做差距。概念 4-0-9 叫做关系的紧密性,实际上也是距离。所以,在 j0 中引入距离的概念是符合语言需要的。这些,就是我"酝酿甚久,未敢率定"的想法。近来在填写字义库和词义库的过程中,于惊叹汉语的概括性及辩证性之余,终于获得了豁然开朗之感。于是补充定义:

j0-1　　向

j0-2　　距

1994.11.23

译"忘,望,妄,旺"。

"忘却"与"记住"不是对偶性概念,而是一对相互否定性概念。若"记"映射为 6-8-0-1,"忘"只是它的否定,用　v6-8-0-1 表示,符号"　"表示否定。(注:后来改为"!"。)

译“威”为,违”。

“威”的词典释义是:“压服的力量或使人敬畏的态度”,其映射符号的高层近似分别是:r0-0-0-9,r7-1-1。但“威望,威信”也可近似映射为 r13-0。“名誉,信誉”等是典型的 r13-0。13 行的三级节点尚未设计,在二级近似下,“威望,威信”与“名誉,信誉”是同义的。

“wei”这个音节,有三个常用 13 行概念,除“威”之外,还有阳平的“为”和“违”。对“违”字,先赋予三级字义 v13-0-0-2。“遵守”和“违反”是人类行为最基本的概念对,这比较自然,如同人类理智反应的基本概念对是“同意”与“反对”(vg9-0-2-k,k = 1,2)一样。在第一期词义库中,“违背,违反,违犯,违抗,违拗”都记为 v13-0-0-2,有些就是近似。“违章,违法,违纪,违禁,违例,违心,违宪,违约”都采用方程表示 7-5-0-5,类别表示一律为 vc。显然,信息的表达不够充分,但在 2,4 级表示里还有潜力可挖。所以,结构方程的 2-3 规则和 4-1 规则有待补充。

“为”是一个很特别的字,特别是阳平的“为”,其独立义项值得详说。

第一,阳平“为”字使用频度最高的意义是充当程度副词的后缀,其语法功能相当于英语的“ly”,映射符号是:(h↓uu)|juu6-0。词典里举了很好的例子,转录如下:大为高兴,广为传播,深为感动,极为重要,甚为便利,颇为可观。词典里还具体指出:“1.附于某些单音形容词之后;2.附于某些表示程度的单音词后”。这两点具体特性,上面的层次网络符号大体上有所体现了。符号“h↓(五元组符号)”可视为词性变换符号,如下面的“化然性”。“h↓”后跟其他类别符号如 p,w 的意义,对照相应汉字自明。

化	h↓vg
然	h↓gu
性	h↓g
员,手	h↓p
家	h↓p10-0-0
子	(h↓w;h↓p5-0-10)

我举这些例子希望表明一个事实,就是词性后缀的说法用五元组符号的组合表示更为确切,例如“化”和“然”,称之为 vg 型和 gu 型概念的后缀,比用传统术语,如动词后缀、形容词后缀来表达,显然要好一些。

第二,相当于英语介词“by”的意义。但映射符号是 1q2-1,而不是 l0-1,因为它要与“所”搭配使用。

第三,相当于英语介词“into”的意义。映射符号是 l0-3-3-0;l0-2-0-0。此义词典未给,而是包含在“变成”的义项里。“为”有“变成”的意思,但独立性为 3 级,仅用于某些成语里,如“洋为中用”里的“为”。至于词典里在此条义项下给出的例子,“一分为二,化为乌有,变沙漠为良田”等例句里的“为”并没有变成的意思,仅是一个逻辑符号,词典编者显然是把“分,化,变”的意思转嫁给“为”了。

第四,以阳平“为”字构成的下列词汇:“为重,为主,为难,为期,为人,为生,为首,为数,为限”,体现了汉语表达的特殊简化风格。相应的意义表达在英语里要借助于一个短语,因为英语没有对应的词。在这些词汇里,“为”字实际上只有语法功能,其作用是使后面的字变为动词。按层次网络符号的约定,可记为 q↓gv,并按 1 号结构方程构词。虽然这种方式可以准确地表达“为”的这一义项,但有些不如采用直接表示更为简明,如:为重 jv7-2-5;为主 jv7-2-1;为期 jv4-2-4|j1-2;为数 jv3-0;为限 jv4-2-4;为首 jv12-4-4-1。不及物性通过 0 级独立性来表示,但前搭配信息“以”就不能表达了。

1994.11.26.星期六

译“未,为”。

汉语里有五对逻辑性概念同音同调,它们是:“及,即;未,为;以,已;由,犹;在,再”。四年前当我第一次注意到这一现象时,颇感困惑,甚至产生过这些字的古音或许有所不同的想法。说明当时还未彻底摆脱两点“迷信”,一是“一词一本义”的迷信,二是“仓颉造字”的迷信。这两点迷信,来源是一个,将词主要理解为命名,就像每个人有一个名字一样。实际上,词的作用不仅是命名,更重要的是充当概念联想的指示符号。语言的根本使命是概念关联集合的整体表达,而不是孤立的概念的罗列。语言不要求词的单义性。词的具体意义依赖于上下文,是整体表达即语言深层结构的天然特征。因此,词的多义应视为正常现象,而单义才是特殊情况。消除词的多义模糊,是语言理解的基本操作。概念层次网络符号的设计正是以这一认识为出发点,产生如下的基本符号优化准则:为消除词的多义模糊提供最充分的信息。

词性的概念立足于对词作狭义的命名理解。在明确了五元组和概念的类别性、层次性和网络性以后,回过头来看词性的概念,有点类似于从现代科学看“金木水火土”的学说。这个学说过于简化,但不能说它毫无道理,这类唯象理论有时也有其简明的优点并蕴含深刻的哲理。不过,词性概念的这一优点被人们过于夸大了。从消除词义模糊来说,词性的作用甚微。从口语的听觉感知来说,消除上列五对同音同调逻辑概念的模糊,词性知识是远远不够的。从书面语的视觉感知来说,也是如此。这五对逻辑字的错别字,你很容易辨认,但决不是仅依靠词性知识。

当然,语言对词义模糊的容纳是有限度的,汉语拼音化之不可行就因为它大大超出了这一限度。但你不必对当年“汉语拼音化势在必行”的结论感到惊讶,因为它反映了 20 世纪中国思想文化界的某种状况。这一状况的后遗症远未消除,本课题当前的困境多少与此有关,故情不自禁写下了上面的话。

“未”与“为”这一对同音同调字,都是高层逻辑概念,但义距甚大,分属于 j11 和 l1。j11-k 和 l1-k 的设计都非常复杂,是所有二级节点中最难把握的一对,反复多次,迄今仍有疑问。我觉得非常需要了解更多语种对 j11 和 l1 的表达,但这已是心有余而力不足了。

对 j11,容易想到的是“是否有无”,简单的“是”字句和“有”字句大约是频度最高的口语

句型,也是幼儿学习句子的起步句型之一。“是否有无”句看来简单,但“是否有无”这四个概念却很不简单。首先,它们是属于同一层次的概念么?这个问题就很不好回答,因为,它们的层次究竟谁高谁低就很难判断。一方面似乎可以说,“是否”的层次高于“有无”,因为,“有无”是关于存在性的“是否”判断。但另一方面,“存在”又是更基本的概念,古希腊哲学家巴门尼德对“存在”的概念曾进行过深入的探讨,我们同时代的祖先在这一点上应该是自愧不如的。如果不“存在”,“是否”从何而来?

其次,它们属于对偶性概念么?这又是一个难题。似乎“是否”是对偶性的,而“有无”不是,因为,“肯定”的非不等于“否定”,而“存在”的非,就是“不存在”,也就是“无”。但是,逻辑上有“否定之否定”的重要概念,却没有“肯定之肯定”的概念,说明“肯定”与“否定”又有非对偶性的一面。这里我们引入了“肯定,否定,非,不”这四个汉语概念,“是”与“肯定”,“否”与“否定”,“非”,“不”(此外,还有正在讨论的“未”和口语常用的“没”)又是什么关系?

第三,从句类分析来看,如果对“是否有无”的有关概念的符号映射处理不当,将会带来混乱。因为,我们定义以“是否有无”等基本逻辑判断概念为特征要素的句子为基本判断句,可是,大部分句子都可以转换为否定的形式,那么,句子的否定形式与否定型基本判断句如何区分?

第四,情态动词与基本逻辑判断“是否有无”有密切的联系,这些情态动词的符号映射如何处理?

jl1-k 的设计,主要涉及上述四方面的问题。我的体会是:这些问题的处理难点不在于问题本身,而在于把侧重点放在哪里。开始的时候我并不体会这一点,因为我认为把侧重点放在有利于理解的基本操作(即消除多义模糊处理)是理所当然的。但年轻的同志们不断提醒我,理论上的理解基本操作与程序的理解基本操作有所不同。虽然对这一点我依然坚持自己的基本观念,即层次网络符号计算机绝对不难把握,但感到词义的表达确有浅层及深层两种形式的必要。人的理解过程也是如此,并不是每个概念都需要到深层去追根究底。第一期词义库的建库过程促进了浅层表达方式的发展。下面就对 jl1-k 有关概念的符号映射问题作具体的说明,但在叙述的次序上正好与上列的问题次序反过来。

第一,情态动词,如英语的“may,can,will,need,should,must,ought to, have to”,汉字“可,能,应,需,必”以及由它们组成的词汇“可能,能够,必能,应该,必须,不得,不得不”等都与势态有关。(情态的意思就是势态,将 modal 译成情态的始译者,我猜想他是过于重视了 should,have to 里的情理成分,因而创造了情态这个词。)对情态类概念,以前我是通过 jl1 与 5-3 以 7 号结构方程组合的方式来表达的,意思是关于势态的基本判断。现在看来,这种方式适合于深层表达,浅层则可直接纳入 jl1-k 的设计。因此,jl1-k 的最新设计是:

jl1-1-1	是,肯定
jl1-1-2	否,否定,不,非也
jl1-1-5	存在,有
jl1-1-6	无

jl1-2	情态动词“可,能”
jl1-2-12-3-1	可能
jl1-2-12-3-2	能够
jl1-2-12-3-3	必能
jl1-3	情态动词“应,必”
jl1-3-12-2-1	应该,应
jl1-3-12-2-2	必须,必

这里,将情态动词分为两组:jl1-2 表示纯粹的客观势态;jl1-3 则同时含有主观成分,或者说情理部分。

第二,在这个表达方式里,未突出否定的概念,这一点,不及原设计。与此相联系的另一缺陷是将“是否,有无”都按对偶性概念处理。如上所述,“肯定”与“否定”是一对特殊的对偶性概念,而“有无”只是对称性或镜象性概念。对偶与对称的区别在于:第一,对偶存在中间状态,对称不存在;第二,对偶一方之非不等于另一方,而对称一方之非则等于另一方。“有无”是典型的对称性概念。对偶与对称的上述差异,属于语义深层的隐知识,人类的理解过程很少涉及这一隐知识。在那个只强调“斗争”是纲的特殊时期,甚至只强调对称性而完全漠视对偶性。所以,目前不区分对偶和对称的做法也许更符合自然语言的习惯,而原来的精确表示反而显得过于书生气了。

至于句子否定型的表示,仍采用原定方案,即放在节点 7-4-1-k 里,具体的 k 值未定,尽管现在已约定 7 行的高层层次数,但这个 k 值仍缓定为妥。这不影响语义库的建设,但另一方面,“非”的观念在语义表达中经常使用,因此,特地为它引入了一个结构符号 以简化表示。

第三,自然语言对否定的表达有多种样式,这到底是自然语言的冗余性表现,还是语言表达的实际需要?两种因素都有,至于是哪种因素为主,倒不必深究。问题是自然语言的各种否定性表达如何在映射符号上加以体现。基本的问题是动态和静态的区分,西语对这一区分比较严格,汉语不那么严格。回到引出这一大段议论的“未”字,它基本上是用于动态否定的表达,因此,可记为 jlqvu1-1-2。

第四,“是否”与“有无”的层次高低问题,新设计采取回避策略,这比原来规定“是否”的层次高于“有无”要高明一些。

最后,简单说一下去声“为”的符号映射。这是一个典型的高层 l1 概念,就是说,只需要本体层,不需要挂靠层,共有三个 0 级独立义项:l1-0,l1-6,l1-7,分别表示广义逻辑对象、动因和目的。

1994.11.28.星期一

译“温文稳问”。

“wen”的四声各有一个常用字,是否还有其他的音有这个现象?不妨留意一下。对汉字

的表述,似乎应该引入“同音”和“同调”的说法,同音字仅表示拼音相同,同调字音调皆同。

“温度”的映射符号是 jx0,把它作为基本物性之首,大概是无可争议的。汉语从“温”联想到性格和态度,很有趣,是否多数语种也如此?

“文”是 10-3 的高层概念,汉语由它派生出一系列的低层 10-3 概念,是汉语字义基元化的又一生动体现。10-3 的三级节点尚未设计,所以,这些低层概念都以直接方式放在 10-3,仅表现了它们的类别性差异。汉语由 10-3 的“文”联想到人类行为 13-0 的基本特征,这就是“文武”之文,并进而联想到手段 12-10 的基本方式。后一联想十分自然,因为 13-0 与 12-10强关联。词典对“文”字的义项列举了十余项,却没有这两个一级义项。现代的“和平”概念只是“文”的一部分内涵,儒家的文武学说是治疗现代社会许多病态的良方。近代中国的哲学大师之一熊十力先生在建国之初曾向毛泽东主席进万言书,阐述这一论点,熊先生因此而被贬入冷宫。有感于此,写了上面的话,并在词库中加了“文武”一词,聊表对这位乡贤的景慕之意。

“稳”是状态的基本属性之一。状态的基本属性很多,但主要放在 5-0-0,然后挂靠基本概念,这是层次网络符号设计的基本约定之一。“稳定性”挂靠时间,其映射符号为:gu5-0-0-1-0。为什么作用效应链的六个基元概念只有状态的 0-0 分行挂靠基本概念?简单地说,一是因为 5-0,5-1,5-4,5-5,5-6 都着重于静态表达,把它们的共性集中到 5-0-0 分行,符合层次网络符号设计的基本原则。二是所谓基本属性,不外乎属性的“序时空量质度”表现,这可视为基本属性的定义。那么,作用效应链的其他基元为什么不作类似的挂靠?这是因为我们对 0 分行赋予了普适性,用状态挂靠,最有代表性,其他就不必了。我在〈概述〉中说过,抽象概念的“你中有我,我中有你,需要通过 0 分行的普适性予以体现”。

当然,将状态向基本概念挂靠,只是一级近似。例如,要对“稳定”这个概念作二级近似描述,就需要加上映射符号 vu3-0-9(不变化之意)。但这个一级近似对于理解的基本操作已十分有效。例如,“平稳过程”和“平稳过渡”这样的组合概念就符合广义同行优先准则。

“问答”是信息转移 2-3 的对偶性三级节点。可以设想,“同意”与“反对”,“问”与“答”是人类智能进化的起点。我将这两对对偶性概念分别列为 9-0-2-k,9-2-3-9-k (k = 1,2),与这一设想有关。对“问”与“答”,还应另有义项 6-2-2-3-9-k,因为,动物甚至植物的信息转移,也必然有“问”与“答”。不过,其基本含义是 vg2-3-9-1,目前,在词义库中就以直接方式登记了这一义项,当然标明它是多义词。

疑问句是语法学的基本句型之一。提问的语法手段,不同语种各有千秋。这些语法知识对句子分析和理解都极为重要。层次网络符号设计将语法知识统一安排在 7-3 分行之后,提问的语法知识在 7-4-2。但目前 7-4-2-k 的设计有重大失误,就是为照顾汉语的特点加了一个量的询问,而没有给原因的询问安排一个节点,应该作如下调整:将 7-4-2-4 定义为原因的询问,而将数和量的询问并为 7-4-2-3。

将数与量分为两大类基本概念,乃基于多方面的考虑,对此,我作过多次说明。这里想补充一点,就是在基本概念的排序上将数放在最后更合理一些,在叙述上,也可免除一些不

便。例如,上面说到状态的挂靠,就是跳过数。这是马后炮,遗憾而已。

1994.11.29.星期二

译‘窝 我 握,卧’。

‘wo1’(以后用拼音后的数字表示调号)的大部分词汇用直接表示,如‘涡旋,涡流’映射为 rw1-0-9,意思是运动的效应物。对‘涡旋’来说,这个近似性很差,没有表达它在形态 5-1 方面的意义。‘窝囊’直接映射为 vu7-2-1,意思是‘无能’。

‘我’是 p4-0-0-1—0 的精确映射,但以‘我’为前搭配构成的词,是 p4-0-0-1(我方)的意思。

‘卧’,以及相应的‘躺,站,坐,骑,挺,昂’等字,都是形态的动态表达。形态表达的静态与动态之分,不是类别符号 v(或动词)和非 v 所能完全刻画的。其实,这一特征是任何基元概念的共性,不过,形态更加突出罢了。动态节点 5-2 的引入就是为了弥补类别符号表述能力的不足。引入 5-2 的必要性,谈过多次,以后碰到好的字例,当作具体的详细说明。由于形态的表达更需要区分静态与动态,最初曾考虑将 5-1 的下一级分为静态与动态两个节点,当然后来放弃了这个想法。‘挺,昂’等是典型的动态形态,映射符号是 v5-2-5-1。

1994.11.30

译‘wu1’。

‘(早,午,晚)餐’三词词库都未收,原因是旧版《现代汉语词典》未收这三个词(新版不知),但收了‘(早,午,晚)饭’。词典在字义解释中用到的双字词有的不再纳入词条,有的又纳入,纳入与否编者似未制定统一或明确的标准。在建库时要求同时搜集字义解释中的词,但这一要求远未达到。词库之不全,这是最大的原因。

‘wu1’无高频词,但有‘诬,污,乌,屋’等中频字。中频字的义项大部分是复合概念,而高频字的多数独立义项是简单的基元或基本概念,这当然是预料中的事,否则,层次网络符号的设计就存在根本缺陷了。

‘诬’的词典释义是:捏造事实冤枉人。‘捏造事实’的映射符号是 j8-1-2/vg9-2-3-9,‘冤枉人’的映射符号是 v3-2-2→p,但这两组符号的组合则可以采用三种不同的组合结构。一是作用效应结构,‘捏造事实’为作用,‘冤枉人’为效应。二是 l1-1 结构,前者为方式。三是 l1-7 结构,后者为目的。在字义库中选用第一种方式,因为‘诬’字的核心含义是捏造事实。以‘诬’字为前搭配构成的词,多数采用 0 号结构方程,0-3-0-1 或 0-0-0-1,类别序号都是 vg,独立性都是 1,但‘诬告’采用 4 号结构方程,因为它的效应是法律行动。

‘污’的义项暂定为 gu5-0-8-2,其对偶字‘净’就是 gu5-0-8-1,相应的‘污垢’是 rw5-0-8-2,‘污染’一词则采用 4 号结构方程 4-1-0-2 表示,这里的作用是 rw5-0-8-2,效应是 v3-3-2(它将是‘染’字的一个 3 级独立性义项)。

对‘乌拉,呜呼’的处理,采取最简化的方式,仅映射为 g7-4-3-。这种方式也用于‘污点’

一词,仅映射为 r13-0-,凡后继层次尚未设计的节点都采取这种方式。

1994.12.1

译"无;五,午,武,舞,侮"。

"无"是 jlvg1-1-6 的精确映射,其对称性概念 jlvg1-1-5 汉语的映射是"有"。"无,有"是汉语的万能高层概念字,所以两者都是构成四字词的大户。汉语有 10 个四字词特大户,依次是:不,一,无,人,心,之,大,天,如,风。由它们构成的四字词约占成语四字词的三分之一。四字词所体现的结构美,在宋词和楹联中得到了充分的发展,节点 7-8 就是为了表达这一结构美而专门引入的。

英语没有 jlvg1-1-6 的精确映射词,视不同情况,分别采用"not,no,nil,nothing,without"以及前缀 un,in,后缀 less 等表达手段。这么多的"手段"汉语用一个"无"字都能表达,为什么?是西语的词性语法约束带来的负担,是汉语由于无这一约束而赢得的运用自由。从理解来说,单义而自由运用的词利大于弊。反过来说,多义又不能自由运用的词,则弊大于利。我之坚信汉语更易于理解,基本根据之一就是这一点,"无"字是一个小小的例子。

这里应该谈一下"无所"的问题。这一类的组合,词典不作为词收录,因为根据语法学对词的定义,它们不是词,是语素。我们尊重词典的权威,把"无所"之类的组合另外起了一个名字,叫做搭配。其实是没有必要的,"无所"就是词,而且其独立性并不是 3,而是 0,1 之间。这个词的意义就是"没有逻辑对象",精确的映射符号是 jlv1-1-6|lg1-0。词典里收集的"无所"四字词:无所不包,无所不为,无所不在,无所不至,无所适从,无所事事,无所畏惧,无所用心"中的"无所"全是上列层次网络符号所表达的意思(当然,如果没有引入 l1-0 节点,这些"无所"的共性是提取不出来的)。虽然现代汉语把"无所"延伸为"没有什么",但你如果说"无所喜爱,无所参考"等等,也不至于受到复辟古文的指责。"无所"完全有资格登上词的大雅之堂,否则,"无限,无效"等等都不够资格了,因为它们与"无所"的语义组合结构是完全一样的。

这里需要对〈概述〉第 14 章作一点补充说明。l0 l3 的基本功能是用于语句要素或语义块的指示符号,同时,它们又是语言逻辑的基本概念,〈概述〉对后面的这一点阐述得很不够。句类分析的基本概念是 E,A,B,C;用汉语来说,就是特征要素、作用者、对象、内容;用层次网络符号来表达,就是 lg0-0,lg0-1,lg0-2,lg0-3。往低层次说,作用对象是 lg0-2-0-0,效应对象是 lg0-2-3-0。语法学的谓语、宾语、补语大体与特征要素、对象、内容相对应。主语这个概念则只能说大体与作用者及对象相对应。只有施事这个概念才与作用者完全对应,而受事这个概念是作用对象与效应对象之并。不过,应该说明,施事和受事已不是语法而是语义的概念。显然,语法学的主谓宾补概念不适合于作概念层次网络符号体系的概念基元,主语与宾语的可交换性就足以表明这一点,我们不得不专门引入上述 E,A,B,C 的概念系列。

l1 行概念中的 0 分行也是专门引入的,称之为广义逻辑对象,它包括对象和内容,用

lg1-0 表示。从技术结构来说,这样安排不甚合理,逻辑概念本体层的两层的约定必然会带来一些弊病,这几乎是无法避免的。l1 行的设计目的是指示辅要素,但要素的主辅性是可以转换的,对此,尚未找到简明的表示方法。l1-0 的上述定义只解决部分问题。辅要素中的途径、手段、条件、工具四大项如何安排,屡经周折。现在的安排方式仍只能说差强人意。途径、手段、条件安置在 12 行的 12-9,12-10,12-11(注:这三类概念最后独立成为一个概念类别,并加上了工具,命名为综合类概念,符号为 s),与 6,9 行相对照,它们是 12 行独有的三个分行。为什么把途径排在最头里,而舍去工具?因为,途径是战略性的,而手段是战术性的。工具未与途径、手段、条件并列,因为它不具备并列的资格。工具既可视为手段的一部分,也可视为条件的一部分,另外,表达工具的抽象概念很少,所以,最终决定将抽象的工具概念映射为 lg1-2,具体的工具可考虑用 wl1-2 来表达,汉语的"工具"一词是两者的并。

l 后面跟五元组符号表示它是一个实词,单用(不跟五元组符号)表示虚词。不过,在直接表示中有可能未严格遵守这一约定,如果发现这种情况,必须改正。

顺便一说,j 的后面可省去 g,因此,ji-k 与 jgi-k 等价,但其他的五元组符号不能省。

1994.12.4

访北大老友,在秉乾家午餐,在子钊家晚餐,畅谈竟日。听子钊宏论,乃一大乐事。今后拟多译而少记,太累了。

1994.12.12.星期一

译"瞎 狭,侠,峡"。

"瞎"字涉及到知觉或感觉的表达。感知的映射符号是:r6-2-0-2。6-2 表示动物(包括人)的生理本能,0-2 表示反应,感知就是动物的生理本能反应。注意!这里类别符号必须用 r 而不能用 g,否则就糟蹋了汉字"觉"的奥妙。"觉"不是指反应本身,而是指反应的效应。在心理学里,将 r6-2-0-2 分为感觉和知觉,感觉是知觉的基础,知觉是感觉的深化,两者不是对偶关系,也不是包含关系,接近对比关系。两者的区分可近似表示为:

感觉　　r6-2-0-2-12-2-1

知觉　　r6-2-0-2-12-2-2

意思是知觉的"值"高于感觉。目前两者都采用高层近似表示:r6-2-0-2。这个近似是足够的,因为自然语言并不也没有必要严格区分知觉与感觉。

在〈日记:1994.11.8〉里我提到语言有十几条联想主脉络(清单),当天只概述了两条,一是专业活动清单,二是行为清单。今天谈一下反应清单之一。

6-2-0-2,6-5-0-2,7-1-3;2-1-11,i-7,i-8 (i = 6,9,12)

jw6-2—

联想清单的源头都是基元概念,其中作用基元的每一个二级节点都有相应的联想清单。上次讨论的专业活动和行为清单的源头分别是 0-0 和 0-4,以后会讨论 0-1,0-3 清单。0-0 清

单不止一个,专业活动只是其中之一。0-2 也不止一个,今天讨论也是其中之一。这一点不难理解,因为 0-0,0-2 概念是整个作用效应链的基石。今天讨论的反应清单以生理反应和心理反应为主,不包括理智反应,后者另构成一个反应清单。

这个清单的前三项分别代表生理反应、本能反应、感情反应。后继的三项代表这些反应的主要运作、宣泄与需要。第二行是参考项,与前两个清单的最后两项对照一下,就不难理解,jw6-2— 代表有关的感觉器官。

生理及心理反应的基础是视和听,这两个概念在 6-2-0-2-k,2-1-11-k 两个低层节点里都作了安排(但 k 的具体设置未定)。

接俞平信,为方代复,我仅附小诗一首,录于下:

欣知百事尚开心,遥祝明年更有成。

巨变风雷窗外事,我行我素一痴人。

1994.12.18.星期四

近一周所译,仅就其要者,简略补述如下:

(1) 关于广义位置、广义方向的概念,最后确定映射符号如下:

广义位置　　j0-1-4

广义方向　　j0-1-8

广义位置除了狭义空间位置 j2-1-0 之外,还包括:转移的空间序列 2-0-4,一般状态 5-0-0,层次 5-5,等级 5-6,数的位 j3-1。汉语的"位"字大体符合广义位置的意思。

广义方向除了狭义空间方向 j2-1-8 之外,还包括:过程的趋向和转化 1-3,转移的定向 1-0-9 及传输 1-0-10,效应的推动与抑制 3-6,人的希望与追求 7-1-2 与 11-0。汉语的"向"字大体符合广义方向的意思。

"位"有上下之分,汉语词汇"上面,上边,下面,下边"就是广义位置的上下,所以词典里都详尽地给出多义解释,但其多义范围不外广义位置。因此,将它们分别映射为 j0-1-4-1,j0-1-4-2 似乎更为简明并便于理解。同理,多义词"上去,下来"也可映射为单义词 jv0-1-4-1,jv0-1-4-2。当然,考虑到这两个词还有充当后缀的语法功能,把它们映射为上列符号与 hv 之或更为确切。

在词典里,"上下"两字是义项最多的冠亚军,分别高达 22 及 25 项。但其中一些义项可并为广义位置表示。上下有 v9-1-1-k,k = 1,2 义项,即"开始某事"及"结束某事"的意思,从打斗到专业活动都可用,前者可简单地说"上",后者可说"上项目,坚决下,×上×下"等等。词典里有"到规定时间开始(结束)日常工作或学习等"的解释,举例上(下)班,上(下)课,这似乎有些不妥。"到规定时间"是"班,课"的属性,与上下无关,所以,字义库中无此义项,它已包含在上列义项中了。

"上下"的词汇甚多,大部分未填,留作大家练习之用,这是难得的教材。要运用层次网络理论搞自然语言理解,不亲自填写部分字义和词义,难以把握要领,程序设计更难有所创造。

(2) 感知 6-2-0-2 的低层设计：

6-2-0-2-8　一般感知，含触觉

6-2-0-2-9　视听觉

6-2-0-2-10　味觉

6-2-0-2-11　嗅觉

视听觉 6-2-0-2-9 分别与 jx1(包括 jwz1，jxz1)，jx3 强关联，味觉与食物的属性 px6-5-2-2 强关联。一般来说，6-2-0-2 的联想脉络首先通向两类高层概念节点。一是基本物性 jx，例如，一般感觉中的冷热，潮湿，干燥等等与温度 jx0、气候 jx5-1-8 强关联。二是一般环境 r5-0 及自然环境或景象 r5-0-8，后一点是不言而喻的。这里不妨提请大家注意，环境这个概念必须采用类别符号 r，而不能采用 g，在五元组想法的形成过程中，“环境”及“概念”等是重要的启发者。

味觉的“酸甜苦辣咸”可表示为：

酸　zu6-2-0-10-12-5-1

甜　zu6-2-0-10-12-5-2

苦　zu6-2-0-10-12-5-3

辣　zu6-2-0-10-12-5-4

咸　zu6-2-0-10-12-5-5

嗅觉的香臭则可表示为：

香　zu6-2-0-2-11-1

臭　zu6-2-0-2-11-2

这些体现了对偶性及对比性概念在中层设计的典型应用。

(3) 关于 0-4 的联想脉络

以前曾简述三个联想脉络：专业活动脉络、行为脉络、本能反应脉络。这些联想脉络的源头分属于基元概念 0-0，0-4，0-2。曾经指出，来于这些源头的联想脉络都不只一个。这里介绍源于 0-4 的第二个脉络：0-4，j4-2，j6-；3-6-2；13-，7-1-1-k；12-11。

脉络中的节点 7-1-1-k 代表对自己的态度，k 的取值待定(曾定过，作废)。约束的基本内容是范围 j4-2 和度 j6 的概念，约束涉及的效应对象主要是人类的行为和态度，所谓条件，本质上就是约束。所以，这些概念形成一个天然的联想脉络。理智的内涵之一就是人要懂得约束自己，而禽兽不具有这种本能。将“遵守”作为行为的第一号对偶性概念赋予节点 g13-0-0-1 就是基于这一考虑。东西方文化的根本区别之一就是东方重视 0-4，而西方重视 0-3。

(4) 关于地域表示的说明(来于对“县”字的说明)

基本概念的“物化”及“人化”，是运用层次网络方法理解物和人的根本途径之一，实际上就是将物与人向基本概念挂靠，如同向基元概念挂靠一样。其中，地域的挂靠最为复杂，原一般表示式中的符号未统筹安排，容易引起混乱。主要是两个问题，一是挂靠层次的标志，

二是一般表示式中通用符号与非通用符号的区分。

在本体层与挂靠层之间,已有天然约定,但地域的语言概念甚多,后本体层又必不可少,因此必须加一个挂靠层结束符号“ * ”。根据这一设想,我建议在“ * ”之后将第一级层次符号的 1 3 赋予水域,4 以后赋予陆地,因为陆地的形态比水域丰富。

关于地域的表述,这里需要写几句话供具体设计参考。地域 wj2— 定义为生命体赖以生存的空间。地域基本上是两类,陆地和水域。虽然鸟类及其他可飞行动物常翱翔于天空,但其生存基地仍是陆地和水域。

(5) 关于组合规则 2-2 的补充说明(来于“献”字的说明)

结构方程的组合规则已约定了 5 个,分别称为 2-1,2-2,2-3,4-1,4-2 规则。这些规则用于结构方程 i1-i2-i3-i4 的 i2,i4 定义。2-m 规则用于 i2 的定义,4-n 规则用于 i4 的定义。其中 2-2 规则专门用于 6,7 号结构方程。这些规则的含义在〈问答 34〉中已有详细说明,这里补充的是隐含的意义。

就 7 号方程来说,如果该方程构成的 i2 确实是 vr 型概念,则意味着已具备内容,因而在构造语句时需要补充对象,对 6 号方程可作相反的推论。仅就这一信息要求的表示来说,为 4,5 号结构方程设计的 2-1 规则已提供了表示手段,为什么要另行制定 2-2 规则?在庐山时,由于手边找不到适当的“证据”,只好回避了这个问题。现在,“献”字提供了很好的例子,可以把话说透了。

由 4,5 号结构方程构成的词汇,对于跟随对象及内容的要求比较明确,如规则 2-1 所示。但 6,7 号结构方程构成的词汇不具有这种明确性,部分修改 2-1 规则另立 2-2 规则就是为了表达这一区别。大家也许注意到,2-2 规则定义中的“要求补充”本身就是模糊的,事实正是如此。这个“补充”,不是指“对象或内容”,而是“对象加内容”,如果仅是前者,利用 2-1 规则就足够了。但这个“加”又不像 2-1 规则那么确定,实际上是可加可不加。现在,你可以看到具体的例子:

i2	i4		
2	1	献花,献茶	明确地需要效应对象
5	1	献丑	一般不需要补充对象,但不那么明确
6	5	献计,献策,献身	需要作用及效应对象,但不那么明确

这里值得对 i4 提供的类别信息作一点补充说明。i4 = 5 表示作用及效应对象优先于人和事,这里的“人”,通常优先于 p12 ,而不是 p。而“献花,献茶”的优先对象显然是 p,而不是 p12。看来这一重要信息可以在词汇级给出更明确的指示,可惜现定的表示空间不够了。

最后,4-1 规则(上面的 i4 取值即来于该规则)有一个遗漏,就是“物和事”的组合,现定义该种情况的 i4 = 6。

1994.12.22.星期四

译“相”。

“相”的映射符号是 u4-0-8。我没有把它定义为 u4-0-0 或 u4-0-9,请体会这种细微之处。在〈概述之关系章〉中曾谈到关系的单向性及双向性,这将用层次符号 4-0-8-k,k = 1,2 来表示。“相”的第二个义项是 u4-0-8-1,这是“相”与“互”的不同之处,“互”一定是双向的,而“相”既可以是双向的,也可以是单向的。英语词汇里只有“互”的概念,没有单向“相”的概念,似乎汉语的词汇表达高明一些。可是汉语又把“相”搞成模糊的两可概念,高明变成了理解的困难,这是非常普遍的语言现象。为什么?因为语言与艺术有天然的亲和力,有时为了满足艺术性的要求,不得不牺牲科学性,语言绝不作反向的牺牲。因此,仅仅从科学性上来评价不同自然语言的优劣,是不懂自然语言的表现。

“相”有与“互”共性的成分,但我们并没有仅将“相”映射为 u4-0-8-1,还映射为 u4-0-8,这是试图突出“相”的层次性模糊。关系有三项基本特性:相互性(方向)、紧密性(距离)、传递性(序),它们构成关系节点 0 分行的三个低层节点。“相”作为 u4-0-8 的映射符号,不仅意味着它具有双向及单向的模糊,还意味着它隐含着传递的意思,上面说的细微之处就包括此层隐含意义。词典的解释里并没有这一条,但这一含义在“奔走相告”里是非常明显的。这类精细的含义在字义库中都可以表达。但我宁愿劝告程序设计者先不要去把握这些精确含义,而先深入领会概念表达的层次性和网络性区别以及基元概念 0 行和非 0 行的区别,并转化为程序运作的启发性知识或规则。概念层次网络理论如果没有这项转化,当然只能是一个束之高阁的漂亮理论。

“相”字的高层性或模糊性决定了“相”字的组合能力远大于“互”字。我曾列举过汉字的四字词大户,“相”字不在前十名,但肯定在前二十名之内。

我在〈日记:1994.11.8〉中曾说到“联想主脉络大约一共有十几条”。但当天的日记草草结束,对这两句话并没有作出交代。“十几条”之说根据何在?就是作用的五个二级节点加其他主体基元概念的 0 分行再加几个基本概念。当时我脑子里所谓“粗略清单”就是指这个。

节点 4-0-8 是联想脉络的源头之一,但不是主脉络。以前曾谈及四个联想主脉络,并曾命名第一号联想脉络,用意是突出这些主脉络的语用重要性,其实没有这个必要。而将源头节点放在第一个倒是必不可少,所以,应将前面的联想主脉络书写格式改为:

专业活动主脉络	0-0,10
行为主脉络	0-4,13
本能反应主脉络	0-2,6-2-0-2
客观及理智约束主脉络	0-4,7-1-1

现在我们碰到了一个以 4-0-8 为源头的脉络,其格式内容如下:

u4-0-8-1,2-4,9-7-1,11-3,11-4,10-4,10-7,10-5-k,j10;7-1-1,7-1-3

这个清单的第一个节点前加了层次符号 u,不仅用以体现联想脉络的主次之分,还表示该节点具有“源头”的属性。当然,一个清单内部的具体特征还有待作出更细致的约定,这是后话。这个清单表示,交换及替代,交往,竞争及协同,军事活动,教育,法律活动中的诉讼

(其 k 值未定),比较判断等都具有相互性,态度及情感也具有相互性。前者的相互性是绝对的,而后者的相互性是相对的,故两者用分号“;”分开。这个清单未囊括一切。清单外的节点并非一定不具有相互性,只不过程度较弱罢了。

1994.12.23.星期五

译“详,降想向”。

“详”的第一义项“详细”暂时映射到 u9-2-3-0-2,其对偶性概念“简略”为 u9-2-3-0-1。通常对偶性概念的 1,2 有积极与消极之分,但这里的“简略与详细”无此含义,与“详细”同一映射符号的“罗嗦”倒是精确符合该符号的通常意义。关于对偶性概念的这一特性目前未找到简明的表达方式。

“想”字具有与“相”(xiang1)字类似的模糊性。字义或词义的扩展(从而形成模糊)不是任意的,绝大多数情况是在联想脉络内扩展。汉语主要以“字”为基点进行扩展,沿着联想脉络尽情扩展字义的模糊区,然后又以双字组合的方式来尽可能消除模糊,获得模糊度小而信息量更大的词。这样,一个概念的联想脉络就往往凸现在某一个字的前后组合词的集合里。汉语的这种知识结构堪称奇妙,硕士卡的数据结构来于这一启发,层次网络符号的基本设计实际上主要是依靠这一启发。我有时想,对人类大脑里知识结构的探索也有可能得益于这一启发。

“想”字的义项集合就是一个典型的小联想脉络,它涉及下列概念节点:

v8-0,7-1-2,11-0,6-8,9-8

在这个小联想脉络里的代表性汉字有“想,思,念,忆,记,望”等,每字各有侧重,但都在脉络内扩展。当然扩展范围有很大区别,“想,思”的范围最大。

“向”字是前面所说广义方向的典型代表。当然,在字义库中,除了这一高层义项外,还要给出具体的义项,以供某些取特殊含义的组合词选用。例如“人心向背”里的“向”就取特定含义 g12-1-3-9-1,它是广义方向联想脉络里的一员。

“向”字的第一义项是〈日记:1994.11.22.〉里所说的逻辑对象指示符(l0-2;l0-3)。无论是独立使用和在组合词中,绝大多数情况是选用这一义项。

“向”字的上述两义项在与其他概念组合时以前习惯于用 7 号结构方程,这是由于当时对该方程有偏爱所致。实际上采用 1 号方程更为适当。但为 1 号方程配置的 2-3 规则过于顾及全面而造成重点不突出,所提供的信息过于宽泛,可以作出更好的设计,目前只好“委曲求全”。另外,4-k 规则原来 k 值只约定到 5。上次对 4-1 加了 6,现在 4-2 也加 6,表示该词“优先于充当特征语义块的非核心成分”,而原来的 5 自然就是“非特征语义块”的相应成分。由“向”字的定义可知,它一定可与下列方位“字”——“上下前后左右东南西北”组词,词典只收了“向前”(我们的词库加了“下前后”),其他九个属于我们常说的“新词”。“新词”基本类别是一个重要话题,我将在适当时机详细讨论。这里只说一点,由“向”与上列方位字构成的词不仅具有副词特性,还具有 vr 特性,词性这个语法概念对这类词就显得无能为力。这只

不过是词性概念不适合于汉语的例子之一。语义结构方程采取的做法是:通过方程规则、类别符号、独立性三者加以综合表述。以这里的上列“向”字组合词为例,它们的结构方程是 1-4-i3-6,类别性是(v,u,r) = 1,独立性 = 0,这就把上列词汇的语义、语法、语用知识都包含进去了。词典编者实际上感受到了词性说的这种困境,我认为只收“向上”一词与此有关,而且,对“向上”释义时,不提副词的意义,不完全是疏忽,也与这种困境有关。

1994.12.24.星期六

译“消”。

“消”字属于 v3-1-2。3-1 这个节点最初的编号不是 1 而是 6,最初的 1 是现在的 2。这个改动极为必要,因为虽然人类社会发展的基本效应是狭义的利害,但宇宙发展的基本效应却是“生与消”,达尔文的进化论就是关于生物的“生与消”理论。

3-1 概念联想脉络之广,虽不宜擅用“各概念节点之冠”的说法,但名列前茅是毫无疑问的。这里只列举四个关系最密切的脉络点。

3-1;1-4,3-5,6-6-1,6-8-0

过程的 1-4 与效应的 3-1 是同一概念的两种角度表述,类似的典型例子还有 1-3 与 3-0-9,3-9 与 4-1,3-10 与 4-6 等等。3-5 与 3-1 强交式关联,3-5 的概念内涵是最近一次大学生辩论竞赛的命题“破与立”。“立破”自然伴随着“生消”。之所以分为两个节点,主要是两个考虑:一是两者的作用色彩及过程色彩有很大的差别;二是希望建立各有侧重的联想脉络,这就是说,3-1 的侧重点是从无到有及从有到无,3-5 的侧重点是质的变化,而质的重大变化就表现出“生与消”。

人类创造物质财富及精神财富的一切活动,人类的思维活动,包括记忆,本质上都伴随着“生与消”。这里,仅选取 6-6-1(自然包括 9-6-1,12-6-1)和 6-8-0,也许过于迁就了汉字的联想特征。以前写〈概述〉和〈问答〉的时候,各种环境约束常使我苦恼,有时不得不采取回避策略,不得不为“高明”的“闪烁其词”动一番心思。〈日记〉的形式可免除这一苦恼,自由驰骋,但求“新意”,不求成熟,错了就改。

1994.12.25.星期日

译 xiao1 的其他诸字。

“销”义项之一是 v5-2-12-2-4-9-9-2,看起来层次很多,但其高层表示实际上只有两层,即 5-2,12-2-4,程序设计者必须熟悉这个窍门。后续的 9-9 是低层表示,2 是中层表示,9-9 表示物的交换,2 表示输出。因为,现代的销售也包含信息和知识,贸易有“商品,服务,经常项目”三种定义,语言概念必然随着社会生活的发展而发展,层次符号的设计要考虑到这一需要,第二个 9 也可改 8 即是一例。“销”与“卖”的区别在于“销”的动态性更强。

“萧”是一个表达景象属性的字,组合词有“萧然,萧飒,萧瑟,萧森,萧索,萧条”。景象自然是 r5-0-8,是自然状态 g5-0-8 的效应而已。故上列诸词,除“萧条”以外,都记为 ru5-0-8。

也许现在就可以记为 ru5-0-8-2,表示是一种消极的景象,而‘蔚,旺,盎,勃’诸字则是 ru5-0-8-1。景象密切联系于物,将它映射为 rw5-0-8 亦无不可,这样,关于景象属性的词汇可映射为 rx5-0-8。对这两种方案的选择,希望听取大家的意见。

“硝烟”一词自然是映射为 rw0-0-9,因为,燃烧的映射符号是 v0-0-9。

“消磨,消闲,消遣”等词,映射到 v5-0-10-2,其对偶概念是‘忙”。

1994.12.26

西语的词根相当于汉语的偏旁及充分基元化的汉字,是研究联想脉络的重要依据。人类创造了几千种语言,几百种文字,这些语言和文字的共性表现在‘词根”。语言和文字中有形的‘根’对人的实际思维过程作用甚微,那是因为大脑已将这些有形的‘根’转换成无形的脉络。脉络就是思维的程序,语句表示式不过是这一脉络的第一个可操作理论模式。电脑也需要完成这一转换,从这个意义上说,层次网络理论真是万里长征刚迈出了第一步。

1994.12.27

译‘小 效,校”。

“小’属于基本概念 j4-0,其映射符号有三个,一是 jgu4-0-12-3-1,二是 jgu4-0-12-2-1,三是 ju4-0-12-2-2,即一个三分对比及两个两分对比。相应的概念分别是‘小中大”,“小大”及‘微与小”。“小’的另一常用义项是q|j0-0-12-2-1。汉语对基元概念 j4 的表述值得玩味。j4 的三个二级节点都安排了两个常用汉字,j4-0 是‘大小”,j4-1 是‘多少”,j4-2 是‘内外”。

“小’字构成的词汇多数与上列词义挂钩,但也有少数例外,如‘小人,小丑”,它们的映射符号是 pj8-3-2。人的划分标准很多,j8 的各二级节点都可作为人的分类标准。中国以前的传统重视标准 j8-3,即所谓君子与小人。这本来是一个很深刻的标准,不幸后来给加上了封建的恶谥。诚然,在封建时代,君子与小人的含义有阶级歪曲的色彩,有些偏离 j8-3 的原意,但绝不是这个词实际语用的主流。可惜,这个合理的内核连同它的污点一起被彻底抛弃了。以至人们陷为小人而不自知,不以小人为耻。我曾在一个郑重的场合郑重地警告某君是‘标准小人”,而某君含笑受之,因某君深知“不做小人,就做不了能人”的现代‘哲理”。

喜遇‘效’字,因为它可为五元组概念提供强有力的合理证据。“效应”是 g3-0,“效率”是 gz3-0,“效果”是 r3-0,“效益”是 r12-3-0,“效力”的义项之一是 rz3-0。试想,如果不引入 r,z 的类别概念,怎能将上列概念的异同表达得如此简明?真是难以想像。

1994.12.28.星期三

译‘笑,孝 歉 沏 泻 谢,懈,泄”。

“笑”的映射符号有:v6-2-2-3,v6-5-2-3。“哭”相同。两者的高层符号是一样的。它们作为一种信息转移的定义,既是生理本能的表现,又是人类特有的本能,因为只有人类会发出“嘲笑,奸笑”之‘笑’及猫哭耗子之‘哭”。至于两者的区别,我倾向采用复合结构的方式,而

不采用低层设计的方式加以表达。

信息载体有两种基本类型,音和形。笑和哭兼而有之,以音为主。人类转移信息有两种基本方式,说和写。说是对音的运用,写是对形的运用。但汉语的"写"字是一个高层概念,其精确映射符号是 v9-2-3-8,它包括写字的"写",写真的"写"(即绘画)和一般的描写。

1994.12.29

译"心"。

"心"是 g7-1-3,g7-2-0,g7-2-2 的精确映射,就是说,它是这三个节点的高层表达。因此,"心"字既是双字词的大户,又是四字词的大户,就一点也不奇怪了。"心"的本源语义心脏 jw6-3一,组合词不少,涉及心脏各部分的表述。

抽象"心"的组合词不少采用内容结构,如"心窄,心毒,心烦,心诚,心急,心焦,心软,心酸,心硬"。按语法学的分类,这些都是主谓结构,相应于我们定义的 9 号结构方程。这里顺便说明一下,9 号方程是后加的。我原来的想法是将主谓结构分别纳入内容和反偏正结构里去,7 号的第一概念 X,容许所有概念类别,包括五元组和 l,j,w,p。后来在制定结构方程的二级及四级规则时,感到 7 号方程的包容度太大,不利于二、四级规则的制定,于是,决定保留主谓结构,并命名为 9 号方程。但当时对 7 号方程的偏爱心理仍未彻底消除,觉得像"地震,雪崩"这样的词汇固然是比较典型的主谓结构,但"心窄,心软"这类的词汇还是用内容结构较好,后者的 u 特征非常明显,而前者完全不存在这种特征。介于两者之间的情况还是用 7 号方程表述比较适当。

新旧的概念纳入 j7 的二级节点之一较为明智。这样,与原始设计相比,一共增加了三个二级节点,即"正常与异常","简单与复杂","新与旧"。这三个节点是分别从 j7-3,j5-1 和 1-4 独立出来的。这一步很有必要,因为三者确有"属性之属性"的特征。将来还会补充新的成员么?估计可能性不大,因为在这个问题上的全部疑点已不复存在了。

1995.1.4

译"可能似同异差(cha1)"。

"可"的第一义项是 jlvu1-2,既有"可能"的意思,又有"能够"的意思。我在〈日记:1994.11.26.〉中曾将"能"与"可"并列,作为这一高层意义的代表字,这是一个典型的粗枝大叶错误,"能"不具有这一资格。

"可"最灵活的用法是与 v7-1 搭配构成 gu 型概念,词典对这些搭配搜集得比较完全。有关 v7-1 特别是 v7-1-3 的各种表达,可通过对"可"行的搜索得到相当完整的联想。层次网络的大部分一、二级节点,在汉语里都可以找到像"可"这样的代表字,这类字可命名为"联介字"。建立第一期字义库的附带任务之一就是找出这些联介字,将随时在日记中说明,"可"字是第一个。

"可"字上述意义的映射符号是:(q↓gu)|jlvz1-1-1,g7-8-1。符号的第一部分表示:"可"

作为前缀的语法功能是使组合词 gu 化,具有与〈日记:1994.11.24.〉中的“化然性”等字同样的语法功能。我在写东西的时候,以达意为主,为图书写之便,对表达的细节往往注意不够,例如〈日记:1994.11.26.〉中关于 jl1-k 的示例都省去了五元组符号,这一点,望读者留意并予以体谅。

“同异似”的语义地位安排在 jl0-0,我有时也怀疑是不是太高了。但仔细想来,它们确实当之无愧。思维的起点毫无疑问是从比较开始。比较的基本结果就是“似同异”。可以说,一切判断都是以“似同异”为基本出发点。三者互为对偶,根据我们对对偶性表示的约定,“似”为 0 是不言而喻的,因为,“似”意味着有同有异。但同异的取值则曾反复多次,最后取同为 1,异为 2,这与改定“特殊”为 1,“一般”为 2,“珍奇”为 5,“普通”为 6 是一致的(后者属于基本概念 j7-3)。

“似同异”的类别符号理所当然的是 rv,而不能是其他。“差 1”的类别符号则是 zv。在层次特性上,两者的差别也十分鲜明,前者是中层概念,后者是高层概念,其层次符号分别为 0-0-k,0-0。

词典对“差 1”给了两个义项,一是“不相同,不相合”,二是“甲数减去乙数剩余的数”。对“异”有关义项的释义是“不相同,有差别”。显然,“异”与“差 1”的第一释义具有可互换性。从类别性和层次性的观点来看,这个释义不妥,因为它模糊了“异”与“差 1”的不同类别性和层次性。这个差别是不容忽视的,因为它与语用性有密切关系。汉语将“异同”作为对偶性概念来使用,“求同存异”、“大同小异”等精彩的汉语表达方式,都是生动的例子。而对“差同”就不能这样运用。词典采用了可以互换的释义,对“异”字又不加“对偶”性的说明。为什么?我猜想,这与编者们当时不能不接受“一分为二”的狭隘对偶概念有关。

词典“差 1”第二义项是第一义项的特例,字义库一般不为这种特例立项。这一做法的利弊如何,请大家思考。词典漏注了贸易顺差、逆差中的“差”义,字义库中补上了。上述“似同异差 1”四字只涉及 jl0-0 的 rv 及 zv 类别概念。jl0-0 的 vz 类别概念有另一个三重对偶,与 jlvz0-0-4 相关的有“当,匹,敌,抵消”等;与 jlvz0-0-5 相关的有“胜,强,超,高,赚”等;与 jlvz0-0-6 相关的有“差 4,弱,欠,低,赔”等;在数的意义上还有“等于,大于,小于”等。这些字的含义,或多或少都与 jlvz0-0 有关,但必须与相应的基元概念并合,才能获得完整的理解。例如,“抵消”联系于效应,“赚赔”联系于买卖。这些并合是否简化为 jlvz0-0 的低层设计,到建相应汉字的字义库时分别讨论较为适当。这里先说一点,就是 jlvz0-0 的动态表达比较复杂,因为它有一个方向问题,其一般动态表示“逼近”可映射为(v5-2,jlvz 0-0),但“赶,追,超”就不同了,它们是从 jlz0-0-6 向 jlz0-0-4 或 jlz0-0-5 转化。有关表达符号都已具备,但具体表达方案有深浅之分。在〈日记:1994.11.26.〉的第一个“第四”中曾谈到这个问题,请大家作进一步的思考,我热切盼望得到反馈信息。

1995.1.5

译“假如若则”。

前三字的主要义项是 l11-1-1-2-1,“ 则 ”的主要义项是 l11-2-1-2-2。这里顺便说明一下曾经用过的另一表示符号,前者是(l11-1,lg6-0-1-2-1),后者是(l11-2,lg6-0-1-2-2)。这里的两个 lg6-0 分别表示“ 前提 ”和逻辑意义的“ 结果 ”。现在字义库中采用的表示实际上是一种简化。

大家知道,l 概念的设计曾经历过一次大的调整,主要内容是将基本判断逻辑概念独立出来成为 jl 类。至于“ 与基元及基本概念挂靠 ”,“ 本体两层 ”,“ l 分为三级,并用第一层的数字 0-3,4-7,8-11 予以区别 ”这三个基本点是始终一贯的。局部调整限于本体第二层,至今尚有待定部分。不过,在建设词义库的过程中,随着“ 简化表层 ”手段的扩大应用,原有的一个想法也发生了变化,这就是放弃了以 l4 为语义块并合标志的想法,因为并合不如用类别符号 hq 表达更为自然简明。这样,以后也许见不到 l 之后直接跟数字 4 的机会了。(注 :后来又恢复了,见【1】和【6】。)

1995.1.6

译“ 把,被,使,于,相,互 ”。

“ 把 ”字和“ 被 ”字是汉语里两个最重要的 l0-k 概念,包含了 E、A、B、C 的指示。有趣的是,两者分工明确。“ 把 ”字用于对象及内容 B、C 的指示,“ 被 ”字用于作用及作用者 E、A 的指示。其映射符号分别是:

把　　(l0-2-0-0 ;l0-3-2-0 ;l0-3-3-0)

被　　(l0-1 ;(l0-0-0-0,v7-9-2|lg0-1) ;qv↓ug)

为加深读者对这些重要概念的印象,这里将词典中有关“ 把 ”字和“ 被 ”字上述义项的示例转引于下:

作用对象:把衣服洗洗。把他急疯了。

转移内容:扭转身来把话讲。把这封信贴上邮票寄出去。

效应内容:领导人民把身翻。

作用者:　这部书被人借走了一本。

作用:　那棵树被(大风)刮倒了。

词性转换:被害人

逻辑符号,特别是要素指示逻辑符号有广义与狭义之分。狭义逻辑符号带挂靠层,广义不带,只有本体层。但自然语言对这个界限的把握十分模糊,这就是我常说的自然语言的弱点,或理解的难点之一。但“ 把 ”和“ 被 ”基本属于狭义型,语义模糊不大。词典对“ 把 ”字上示义项的解释用了 300 余字。如此大的浓缩,也曾使我产生怀疑,这么几个符号就能表达“ 把 ”作为介词的丰富内涵么?在接下来译“ 把 ”的其他义项之前,呆坐良久。华罗庚先生关于“ 厚薄 ”的名言使我的不安心情稍减。当时我反复思考的是,“ 把 ”作为对象指示究竟是狭义还是广义或兼而有之?词典的例子都属于狭义,但“ 把他叫来 ”、“ 把马列主义的普遍真理与中国革命的具体实践相结合 ”里的“ 把 ”,不是很有点广义的味道么?自然语言的符号设计虽然不

把科学性放在第一位,但终究有其特殊的科学性。语言的四大主角及七大配角需要配置指示符号,以利于角色位置的灵活安排,这就是语言的基本需要。这些符号必有广义与狭义之分。如果都是狭义,看似科学,但与语言的灵活本性相违背,所以。部分广义、部分狭义的“设计”,正是‘天工’之巧。我想,通观语言逻辑指示符的全局以后,会看到汉语的‘天工’是将‘把’与‘被’作为狭义指示来使用。在我举的两个例子里,“把”不正是用于体现作用的隐含意义么!

“把”字量词义项的映射符号如下:

(h(j3-0-8-1;xjzz2-0-;xjzz5-1-8))↓z

pw6-5-4-5-1/zz

第一义项的意义是:作某些概念节点的后缀,产生该概念的模糊值。示例说明如下:

j3-0-8-1	如个百千万等。
xjzz2—0-	各种面积单位及长度单位,如亩、顷、尺等。
xjzz5-1-8	各种重量单位,如斤、两等。

对 j3-0-8-1 中的概念,应将‘十’除外,这一细节也不难表达。但重要的是,大家要领会这些符号所试图体现的高层意义:zz 代表量词,z 代表模糊值,↓表示效应并落实到效应。一言以蔽之,就是‘把’将量词转化为模糊量词。词典里一长串的文字说明,现在变成了计算机能够理解的简明符号串。在这里,就是依靠引入了效应及其结构符号↓,和五元组 z 的概念。第二义项不言自明。

第一义项若省去外层括号,似乎不仅无碍于理解,且有利于运行效率,是否如此,请软件设计者考虑。

1995.1.7.星期六

补写‘使于’日记。

“把被使于”这一组汉字都是语义块切分的明确标志,但提供标志的方式各有特色,所以这里放在一起来讨论。

“使”字目前装了四个义项,映射符号如下:

1.(v0-0,lg0-2,lg0-3)

2.v9-4-5-1

3.gv10-1-4

4.(p12-0-1;(j8-2-1,j8-3-1)/p1-0-11-12-5-2)

这里的第一个义项的三项符号都是高层意义。实质上义项 1 本身主要起一个指示符的作用,其第二项表示‘使’后面必须直接跟对象语义块,其第三项表示对象之后必须直接跟内容语义块。这就是说,义项 1 与 4 号结构方程的 2—1—4 规则等效。从义项 1 的本意来说,也许将它的第一项改为 v3-0 更为确切,但是,这是绝对不容许的,因为‘使’后面的对象和内容不能加‘的’字并合成一个语义块。它是汉语作用效应句的一种特殊指示符。在书面语理解

处理时,如果在句首分离出一个单独的“使”,句类分析的第一步就大功告成了。因为,在“使”的所有义项中,不仅只有这个义项具有 0 级独立性,而且在句首的条件频度最高。关于义项 1 的上述特征,建议读者记住“虚心使人进步,骄傲使人落后”这句名言,就能有所体会了。

义项 3、4 的书写格式必须考虑到与词义库表达的配合,因此,也顺便一说。

义项 3 的词典释义是“奉使命办事的人”。“使命”的映射符号是 r12-0-1,“奉使命办事的人”的映射符号自然是 p12-0-1,但现在却映射成“外事活动”(这是义项 3 的意义,读者习惯否?)。为什么?这是为了“使馆,使节,使者,出使,大使,公使,密使,信使”诸词都能共用此义项。“人”的含义在词义库的类别符号中加 p 就解决了。实际上,词典和义项 3 给出的意义都是“使”字所具有的,不过,词典的定义没有考虑到“出使”的意义,而义项 3 则不能包含“学使”的意义,所以,两义项都应该有。我们将词典的定义加上“善良,美丽的少年”一起放在义项 4 中,这样,“学使”和“天使”两词都可“得其所哉”了。

在汉字的所有的虚词中,可以说“于”字最特殊。第一,它是万能指示符,不仅四大主角都能指示,某些配角,如条件、原因、比较等也能指示。第二,它把逻辑指示功能与类别(词性)变换功能结合在一起,而且以后者为主。变换功能主要在两方面。一是将不及物的过程及状态动词变成及物动词,但“及”的是内容,而不是对象。二是将 u 型概念变成 uv 型概念,由于 uv 型概念必然紧跟动词,所以这时的“于”起着特征要素指示符的作用。

万能性决定了“于”字作为逻辑指示符号一定是广义型,即不偏重于某一句类的角色指示,这是“广则不专”的常规道理。但“于”字作为逻辑对象指示符,却有它的侧重点,这就是效应对象。如上所述,作用对象已有“把”字承担,所以,“于”字虽然万能,但不应该插足作用对象,事实正是如此。依据我在〈问答 32〉中阐述的汉语重视作用效应两极的观点,汉语应该安排一个着重指示效应对象的虚词,我觉得它就是“于”字。不过,在字义库中,还是将它记为广义对象。但大家应该记住,它不包括作用对象,而且以效应对象为侧重点。这个意思不难用层次网络符号加以注明,但目前我宁可模糊一点。

我对于“主谓宾补”概念的发展就是将它们变成句类的函数来处理。这一思想又进一步扩展到对虚词的处理。就介词来说,我是把它当做一个二维函数来表述。第一维是四大主角及七大配角,其函数关系用本体层表示。第二维是七大句类,其函数关系用挂靠层表示。

语言的非科学或艺术性表现把这一函数关系搞得十分模糊,如同信号中混有噪声一样,但这个函数的踪迹仍然是可以察觉的。以“对象”这一主角为例,上面我们谈到了“把”主管作用对象,“于”主管效应对象。以后会谈到,“给,向,到”主管转移对象,“向,同,和,与”主管关系对象。而主管作用对象的还有“给,对”(这个“给”比较特殊,属于搭配型指示符,其映射符号是 l2-2-0-0),主管效应对象的还有“对,对于”等。

“于”作为逻辑指示符,还有另一特征,就是冗余性。从理论上说,只有当句子语义块的排序偏离标准格式时,才有必要引入角色指示符。但“于”多数情况是在标准格式下进行指示,故曰冗余。汉语为什么搞这个冗余?我还没有想清楚,一是感觉与汉语习惯于将作用对

象和效应对象或将对象与内容并为一个语义块有关;二是感觉与不容许在特征语义块与对象内容语义块之间插入辅语义块有关。这些只是感觉而已,不作定论。

词典对"于"的解释约 160 余字,但远未详尽其意。以上所说,与词典的角度不同,可以互补。望读者两相参照,阅读此文。语言本身是知识的海洋,而一个"于"字就有"浩如烟海"的丰富内涵。虽已"长言大论",但未尽之处仍多,故在下面的每一义项后面,把我以前在字义表中写下的句例转录下来,并作一点评论,期于有所补充。

1. l0-2,v7-8|(v3,v2)

例:有利于,有助于,有益于,良药苦口利于病。无济于事。忠于人民。

评注:冗余,跟效应对象。或作用、效应对象之并,这时,两者之间要加"的"字。

例:形势于我们有利。

评注:非冗余,跟效应对象。

例:嫁祸于人,问道于盲,老一代让位于新一代。

评注:冗余,跟转移对象。

2.(h↓v,l0-3),v7-8|(v1,v5,v7-1)

例:安于现状。勤于思考。乐于助人。精于牌艺。濒于绝境。处于困境。

同归于尽,青出于蓝,言归于好,源于极左思潮。马克思生于 1818 年。

评注:此义项的主项先写 h↓v,后写 l0-3,表示以词性变换为主。变换符号并没有按照前面说明的意思明确标明将不及物变为及物或将 u 概念变为及物动词,但暗含了这个意思。我觉得这样表示更准确,因为它可包括恒等变换,而不必为这一情况另立义项。实际上,所举的例子,只有前四个具有词性变换功能,后面的例子都是虚变换。此义项的说明项只表示优先与过程、状态及 7-1 概念搭配,并不排斥其他概念。这是概念层次网络理论的基本做法,读者应该习惯这一点了。

3.(hu↓v,l0-2),v7-8|jl0

例:重于泰山,合于实际,大于,小于,多于,少于,轻于,重于,快于,慢于,优于,劣于。

评注:对此义项无需解释。但应指出一点,这里列举的词,大多数未收入词典,也未收入硕士卡的词库。列举它们的目的是希望给读者一个印象,即"于"是构造新词的活跃字之一,其活性程度可与"到,出"并列。"于"字新词的取义主要是本义项和义项 5。

4. hu↓vu

例:敢于斗争。善于外交。易于处理。难于回答。

评注:这里把"敢于,善于"等,当做 vu 型概念来处理,与前面的说明不完全一致。uv 是指专门修饰动词的副词,而 vu 则是指具有 v,u 双重特性的词。汉语这一类的词,比较发达。把它翻译成英语时,往往要用"it is (表语) to…"的句型来表达。词性的原有分类标准把词性看作是语言的常量属性,而实际上它是一个变量属性,汉语尤其如此。所以,我改用五元组符号组合的方式予以全方位的表达。通过这种组合可以比较自然地给出十分丰富的类别信息。

5. l0-3-1-2

例:波于奔命。死于非命。行成于思。

评注:此义项在古汉语和现代汉语中都很活跃,对它加 lv6-0-1-2-5 的含义也许更为恰当。似乎"于"字本身暗含着"比较"和"因,果"的意思。这只是姑妄言之。

"于"还有 l1-5 等其他义项,这里就不一一说明了。当需要使用 l1-5 意义时,汉语口语习惯用"在"(英语用 in),书面语才较多用"于"。如果你在书面语中分离出单个的"于"字,而其前后又无其他单字,则绝大多数情况"于"应取此义项。但你不必把这项知识变成一项软件运行的规则,如果是这样,概念层次网络理论的优越性就不大了。实际上,这仅仅是广义同行优先准则的简单应用。如果分离的"于"后跟 j1,j2(含 wj2,pj2,pwj2),则"于"一定取 l1-5 如果前跟 v1-2-1,则"于"一定取 l0-3-1-2。这只不过是一点小花样,把 l 层次符号中的挂靠层取出来与其前后的词分层匹配就是了。难道不就是这么简单么?所以,碰上"源于"、"缘于"这样的新词,我们是能够应付裕如的。当然,如果一个文化素养较高的人,来上一句像"行成于思"这样的精辟语言,那就不是这么简单了。

"于"字谈得太长了。最后简单说一下三个由"于"字构成的极高频词:"由于,关于,对于"。它们都用直接方式表示,映射符号分别是:l1-6,l1-0,l0-2,简明而又足够精确。"关于"与"对于"的差别,词典里大约用了 300 字的说明(见"关于"词条),读者不妨参考,这有助于加深对层次网络符号体系的领会。

"相"和"互"请参看〈日记:1994.12.22.〉,不过,该日记给出的映射符号有重大遗漏:"相"和"互"的类别性不是 u 而是 uv,现在字义库中已改正过来了。uv 这种表达才能告诉你一个明确的信息,就是它的后面应跟着特征要素(请与"于"对照)。如果仅用语法的副词概念,你就不能明确地获得这一信息。

1995.1.8.星期日

译"了着曾(ceng)已正:江湖河海山"。

"了着曾已正"是一组关于时态的概念。这里又分两组,一组命名为逻辑简化,另一组命名为逻辑替代,分别映射为 l4-0,l6-0。逻辑简化的语种个性最强,这一点实际上成了(在设计之初,仅有这个念头)划分 l4 与 l6 的唯一标准。凡汉语有词而英语不以词表达(以形态变化来表达)的逻辑概念一律纳入 l4-0。反之,英汉都用词表达的纳入 l6。根据这一准则,"了着正"纳入 l4,"曾已"纳入 l6。当然,这个标准并不是那么容易掌握准,我在不同时间填写的字义表就有差异。

现字义库中这五个字的逻辑意义映射符号如下:

了	hv3-0-10
着	hv1-0-0-8
正	qv1-0-0-8
曾	ljuu6-0-1-1-1,-v4-0|j1-1-0

已　luu6-0-3-0-10

过　hvj1-1-1,-v4-0|j1-1-0

将　ljuu6-0-1-1-2

这里需要对‘曾’的映射符号加以说明。当初设计逻辑符号的时候,分别选用了直接方式和间接方式向基元概念和基本概念挂靠,并以为这样最灵活。但实际上两种挂靠同时需要的情况极少出现。纯基本概念的间接挂靠也要白白浪费一个逗号和一对括号的空间,所以后来决定统一用直接挂靠方式,只是在向基本概念挂靠时,l 之后加 j 以示区别。

示例中同时给出了‘过,将’的相应逻辑意义,以资比较。

1995.1.12

译‘再又也’。

这三个汉字是 l10 的汉语代表,我命名为带逻辑意义的副词。副词这个概念十分宽泛,国内语言界曾为某些汉字或词是不是副词或在句中是否当副词处理有过争论,可见仅囿于副词概念,人都不能达到无模糊的理解。我对副词的处理是将它们分为挂靠和非挂靠两类。前者集中到 l10;后者散布在基元概念和基本概念中,用四种类别符号 uv,uu,vu,u 来表示。uv 只修饰动词,uu 可修饰动词、形容词或副词,vu 兼有动词特性,可充当(若独立性为 0,则优先充当)C 语义块的核心成分,u 两可,可副可形。这样表示既符合概念的五元组特性,又符合语言的灵活性和弹性。上述争论也就自然不存在了。

在非挂靠类中有两个特殊的节点,就是 j6 和 jl1-m (m=2,3)。一般概念节点的五元分布是比较均匀的,但这两个节点却 delta 集中于 uu,所以 ju6=juu6,这是顺应天然的约定。如果程序不喜欢这种约定,可以去掉。不过,在我已是积习难返,书写时就难免顺其自然了。

汉语原来没有为副词搞偏旁或‘语素字’,现代汉语的‘地’是引进学派的创新。这个引进有点意思。“地”相应于英语的后缀‘ly’。为什么说有点意思?因为这个后缀似乎主要用于基元概念的副词,我一直想把这件事弄清楚,但始终没有抽出时间。

回到挂靠类副词,因为已定义 l10 为副词,所以 l10=luu10。字义库就是按这一自然简化方式装建的。其本体第二层尚未定义,一律用 0 暂代。

“再又”的共性是‘重复’,所以它们都要挂靠 1-0-0-10。现将两者的前两个义项转录如下:

再　1.(l10-0-1-0-0-10,j0-0-12-0-2)

　　2.(lvu10-0-1-0-0-10,j1-1-2)

又　1.(l10-0-1-0-0-10,3-0-10,j4-1-12-3-3)

　　2.(l10-0,(jlvr0-0-1/j1)/jlv1-1-5),g7-8-2-1

我们看到,“再”、“又”都是复合挂靠概念,“再”与第二次或将来相联系,而‘又’与多次或‘同时存在’相联系。“又”的第一义项对基元概念双重挂靠,既挂靠过程的重复性,又挂靠效应的完成性。“又”的第二义项还附加了注解项:自搭配。这里把词典中搜集的例子也转录如

下:又红又专。又白又胖。又哭又闹。又打又拉。又聪明,又漂亮。又便宜,又好。”

“再”与“在”同声(即同拼音同调),所以,下面的两句话里的 zai 有音字转换模糊,但一映射到层次网络符号空间,就变成一个简单的同行优先匹配的问题了。

张三正 zai 上大学

张三将 zai 上大学

“正”、“将”字义已见上文,读者可试作练习否?

“也”与 j10-0 挂靠,书写格式仍采用最早的约定,示例如下:

也 (l0-0,jlvr0-0-1)

意思就是:“也”是“同”的逻辑副词。

1995.1.23

译“的得 de5”。

“的得”是现代汉语里两个非常特殊的词。

字义库中目前装了“的”的六个义项,其中只有第一义项与英语的“of”相当,其他都是汉语的特殊用法。层次网络符号的表示也比较特殊,这里稍作解释。先给出各义项的映射符号:

1.(hq,/)

2.(hq,/,v7-9-2|h)

3.(g7-4-1/jgw10-3 * 0-12-6-4)/h,v7-11-1|j11-1

4.(lg0-0/hq)|lg0-2

5.(v,lg1-0)/hq,v7-11-1|(j1;j2;lg0-1)

6.((v,lg1-0)/hq)|(lg0-1/lg9-0),jlv1-1-2|vg9-4-0,
(jlvu1-2,v7-9-2)|lg1-0)

义项 1 和 2 相当于给出 hq 的两个特定定义。第一个,定义 hq 等价于偏正结构符号,第二个也含此意,但省略了“正”的部分。hq 永远作为一个符号看待,这是约定。在这两个定义里没有说明对什么插入。而在后面的三个义项里,则对这一点作了说明。第四义项指明它是对特征要素的插入,第五、第六义项指明它是对动宾结构的插入。第四、第六义项还对插入的内容作了说明。第五没有这个说明,表示它是直接插入,不加其他的词。读者可能对这些说明感到枯燥并难以理解,下面举一些例子以助说明。

义项 4:开他的玩笑。找大家的麻烦。搞硕士汉卡的开发。

这里,“开玩笑,找麻烦,搞开发”本是不可分割的特征要素,可是汉语却用这种插入把一个整体强行分开。这是汉语常用的语法手段之一。在进行特征语义块分解时要利用这一重要信息。西语比较少用这种手段,但也不是绝对不用,例如英语的 get 是高层动词,需要低层概念的补充,而有时这个补充的动词就会与 get 分开。

义项 5:他昨天进的城。我在车站买的票。是我写的稿子。是他打的人。

这里“进城,买票,写稿子,打人”都是动宾结构,汉语在其中插入“的”,起修辞的强调作用。例句实际上是对下列问话——“他哪天进的城?”、“你在哪儿买的票?”、“谁写的稿子?”、“谁动手打了人?”的回答。可以简单的说:“昨天”、“在车站”、“我”、“他”。后面两句话也可以说:“稿子是我写的”、“人是他打的”。总之,“的”字在这里起着重要的表意作用。

义项 6:你干你的事。你干你的。你走你的阳关道,我过我的独木桥。

这里插入的内容一定有“主语的代词”。此义项有两项附注,一是否定参与,二是“宾”成分可以省略。似乎这个表示把有关例句的意思表达得很清楚,其实情况并非如此。例句中的“的”都紧跟在插入代词的后面,但实际的语言可以打破这个规范,把代词后面的“的”省掉。例如说“让孩子们干他们自己喜欢干的事”,这里“他们自己”后面的“的”就省去了。如何克服这个困难?请大家思考。

“的”字还有其他用法,但目前字义库只装了上述六条。其映射符号中包含一些自然的约定,已如上述。“得”字也有类似情况。

“得”字目前装了四个义项:

1.(hv,l0-3-3-0;(l0-2-0-0,))

2.(j7-1-2,/)|j6

3.(h,(jl1-2-1;jl1-1))

4.(hv,jl1-2-1)

义项 1 和 4 都表示“得”是动词的后缀,但作用却完全不同。义项 1 又含两个义项,一是仅作效应内容的指示符,二是作作用对象和效应内容的双重指示符,构成作用效应句。义项 1 有浓厚的汉语特色。为了说明这一点,应当回顾一下关于作用表达与效应表达的两极观(见〈问答 32〉)。对于作用表达,作用者是不可缺少的,自然的语义块顺序是作用者、作用、作用对象。在西语,符合这个自然顺序的句子叫主动式,否则叫被动式。主动式的主语是作用者,被动式的主语是作用对象。对于效应表达,作用者是隐含的,不是必不可少的,必不可少的是效应对象,于是,效应对象成了主语“常委”。从语法的“主谓宾”结构来说,这时“宾”已无存在之必要。汉语针对这一情况,将“效应对象 + 效应”视为效应句的标准格式之一,西语则统一按被动式处理。如果一个句子只有两个语义块,或在逻辑上容许使用被动式,中西语言的这种差别不难处理。但是,如果一个句子有三个或四个语义块,或被动式在逻辑上不容许使用,则对中西语言结构差别的处理就比较费事。“得”字的作用就不是与西语的一个词对应,而是与一种结构对应。例如下面的句子:“这件工作做得很出色”,“这件事办得大家都很满意”,“他的这一番话说得大家心服口服”。

1995.1.24.星期二

译“这那;从到”。

至此,逻辑概念常用汉字的第一期建库告一段落。

除 18 之外,其他逻辑概念,汉语以单字词为主。

按词频统计,前 10 个高频词累计占有率为 17.7,前 20 个为 24.5,前 50 个为 34.7,前 100 个为 42.9,前 200 个为 52.1。在前 10 个高频词中,逻辑概念占了 7 个。这个统计是混合统计。假定单字词与非单字词的比例各占一半,则将混合统计结果扣除双字词的累计占有率以后乘以 2,即得单字词本身的累计占有率。前 10 个和前 20 个里都没有双字词,可用上面的数据直接换算,相应的累计占有率分别是 1/3 和一半。前 50 个、100 个、200 个里双字词的累计占有率分别为:1.02224%,2.402769%,4.705260%。经粗略换算可得单字词的累计占有率分别为:68%,84%,97%。这个结果与我们曾独立作过的单字词词频统计相当接近。可见单字词占一半的假定大体不差。但这个比例与语体或文体的关系很大,就现代汉语的文件体来说,单字词的比例显著小于二分之一;但对小说来说,则大于二分之一。但大家不妨记住单字词累计占有率三分之一、一半、三分之二的相应常用汉字数:10,20,50。这个数字好记,并有知识性。

这里应该说明一点。词频统计数据中的高频字分别给出了不同词性的统计结果,例如"在"字分别给出了它作为介词、动词和副词的词频,所以,在扣除重复字及双字词以后,进入前 200 名的汉字只有 153 个,补充"该及即极于与以已之"9 个逻辑字,总计 162 个。这些就是字义库第一期工程的主要目标了。

这里还应该特别指出下列极易构成新词的字:"出到去来起住得上下成开"等。这些字用在动词的前后,产生附加的意义。例如"出"在动词后表示"趋向或效果"的意思,"到"在动词后表示"效应完成"的意思,"去"在动词前表示"去做或要做"的意思,在动词后表示"趋向持续"的意思。大家应注意到,这些意思都可以用基元概念予以表示,其层次网络符号就是 hvi-m-k 或 qvi-m-k。这类组合词不胜搜集。以"到"字为例,在 5000 个常用词里收集的"看到,做到",《现代汉语词典》就没有收集。至于"见到,听到,送到,传到,走到,赶到,涨到,跌到"等等更不可能一一收集。此外,它还可以与双字词搭配形成词组,例如"了解到,关系到,涉及到,推广到,深入到"等等,这里的"到",除了"了解到"之外,其他并没有"效应完成"的意思,而只是一个冗余性逻辑指示符号 11-0。"到"字的这种多义模糊,层次网络符号也不能给出明确的解模糊知识,需要联系上下文才能作出可靠的判断。

1995.2.16

上午与季宏谈"机器翻译"基金申请事,记其要点如下:

首先要明确翻译过程的不变性、稳定性、可变性、唯一性及不确定性。

不变的是语义块内涵及个数。也就是说,在翻译过程中要保持语义块内涵及个数的不变性,因为这两者与语种无关。

稳定的是句类。句类基本上与语种无关,但不同语种对强相关句类各有偏好,例如对于作用句和效应句,汉语偏好效应句,而西语偏好被动式作用句;对于基本判断句和基本状态句,汉语偏好基本状态句,而西语偏好基本判断句。但是,作用效应句、反应句、过程句、转移句、关系句具有很好的稳定性,即源语句和目标语句的句类相同,不需要转换。这就是句类

的稳定性表现。

可变的是语义块的排列次序和语义块的构成形式。这些与语种的语法、语用习惯强相关。这就是说,语义块排列顺序和语义块的构成方式具有强烈的可变性。

唯一性及不确定性,在这里是指从层次网络符号到自然语言反映射的两种表现,也可称为单值性及多值性。一般说来,这个反映射都是多值的,但这个多值性主要来于语法、语用习惯或艺术性的需要,而不是语义或科学性的需要。以概念节点"肯定性基本判断"jl1-1-1为例,汉语的映射符号有"是,为,乃,即"等,而英语有"be,is,are,am"。以"结束"v1-1-2这个概念为例,汉语有"结束,完,毕,终止",英语有"end,terminate,finish,conclude,close,stop"等。如果不考虑语用习惯及艺术性要求,就可以人为地将一些反映射单值化,例如将jl1-1-1单值反映射为"是",将v1-1-2单值反映射为"结束"。至于西语里源于人称、性、数、格的多值,实际上可认为是单值。这里的"人为"是"简化和近似"的意思。"为"作为"是"的近似,兼有"等于"的意思;"乃,即"作为"是"的近似,等价于用"是"替代"就是",都是比较精确的近似。一般说来,反映射的多值性问题比较容易处理,关键是正映射的多值性,必须把它转化为单值映射,这就是词的解多义模糊问题。例如汉语的"是",还有存在及其他逻辑意义,在"前面是一片稻田"、"是集体的事大家都要关心"、"这场雨下得是时候"、"你是坐火车,还是坐汽车?"这些句子里的"是"就不能映射为jl1-1-1,而应该相应地映射为jl1-1-5,(l9-0-3-0-0,jl1-1-1),ljv6-0-6-0,g7-4-2-。

所谓正映射过程是将自然语言映射成层次网络符号,所谓反映射过程是将层次网络符号映射成自然语言。翻译的分析过程是正映射过程,语言生成过程是反映射过程。

分析或正映射过程的具体内容是:

对源语句确定语义块的个数、分出主语义块和辅语义块、确定每个语义块的角色及语义块中的核心成分和说明成分,与此同时,以确定句类为主线,消除词的多义性模糊、语义块切分及组合时可能的歧义模糊,就汉语来说,还包括分词模糊。这一分析过程在一般情况只是实现了对源语句的基本理解,因为上述各种模糊可能未消除干净,而且,即使模糊得以全部消除,某些隐知识还有待揭示,否则仍不能得到高质量的译文。因此,上述分析只是局部分析。在每一次局部分析之后,还要作全局分析,这包括语境生成、要点主题分析及隐知识揭示三项。局部分析是指定性的,全局分析是"灵感"式的,即可深可浅,可多可少甚至可有可无。每一次全局分析之后是否返回去再做一次局部分析也是"灵感"性的,即可做可不做,取决于该次全局分析是否增加了新的信息以及前此的局部分析是否留有疑点。

生成或反映射过程的具体内容是:

根据分析过程得到上列四项信息,即:句类、主辅语义块个数、每个语义块的角色、其核心成分及说明成分。依次确定:句类的具体格式,包括语义块的排列顺序;语义块的构成方式;具体词汇的选定。这个生成三步曲的每一步都存在唯一性及不确定性问题。

一个简化翻译系统的基本思路应该是:将客观的不确定性强行变成主观的确定性,从而把双向翻译的技术性困难降低到最低限度。但利用这一简化翻译系统可进行源语句与双向

翻译后的目标语的语义一致性检验。用这种方式得到的一致率可作为译准率的等价指数。

我们的目标是打破译准率徘徊于 70% 80%的困境。以我们目前已经拥有并日益扩大的理论及语言知识库武器来说,这一点肯定是不难达到的。

然而,想到创新性课题的基金申请之‘难’而本课题处于‘一天吃一斤草,产两斤奶’的困境,不禁感慨万千,再次‘闭关’之愿索然。

人的行为,其内在因素由基元概念的理智、观念、心理、素质和习惯,基本概念的 j8 所决定。理智主导者为强人,情感主导者为痴人,习惯主导者为庸人。多数人介乎三者之间,我则一典型痴人耳。愿有缘读到此文者以做‘事业中的强人,生活中的痴人’自勉。

后　记

在创立 HNC 理论的过程中，曾五次进行过阶段性总结。

第一次是 1992 年下半年，围绕着基元概念的 13 个一级节点、基本概念和语言逻辑，共写了 15 篇专文，另外有一些关于知识库的短文和传统语言学的短评。这批稿子都是手写的。当时很穷，个人计算机很少，虽然我们拥有自己发明的使用极为方便的汉字输入方法——硕士卡，可是我个人还不具备享受这一成果的条件。这批手稿都已荡然无存，打印稿也残缺不全。这是 HNC 理论青年期的文字，锐气有余，深度不够，无存损失不大。

第二次是 1993 年冬到 1994 年夏，是预定“闭关”十年的暂休时间，写了〈HNC 理解处理问答〉，试图系统阐述 HNC 对自然语言理解处理的总体思路和方案。大部分仍然是手稿，小部分是我自己直接在计算机上写的。共有几问几答，已没有确切的数字了。这是 HNC 理论进入成熟期的文字，有点保存价值。感谢张全博士，他保存了比较完整的打印稿。这次把这批稿子以原貌整理成文并编入本书，其中必然有应该淘汰甚至错误的东西，但一定要保存这些反面的印迹，因为它们对于 HNC 未来探索的参考价值，或许超过正面印迹。

第三次是 1994 年冬到 1995 年春末，写了〈语义学日记〉，试图通过对概念节点及其反映射汉字的阐述，剖析各局部联想脉络的内部结构和外部连接，并希望通过这一写作方式再次进入“闭关”状态。原文都是在荧屏前写的，没有手稿，也没有全部打印。当机内原始文件毁于一次机器事故时，因当时心情很坏，也没有及时采取补救措施，例如搜集和保存组内的拷贝文件及已打印稿件。这一损失是不堪回首和难以弥补的。这次仅找到残存的一小部分内容以〈语义学日记选录〉为题编入本书。

以上三次都只是为内部需要而写，为了让我当时的研究生和助手了解 HNC 的来龙去脉，为他们牵线搭桥，期望他们按照 HNC 的思路去勇敢地探索自然语言理解处理的新路。

第四次是 1995 年，由于马雄鸣先生的鼓励，开始产生走向社会的意识。拟定了一个〈HNC 论文选集〉的写作计划，预定 21 篇，其中部分论文此前已有初稿或写就，目录如下：

＊1　自然语言语义网络的基本构成及其特性
＊2　自然语言的深层结构及句类分析
＊3　HNC 理解处理系统的基本框架
{4}　解模糊及纠错处理
5　关于汉语词库结构及汉语文本表示的建议
＊6　概念知识和语言知识
＊7　关于汉语 HNC 知识库的建设

{8} 汉语音节感知库及字义库

9 汉语的层选处理

[10] 汉语的新词辨识

* 11 语义块的切分组合处理

[12] 理解处理的环境仿真

[13] 双向及多语种互译问题初探

* 14 作用、效应句的句类知识

* 15 作用反应句及作用承受句的句类知识

16 过程句的句类知识

* 17 转移句的句类知识

18 关系句的句类知识

19 状态句和基本判断句的句类知识

{20} 一般判断句的句类知识

* 21 混合句的句类知识

但这个计划没有全部完成。其中,带[ ]号的3篇仅有提纲,带 {} 号的3篇仅有初稿,实际完成的只有 15 篇。这批论文当时都以 Paper 命名,以区别于过去的〈HNC 理解处理问答〉及其他。对这批稿件,刘志文先生承担了文稿的校订、编辑、打印,最后形成〈HNC 理解处理论文选录〉的繁重工作;杜燕玲女士主持了【5】的写作,并参加了【19】和【21】两文的起草工作。本书选用了带‘ * ’的 10 篇。【5】放在附录中。

随着鼠年‘九五’的来临,HNC 开始时来运转。先是国家语委主任许嘉璐教授对汉字拼音智能输入项目的安排,并表示对 HNC 理论寄以厚望,这不仅使当时面临解体之灾的 HNC 小组得以生存下去,而且给这项研究送来了极大的精神鼓励。接着是中科院高技术局主持了对 HNC 理论(以〈HNC 理解处理论文选录〉为依据)的专家评估会议,并继而把 HNC 列入国家计委‘九五’攻关项目的申请专题,使 HNC 在中文信息处理领域开始有了一席之地。HNC 的漫漫长夜终于出现了希望之光。

〈HNC 理解处理的 52 个论题〉(简称〈论题〉系列)是 HNC 的第五次阶段总结,预定三个月写完。中心目标是阐述 HNC 技术实现的策略,兼及 HNC 思路的形成过程。〈论题〉系列分为 8 组,第一组从论题 1 5,讨论 E 语义块感知;第二组从论题 6 12,讨论广义对象语义块、辅块和短语的感知;第三组从论题 13 17,试图从知性的高度阐述语句表示式的来龙去脉,以期有助于提高 HNC 攻关组,主要是句类分析设计者的理论水平;第四组从 18 20,讨论句类转换;第五组从 21 24,讨论汉语特有的音节感知处理;第六组从 25 34,是整个〈论题〉的核心,讨论句类假设检验及语义块构成处理的基本策略,包括知识运用的基本策略;第七组从 35 39,讨论语义距离计算的有关问题;第八组从 40 51,试图用小散文的形式,而不是论文的形式阐述 HNC 的一些重要概念,以期有助于提高 HNC 联合攻关组,主要是知识库建设者的理论水平。

〈论题〉写作之初定下了两条原则,一是急用先写,二是抛砖引玉。这两点的直接对象都是 HNC 联合攻关组,而不是一般读者。“急用”之作主要是服务于句类分析技术的提高和完善。“引玉”之作的本意则是出题目,活跃联合攻关组成员的思路,激励他们的写作热情。

由于身体状况欠佳,〈HNC 理解处理的 52 个论题〉的写作计划骤然终止,目前仍力不从心。略感宽慰的是,“急用”部分所缺甚少,未写或仅有提纲的论题只能有待于他日。然而,更期望由战友们来完成,题目仅仅是题目而已,不是专利,我的这一期望绝对是真诚的。

HNC 联合攻关组是 1997 年 3 月 3 日成立的。这是一个值得纪念的日子,它标志着 HNC 从小作坊时期进入了符合现代要求的发展时期。陈力为院士、许嘉璐教授、林杏光教授、张普教授和中科院桂文庄局长为促成这一转变起了决定性作用。特别是林杏光教授以其全部学术精力投入了 HNC 联合攻关组的工作。他作为一个跨接语言学和计算机科学的语言学家,以其广阔的视野和敏锐的目光,从 21 世纪语言信息科学和语言信息产业发展大势的高度出发,为联合攻关组规划了远景发展目标,制定了近期工作纲要,采取了一系列重大举措,本书的出版就是其中之一。

本书反映了 HNC 预定建立的自然语言五层面理论模式(见弁言)中前两个模式的研究成果。这一成果为后续三个理论模式的探索奠定了基础。这两个理论模式的基本结论可概括成一句话:自然语言无限的语句可以用有限的语句物理表示式来表达。

这个结论是 HNC 第一理论模式的推论,是在 1992 年 1993 年之交确定的。对这一推论,一方面我深信康德的“自然的最高立法……在我们的知性之中”的名言,另一方面,又深知不能违背“实践是检验真理的唯一标准”的科学论断,自然语言更应如此。然而,大规模进行这一实践检验所需要的技术条件,在当时看来是极为渺茫的,只能采取人工方式。于是,决定从“闭关”状态中休整一下,进行这一验证工作。这项工作断断续续持续了三年之久。这里应该特别提到杜燕玲女士,她不辞繁琐,为此付出了艰辛的努力。但我们从她收集的语句中未曾发现一个不能用句类物理表示式进行分析的句子,这使我们感到欣慰。

一个普通而天然的疑问必然在人们的心中徘徊:自然语言无限的语句能用有限的物理表示式加以表述么?对有限语句未发现“例外”,就能肯定上述推论么?这违反逻辑论证的基本原则。

当然,在严格的意义上,对上面的推论,应该像对许多著名的数学猜想那样作严格的证明,这确实是必要的。但这不是自然语言理解处理当前的急需。重要的是,HNC 开拓了一条模拟大脑语言感知过程的新路。这条新路的路基就是有限的基本句类和混合句类的物理表示式。同时也应该清醒地看到,目前终究还仅仅是一个路基,在这个基础之上的上层建筑是一个浩大的系统工程。《HNC 概念符号体系手册》和《HNC 句类知识手册》的编定,HNC 句类分析技术的完善,HNC 六类知识库对汉语和其他大语种的全方位建设,都还需要付出极大的努力。至于 HNC 预定的关于句群与篇章、记忆与学习理论模式的探索和建立,HNC 九大处理模块尚未着手部分的设计和实现,仍然需要知性的弘扬和创造的艰辛。

前几天，张全博士请我为 HNC 产品的窗口画面写一首七言律诗，我很支持这个想法，于是欣然命笔，现转录如下，作为本书后记的结束语：

科技朝霞起大西，东方长恸失先机。
汉语神奇寓真谛，当仁不让领红旗。

黄曾阳
1998 年 8 月 30 日

主要参考文献

Bookman L A. 1994. Trajectories through Knowledge Space. Boston: Kluwer Academic Publ

陈群秀,张普. 1995. 信息处理用现代汉语语义分类体系:属性分类. 见:陈力为,袁琦主编. 中文信息处理应用平台工程. 北京:电子工业出版社,206 214

Chomsky N. 1965. Aspects of the Theory of Syntax. Cambridge, MA: MIT Press

Chomsky N. 1975. The Logical Structure of Linguistic Theory. New York: Plenum Press

Chomsky N. 1981. Lectures on Government and Binding. Dordrecht: Foris

Chomsky N. 1986. Knowledge of Language: Its Nature, Origins and Use. New York: Praeger

Church K W. 1990. Word association norms, mutual information and lexicography. Computational Linguistics, 16(1)

冯志伟. 1983. 汉语语句的多义多标记树形图分析法. 人工智能学报,(2)

Fillmore C J. 1968. The case for case. In: Bach E and Harms R eds. Universal in Linguistic Theory. New York: Holt, Rinehart and Winston

高名凯. 1957. 汉语语法论. 北京:科学出版社

Halliday M. 1985. An Introduction to Functional Grammar. Edward Armold Publ. Ltd

黑格尔. 1980. 小逻辑. 贺麟译,北京:商务印书馆

黄侃. 1983. 文字声韵训诂笔记. 上海:上海古籍出版社

黄焯. 1981. 诗义重章互足说. 见:诗说,卷 2. 武汉:长江文艺出版社. 33 36

康德. 1957. 纯粹理性批判. 蓝公武译,北京:三联书店

Lenat D B. 1995. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11)

李耳. 老子

林杏光. 1990. 汉语多用词典. 北京:中国标准出版社

鲁川. 1995. 现代汉语的语义网络. 见:陈力为,袁琦主编. 中文信息处理应用平台工程. 北京:电子工业出版社, 233 252

陆俭明. 1989. 十年来现代汉语语法研究的理论与方法管见. 国外语言学,(2)

罗素. 1992. 西方的智慧. 马家驹等译,北京:世界知识出版社

马希文等. 1990. 自然语言处理与自动文摘. 见:高庆狮主编. 智能技术与系统基础. 北京:北京大学出版社, 99 117

毛泽东. 1969. 中国革命战争的战略问题. 毛泽东选集全卷本. 北京:人民出版社, 154 225

Minsky M. 1963. Steps toward Artificial Intelligence. In: Feigenbaum E A et al eds. Computers and Thought. New York: McGraw-Hill

Minsky M. 1975. A framework for representing knowledge. In: Winston P H ed. The Psychology of Computer Vision. New York: McGraw-Hill

Schank R C. 1973. Identification of conceptualizations underlying natural language. In: Schank R C, Colby K eds. Computer Models of Thought and Language. San Francisco, CA: W.H. Freeman and Company

Schank R C. 1975. Conceptual Information Processing. Amsterdam: North Holland

Schank R C. 1975. The structure of episodes in memory. In: Bobrow D, Collins A eds. Representation and Understanding. New York: Academic Press

Schank R C et al. 1977. Scripts, Plans, Goals and Understanding. Hillsdale, NJ: Lawrence- Eribaum Assoc.

Schank R C. 1982. Dynamic Memory. New York: Cambridge University Press

Simon H A ed. 1972. Representation and Meaning: Experiments with Information Processing System. Englewood Cliffs, NJ: Prentice-Hall

孙武. **孙子兵法**

索绪尔. 1990. **普通语言学教程**. **北京**: **商务印书馆**

Veronis J et al. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In: Proc. COLING '90, 289 295

王国维. 1995. **人间词话**. **北京**: **群言出版社**

维纳. 1963. **控制论**. **郝季仁译**. **北京**: **科学出版社**

White G M. 1990. Natural Language Understanding and Speech Recognition. Communications of the ACM, 33 (8): 72 82

Winograd T. 1983. Language as a Cognitive Process. Reading, MA: Addison-Wesley Publ. Co.

俞士汶等. 1995. **关于现代汉语词语的语法功能分类**. **见**: **陈力为, 袁琦主编**. **中文信息处理应用平台工程**. **北京**: **电子工业出版社**, 157 164

张志公. 1990. **汉语语法再研究**. **外语教学与研究**, (3)

张普. 1991. **信息处理用现代汉语语义分析的理论与方法**. **中文信息学报**, (3)

庄周. **庄子**

附　　录

致许嘉璐先生的信

小盟学弟并转呈许先生：

汉语拼音智能音字转换系统(以下简称本项目)的研制已取得了关键性的技术攻关进展,它表明许先生大力支持并亲自组织实施的智能方案是可以实现的。

这一关键性技术攻关进展的具体表现是：

第一,汉语词性兼类现象和汉语无关系代词所带来的汉语句法分析的传统困难(如述语的辨识和短语构成分析等),可以通过 HNC 的语义块感知和句类辨识处理得到解决。

第二,汉语比西语更为常见的语义块分离现象和句类转换现象在句类分析的驾驭下可以进行有效的处理。

第三,汉语拼音直接面对的巨大音字转换模糊在上两项成果的基础上,通过 HNC 知识库提供的语义语用和语法知识的综合运用,可以得到有效的消解。

这三项技术进展的关键在第一项。

为取得此项技术难关的突破,我们投入了最大的力量,花费了最多的时间,但同时也影响了本项目的整体部署和进度。在取得这一进展之后,HNC 下一步向何处去,是一个重大的战略决策问题。在这个关键时刻,对 HNC 前一阶段的成果进行实事求是的总结,明确前进的方向是非常必要的。

从机器翻译开始的文字处理技术,从语音识别开始的语音处理技术,都一直回避一个根本性的问题,就是大脑的语言理解过程。西方传统的语法或句法理论一直被误认为是理解的基础,这种误解甚至传染到自然科学领域。神经生理学家威廉·卡尔文在《大脑如何思维》一书里,用‘句法——智力的基础’为章名,并说“很可能,短语结构所使用的那些规则,就是思维机制”。

思维的机制绝不是语法或句法,而是概念联想脉络的建立、激活、扩展、浓缩与存储。计算机智力不应该以图灵检验为标准,而应该以对语言的五重(对语音)或三重(对文字)模糊的消解能力为第一标准。这是 HNC 理论的出发点。

HNC 理论基于这一认识首先建立了自然语言概念体系的理论模式,在这个基础上又建立了 HNC 自然语言语义块和语句的数学和物理表示式,即句类表示式。幸运的是,HNC 穷尽了这些表示式的具体结构。这两个理论模式的建立为句群关联性表述模式、篇章要点表述模式、短期及长期记忆的生成及转换模式、知识自学习模式的探索奠定了坚实的基础。这些理论模式是让计算机获得语言习得机制并模拟大脑思维过程进行自然语言理解处理的基本依托。

但是,后续 4 项理论模式的研究必须立足于前两个理论模式的技术实现,因为开展这些研究的基本条件是计算机能对大规模真实语料进行语句的理解处理。

本项目已有的技术进展,能否为汉语大规模真实语料的理解处理提供软件保障?这一点,在 HNC 联合攻关组内部是没有异议的。因为,文字流只存在三重模糊,HNC 理解处理面临的困难都是不难克服的。但是,本项目面临的是语音流的五重模糊,理解处理的难度要大得多,因此,下一步技术攻关的重点投向哪里,在 HNC 联合攻关组内部存在不同看法。不同看法主要来于对需求和主战场的不同估计。

与两年前相比,以汉语拼音为汉字键盘输入基本手段的软件有了很大发展,有些宣称具有语句处理功能。这些软件的共同特点是运用了较多的统计搭配知识,并提供了较好的人机交互界面。它们对中文信息处理不会带来理论或技术上的促进,但能够基本满足汉字键盘输入的市场需要。

本项目以句子理解为基础的智能方案是一个彻底的一劳永逸的解决方案。但是,要形成用户满意的产品,还要做大量琐碎的技术开发工作。我们的人力十分有限,技术开发又不是我们的专长,如果过多投入技术开发方面,必然会削弱投入主战场的力量,这是不妥当和不明智的。

对于 HNC 来说,当前的主战场应该是:

尽快推进 HNC 的自然语言理解处理策略,全面取得 HNC 理论预期的处理效果。这关系到自然语言处理或中文信息处理的战略决策,关系到信息时代向高级阶段发展的技术基础,关系到知识处理产业的关键技术(比尔·盖茨先生最近在清华大学的谈话表明,他对此已有充分预感和认识)能否在中华大地诞生,关系到对大脑如何思维这一超级科学奥秘重新确定探索之路。

谈到自然语言理解,主要是未来学家和初期的人工智能学者曾作出过乐观的估计,领域内的专家一般都持悲观态度,认为在 20 100 年之内不会产生重大突破。因为,他们深知,前面没有亮光。

然而,HNC 的出现改变了这一势态。

但是,习惯于句法理论的语言学家,习惯于专家系统思维方式的人工智能专家,习惯于数理逻辑思维的数学家,习惯于数据信息处理的信号处理专家,都很难适应 HNC 的理论体系和方法,因而也就很难接受这一新的势态。因为,他们实际上都接受图灵的计算机智能标准(尽管有人可能不知道图灵其人)。但是图灵标准是对大脑的全面模拟,不区分概念、语言和常识三个层面的知识,而对当前的计算机来说,这一区分是至关重要的。HNC 强调三层面知识的区分,并以概念层面的知识为基本依托,建立计算机的自然语言智能。其基本标准和目标是:第一步,计算机能够如同常人那样感知语义块和辨识句类。第二步,在这个基础上上升到句段和篇章的理解,实现记忆与学习。

当前的进展只是证明了,第一步目标是可以实现的。但这一步的完全实现还有大量的具体工作,包括 HNC 知识库和软件两方面的工作。主战场要我们继续投入最大的力量。

但是,这决不是说,我们应该放弃预定的产品目标,相反,应该毫不动摇地坚持这一目标,这不仅是诺言和协议的问题,而且是因为在这个目标里,蕴藏着探求大脑奥秘的重大线索。

不过,在具体实施方面,我建议动用 HNC 的新同盟军——中科院软件工程中心的技术力量。产品的推出时间作相应调整,并制定出明确的版本升级方案。

HNC 理论基本框架诞生已经五年之久了,我深知它生不逢时,本来就没有打算在有生之年公布它的结果,更不曾打算付诸实践。感谢许先生和计算语言学界朋友们的支持和劝导,使 HNC 走上了正常发展的道路。但书生的愚见常使我感到无所适从,以上所说,仅供许先生参考。

HNC 联合攻关组期盼着您的指示。

黄曾阳

1998 年 4 月 28 日

给萧友芙老师的信

友芙阁下：

阁下近日所写的三篇论述，都已看过。这是 HNC 联合攻关组极为可喜的兴旺景象，我希望攻关组成员都积极投入这一写作行列，并形成热潮。

题目都很大，特别是“如何利用知识库所提供的信息”一文，在论述的范围和深度上，很难控制。题目中的“知识库”前面应加 HNC。“关于概念类别”一文理论上涉及百年来汉语语法界争论不休的“词性”问题，工程上涉及如何利用概念类别信息的问题，如果论述主要是涉及工程应用，则题目改为“关于概念类别信息的运用”为宜。“关于主辅语义块转变的问题”恐怕要从明确有关定义入手，另外我想说一句，HNC 曾为必须引入大量的新词或赋予老词以新意而煞费苦心，这里的“转变”就是“变换”，但使用时要注意它们的 v，g 语用之分，题目的转变改成变换为妥。〈HNC 理解处理的 52 个论题〉(以下简称〈52 个论题〉)的 9、10、11 属于阁下此文的范畴，那里的题目用的是“转变”，取其 v 意，而阁下此文则取“变换”的 g 意为妥。文章的命题十分重要，所以我从这一点谈起。

下面分别谈一些看法，从“关于概念类别”一文开始。“概念类别”这个提法，HNC 赋予了特定的意义，在〈HNC 理解处理论文选录〉的【6】中，在〈52 个论题〉的论题 33 中，都有专门论述。在刘志文先生为 HNC 培训班写的学习材料中有详细说明。

问题在于，原定的概念类别符号组合规则，最初受到“语义结构方程”设计的约束，后来，当这一设计思想已被放弃因而这一约束已不存在时，原有的种种烦琐约定并未及时予以彻底消除，当前首先要做的是这件事。

彻底消除的办法就是：

第一，类别符号仅分基元表示和复合表示两种，基元表示用一个字母，复合表示用多个字母。基元表示都已有明确的定义，现在需要对每一个复合表示给出明确的定义。注意，按照这种定义方式，每一项表示都是相互独立的。明确这一点以后，现稿中的第一个问题“包括 uu 和 uv 吗？”，自然就不存在。

第二，以联合方式表示多元性，类别符号之间用逗号隔开。逗号前后的类别符号单元，用基元表示、复合表示或其他表示方式均可。

这两点是确定当前 HNC 知识表示菜单(工作单)时的规定，现在看来，对概念类别多元性表现，还应该加上展开符号“+”。HNC 菜单的“词频及语境”栏目体现了无条件概率和条件概率相互配合的表达思想，可分为三档，相应的数字分别是 0　3，4　7，8　b。对概念类别的多元性表现，建议第一档用“，”表示，而第二档和第三档则分别用“+”和“++”表示。

这里对类别符号引入了“基元”和“复合基元”的概念，这两者及其他表示方式可统称类别符号单元。在概念的 HNC 表示式中只使用“基元”或“复合基元”表示，但在知识库的“概念类别”栏目中，还需要使用其他的表示方式，一种是字母串数字串的方式，另一种是小写与大写字母混合使用的方式。如果对这两种方式取一个名字，建议分别命名为第三类和第四类概念类别。

阁下的中心任务是：

1. 对原来定义的复合表示加以清理，确定哪些保留，哪些废除。

2. 对保留下来的每一项复合表示写出明确的说明。有些复合表示意义比较简明，如 zz，uv，uu，jw，gw，p-，pe 之类，但有些比较复杂，需要精确化，如 vu，vv，ug 之类。这里的提法意味着对原来定义过的复合表示有减无增，不需要增加新的内容，但实际情况是否如此，由阁下裁定。

3. 规范第三类概念类别：“字母 + 数字”，写出简要说明。第三概念类别的提出是出于基本检验的需要。基本检验有两个层次，语义块层次和句类层次，语义块层次的基本检验一定是邻近检验，句类层次的基本检验则是指对 JK 要素的检验，通常是远距离的。目前程序的基本检验能力很弱，亟待加强，我在联合研讨会的发言，把这一点放在“加强”的第二点，阁下应体会此意。狭义空间、社会空间、广义空间、城市、近代及现代产品、生命体等概念的引入都是为了加强基本检验。这些概念都采用了“字母 + 数字”的方式。

但是，应该强调指出：采用“字母 + 数字”的表示方式不仅仅是为了表达上列特别定义的概念，也是为了表达基本概念和语言逻辑概念。这两类概念对语义块感知和基本检验都具有与基元概念不同的特殊信息，很有必要在概念类别栏目给出简明表示，以满足语义块感知和基本检验的需要。这就是说，对基元概念的类别表示只需要五元组，或者说，只需要基元或复合基元两种表示方式，但对基本概念和逻辑概念就显得不够充分，需要“字母 + 数字”的方式。有人会说，数字所提供的信息不是已在概念的 HNC 符号里有充分表示么？为什么还要另搞一套？如果有人真这么想问题，那我建议他回答一个类似的问题，“为什么一本书的前面要另搞一个目录？”为什么一篇文章要分好段落并经常煞费苦心为段落起一个贴切的名字？”

上面阐述的观点并不是新东西，但过去没有把它系统化，这就是阁下当前的任务。这一层意思值得深思，我认为大脑语言感知的秘密之一就在这里。我对汉语音节知识库的设计就体现了这一层意思，实际上这就是分层次激活和存储信息的思想。

4. 规范第四类概念类别基元：vB，vC，Bv，Cv，f1，f2 等等，写出简要说明。这里的 B，C 原来用小写字母，建议改用大写字母。

如果说前三种概念类别乃用于词汇，则最后这一种概念类别则用于短语，包括汉语的多字词。

在上列四项任务中，第二项是中心。工作量比较大，已有的约定有待阁下收集、整理并作必要的修正。

现稿中关于逻辑概念的说明可以删掉,它不属于本文的范畴。

关于基本物的说明也是如此,但应指出一点,jw6m 和 jw6m- 必须分开,原来定义的 jw6m, m = 0, 1, 2, 3 仍然保留,取消 m = 4 b 的原来定义,用 jw6m- 替代。现稿把原来的 jw6m 统一纳入 jw6m- 是错误的。

下面转向‘如何’一文。

此文总体结构比较适当,这就是说,分别从语义块感知、句类假设检验和语义块构成分析三个阶段来说明知识库信息的利用是非常正确的思路。

但是,采用何种格调来写则是一个不易把握的问题。现稿的特点是格调不明朗。我也说不好阁下应该采用哪一种格调。

这里说的格调是指陈述问题的角度和方式,按照毛泽东时代的语言来说,就是所谓立场、观点和方法。就‘如何’这个题目来说,应该是从知识库的立场出发,具体阐述知识库的哪些栏目在理解处理的哪些环节提供了哪些可用的知识,程序如何利用这些知识。但现稿的阐述方式似乎不是这样,而在多种立场之间转移,这样不容易把问题说清楚。

上面两段话似乎有矛盾,后者不是对前者的否定么?其实不是,HNC 理解处理的基本策略虽然是明确的,但具体的操作过程仍有许多环节有待优化,因此阁下的这种自由阐述方式也许会产生思想的火花,这就是‘后者不否定前者’的原因。由黑格尔所深化并被马克思所大力宣扬的辩证法思想,在我们这一代人可谓耳熟能详,但实际上往往是被当做教条而不加以实践,坚持创新思路的 HNC 联合攻关组为什么不实践一下?

以上是一般讨论,下面谈一点具体意见。

概念类别信息是当前程序在知识利用方面的最大弱点,同时也是知识库与程序携手合作最有潜力的环节之一,因此‘如何’一文应该与上面的‘关于’一文互相呼应,相得益彰,在这一点上深化展开,然后推进到其他栏目。

对‘如何’一文,我暂时就写这些。下面转向另一篇‘关于’。

该文目前只给了一个提纲和一些素材,提纲未作说明,这是首先需要补充的。

在工程意义上,辅块也可以这样定义:辅块是句类代码之外的语义块。HNC 理论提出语义块有主辅之分,但从未说过语义块非主即辅,这是完全不同的两个概念。一般说来,‘有甲乙之分’的说法不排除可甲可乙的过渡或中间状态。HNC 理论并不拒绝两分法,但明确排除对两分法的滥用,因为中国人曾深受这一滥用的惨痛伤害,故 HNC 理论对此保持高度警惕,并一再呼吁 HNC 技术也要高度警惕,为取得两可处理的功能而努力。

我希望阁下此文也体现这一精神。

语言逻辑概念的 l4 和 l5,是语义块构成标记,也就是短语标记,今后为陈述方便起见,不妨把语义块区分标记叫语义块标记,而把语义块构成标记叫短语标记。语义块标记有‘前后’之分,短语标记更有‘前中后’之分。对于标记的前后之分,根据汉语的习惯引入了默认规则,也许取消这一默认是一劳永逸的明智做法,但是,短语标记的‘中’我想是绝对可以默认的,当然严格说来这也需要验证,说不定某些土著语言违反这一默认?这是闲话,实际上

当然不必这样杞人忧天。这里顺便再说一点不是闲话的闲话,逻辑概念节点 l5 直到去年仍处于备用状态,为什么?就是由于感到短语的前后标记要慎重处理,麻烦出在问话标记的表示,这个问题至今仍是悬案。现在看来,问话标记可纳入 f22,这样,就可以把 l5 定义为短语的前后标记,把 l4 定义为短语的中间标记,善哉!善哉!

短语前后标记是形式上或语法意义上的说法,在语义上就是对基本概念范围的语言逻辑表示,因此,以上所说并不影响对 l5 的已有定义。

“关于主辅语义块转变的问题”初稿的根本问题就在于阁下似乎对语义块的主辅两可性重视不够,因此写了上面的话。

今天就写到这里。随着三文的进展,我也许会作进一步的评述乃至争鸣。

黄曾阳

1998 年 5 月 22 日

自然语言语句的 HNC 表示*

刘志文　庄咏□　郝惠宁　萧友芙

（中国科学院声学研究所，北京 100080）

提要　HNC 理论提出的语义块感知和句类分析是对人类语言感知过程的初步模拟，这一处理模式的实现首先有赖于语句 HNC 表示式的构成和句类格式的形式化，本文将对这两个基本问题和语义块构成及分离问题进行阐述。

1　引　　言

自然语言语句的 HNC（Hierarchical Network of Concepts）表示是概念层次网络理论的重要组成部分，是模拟人类语言感知过程的一种理论模式，是句类分析的策略基础，在文献[1]中仅作了简要说明，有关论述已经或即将发表[2 6]。本文将就语句 HNC 表示的三个基本问题，句类和语义块的 HNC 表示式、句类格式、语义块的构成和分离作进一步的阐述。

2　句类和语义块的 HNC 表示式

HNC 理论的基本假设之一是：人类对自然语言的感知，是以语义块感知和句类辨识为基础的，语义块是句类的函数。基于这一假设，语句的 HNC 表示式就是语义块 HNC 表示式的线性组合。因此，语义块的 HNC 表示式的构造就成为构造语句 HNC 表示式的关键。

文献[1]指出：自然语言的主语义块有 4 种：特征 E、作用者 A、对象 B 和内容 C；辅语义块有 7 种：条件 Cn（Condition）、手段 Ms（Means）、工具 In（Instrument）、途径 Wy（Way）、参照 Re（Refer）、因 Pr（Premise）、果 Rt（Result）。语句的 HNC 表示式仅考虑主语义块，不考虑辅语义块，因为辅语义块弱依赖于句类，它是否带语义块指示标记不受句类格式（见下文）的影响。因此，把它们排除在语句 HNC 表示式之外是必要和合理的。

但应该指出，E、A、B、C 仅描述了语义块的共性特征。语义块的个性特征是它的句类属

* 本文发表于《语言文字应用》1998 年第 2 期（总第 26 期）。

性。语义块的共性和个性两个侧面应视为语句二维空间的两个正交基底。

按照这一思路,语义块 HNC 表示式的一般构成形式应是:

“个性 + 共性”=“句类信息 + 语义块类型信息” (1)

两类信息都用大写字母和数字的串接形式来表达。句类信息项中,字母代表基本句类,数字代表子类;语义块类型信息项中,字母代表语义块类型,数字代表类型的子类。

表示句类信息的字母有:X、P、T、Y、R、S、D,它们分别表示作用、过程、转移、效应、关系、状态和判断。表示语义块类型的字母有 A、B、C,它们分别表示作用者、对象和内容。对仅含句类信息的语义块称为 E 块,对同时含有句类信息和语义块类型信息的语义块称为广义对象语义块,记为 JK。

例如,X2、X2B、XAC、X2C 分别表示反应句(作用句子类之一)的反应、反应者、反应引发者及其表现、反应者的后续表现 4 种语义块,这里,X2 是 E 块,其他都是广义对象语义块。又例如,TB、TC 是转移句的对象和内容,而信息转移句(转移句子类之一)的对象和内容分别记为 T3B、T3C,关系的双方分别记为 RB1、RB2,等等。

这样,语句的一般 HNC 表示式 EJ 可写成:

$$EJ = JK1 + E + \sum JK_m \quad (2)$$

这个表示式左方的 EJ 就是‘语句 HNC 表示式’的符号表示;右方的 JK1 称为 1 号广义对象语义块,其余类推。形式上,JK1 相当于传统语言学的主语。在表示式(2)中,E 块安排在 1 号和 2 号广义对象语义块之间,这符合 SVO 语言(包括汉语和多数印欧语)的天然习惯,这种语义块排序是句类格式的基本类型之一(见下文)。表示式(2)并未限定 JK 的个数,但对于基本句类,实际的自然语言只需要考虑 JK 个数为 1、2、3 的情况。它们分别相应于两主块句、三主块句和四主块句。

对于四主块句,JK2 一定以对象 B 为主体,JK3 一定以内容 C 为主体(参看下表)。对于三主块句,B 或 C 都可以充当 JK 的主体。对于两主块句,可以没有 E,但这时 JK2 必须以 C 为主体,汉语的状态句经常出现这种情况。这些都是概念层面的最重要、最基本的句类知识。

句类有基本句类、混合句类和复合句类之分。基本句类是指表述作用效应链一个环节的句类;混合句类是指用一个 E 块同时表述作用效应链两个或两个以上环节的句类;复合句类是指用两个或多个 E 语义块表述作用效应链不同环节的句类(这里说的作用效应链是广义的,包括判断)。

混合句类和复合句类的语句表示式将分别用 E1E2J 和 E1 * E2J 来表示。表示式(2)实际上是基本句类的表示式,也就是下文将要说明的标准格式。

自然语言的基本句类有 7 种,其一级子类有 57 种。混合句类有 36 种,其一级子类在理论上应有 56 × 57 = 3192 种,但语言中常用的不到十分之一。

基本句类及其部分子类的 HNC 表示式如下表所示。

句　类	JK1	E	JK2	JK3
一般作用句	A	X	B	
承受句	X1B	X1	X1BC	
反应句	X2B	X2	X2BC	
免除句	X3B	X3	X3AC	
约束句	A		X4	X4B
效应句	YB	Y	YC	
双对象效应句	YB1	Y02	YB2	
过程句	PB	P		
因果句	PBC1	P21	PBC2	
转移句	TA	T	TB	TC
信息转移句	TA	T3	TB	T3C
交换句	TA	T4	T4B2	T4C
关系句	RB1	R	RB2	
单向扩展关系句	RB1	R	RB2	RC
状态句	SB	S		
三交换状态句	SB	S02	SC	
判断句	DA	D	DBC	
比较判断句	DBC1	jD0	DB2	

3　句类格式的表示

本节介绍句类格式、句类代码和句类转换的概念及其表示方法。

1. 句类格式

句类格式的定义是:语句中主块的排列顺序。这个顺序有“标准、规范、违例与省略”4种类型,相应于标准、规范、违例与省略4种格式。

标准格式的特征是:主块按语言的自然逻辑顺序排列。

规范格式的特征是:主块的排列顺序违反了语言的自然逻辑排列顺序,因而偏离了标准格式,但在广义对象语义块之间一定要加指示标记。对三主块句,规范格式有4种,汉语中常用的为两种。对四主块句,规范格式有23种,汉语中常用的和比较常用的共9种。

违例格式的特征是:在广义对象语义块之间部分或全部省略指示标记。对三主块句,违例格式有4种,汉语中常用的为两种。对四主块句,违例格式有47种,汉语中常用的和比较常用的共4种。

省略格式是指句中省略某一个语义块。

2. 句类代码

句类代码是语句表示式的编码,是句类知识表示的总纲,它决定主块的数量、每一主块的基本内涵以及各主块的排列顺序。其基本表示式为:

$$EJkmn \tag{3}$$

数字序列 kmn 就是句类代码。对于混合句类,E 写成 E1E2;对于复合句类,E 写成 E1 * E2。上述四种类型句类格式的代码表示式如下:

EJ0mn	表示标准格式	EJ2mn	表示违例格式
EJ1mn	表示规范格式	EJ3mn	表示省略格式

(4)

对三主块句,没有表示式中的 n。应用句类代码,可表达语句各种复杂情况。

3. 句类转换

一个语句所需要表达的内容,不仅可以采用同一句类的不同格式,而且可以采用不同的句类,这种句类之间的变换称为句类转换。句类转换在本质上是一种特殊形式的混合句类。其表示式为:

$$(E2, E1)J \tag{5}$$

式中,E2 是原句类,E1 是转换后的句类。

例: 中国人民爱戴周总理。 (反应句的标准格式)

中国人民对周总理非常爱戴。 (反应句的规范格式)

周总理深受中国人民的爱戴。 (反应句转换为承受句)

对转换后的语句进行句类分析时,关键是要排除 E1 的干扰,恢复到原句类进行处理,因为各语义块之间的关联性决定于 E2 而不是 E1。

4 汉语中的语义块构成和分离的表示

1. 广义对象语义块的构成

广义对象语义块的构成有良性与非良性之分。良性构成的特征是,各块素的排列顺序是确定的;非良性构成的特征是,各块素的排列顺序是不确定的。

例如,作用句的对象语义块 B = XB + YB + YC,其中,XB 为作用对象,YB 为效应对象,YC 为效应内容。这三个块素的顺序不容颠倒,属于良性构成。又例如,“张先生怕李小姐的脾气,张先生怕脾气乖张的李小姐,张先生怕李小姐发脾气”,这三句均为反应句,前两句 X2BC 的构成方式分别为 X2BCB + X2BCC 和 X2BCC + X2BCB 两种排列顺序,第三句的 X2BC 则扩展为一个语句(即下文要讨论的块扩)。由此可见,反应句的 X2BC 语义块属于典型的非良性构成。

广义对象语义块的这种良性、非良性表现密切依赖于句类。在语句表示式中，ABC 字母连用的语义块一般具有非良性表现，而单字母的语义块一般则具有良性表现（见上表）。

2．E 语义块的构成

特征语义块 E 是典型的良性表现语义块，它的各项构成有严密的排列顺序，一般表示式为：

$$E = QE + EQ + EH + HE \tag{6}$$

式中，QE 为 E 块的修饰部分，EQ 和 EH 是 E 块的核心部分，HE 是 E 块的补充部分。在 EQ 和 EH 之间还可以插入 EH 的说明部分，这个插入成分可以不加表示。QE 通常包含势态、情态、时态和性态四部分，其中，势态、情态和时态在 HNC 符号体系中都有特定表示。

HE 一定是由基本概念的序、时间、空间、数、量与范围、质与类和度构成。

E 语义块的核心部分 EQ 与 EH 有多种构成方式，主要有高层概念与底层概念、vv 概念与 v 概念、动态概念与静态概念三种搭配方式。

以上三点，都是概念层面最宝贵、最重要的句类知识。

式（6）中 EH 的后面，汉语可以另加附属成分 hE，它大体相当于传统语言学的“助词”。EQ 的前面，汉语也可以另加附属成分 qE。

3．块扩与句蜕的表示

语句与语义块可以相互转换。语义块向语句转换称为“块扩”，语句向语义块转换称为“句蜕”。块扩的表示式为：

$$JK := J \text{ 或 } JK = J \tag{7}$$

例如：“我们告诉他这个情况”和“我们告诉他王先生已抵达北京”，都是信息转移句，它的语义块 T3C 可以扩展为语句，这是概念层面的句类知识，例句中第二句的“王先生已抵达北京。”就是块扩。

句蜕的表示式为：

$$JK = \langle J \rangle \text{ 或 } JK = \langle (J) \rangle \tag{8}$$

例如：“张三痛打了李四，被张三痛打了的李四，痛打了李四的张三，张三对李四的痛打”。后面的三个语义块都是由第一个语句蜕化而来。

4．语义块分离的表示

E 语义块一般存在分离现象，但这种分离现象不难处理。广义对象语义块的分离要引起高度重视，这种分离现象在标准格式中不会发生，但在规范格式中却经常出现。

广义对象语义块的分离会带来语义块感知的困难，这时语义块的数目多于句类代码所标明的个数或改变了某些语义块的构成，这个问题是句类分析的难点之一，但实践证明，只要存在语义块分离的指示信息，软件就可以对这一现象游刃有余。

5 结束语

语句的 HNC 表示是 HNC 处理技术的‘语义块感知’和‘句类分析’两大模块[5]的基本依托。相应软件[6]已顺利通过了大量真实语料的检验，我们将于近期发布有关结果。应该指出，对于语音输入或汉语拼音键盘输入，汉语单音词的巨大模糊对语义块感知仍是一个巨大的障碍。我们认为，汉语的音节应作为一个特殊的语言信息单元进行系统深入的研究，并寻求它的特殊知识表示方案。

参考文献

[1] 黄曾阳. HNC 理论概要. 中文信息学报，1997，(4)

[2] 林杏光. 正确引导汉语理解与汉语研究——事关人工智能研究的一个重要前提. 科技导报，1997，(4)

[3] 苗传江. 自然语言理解的新进展——简评黄曾阳先生创立的 HNC 理论. 科技导报，1998，(3)

[4] 黄曾阳. HNC 理解处理论文选录，1996

[5] 张全. 基于 HNC 理论的语义块感知处理. 中国科学院声学所博士学位论文，1996

[6] 晋耀红. 基于 HNC 理论的句类分析系统的设计与实现. 中国科学院声学所硕士学位论文，1998

HNC 的句类分析与传统的句法分析的比较研究*

晋耀红　张　全　杜燕玲

（中国科学院声学研究所，北京 100080）

1 引　　言

HNC（概念层次网络）理论是面向整个自然语言理解的理论框架，它包括概念体系的理论模式、语义块和语句的理论模式、句群关联性的表述模式、篇章要点的表述模式、短时记忆和长时记忆的生成以及相互转换的理论模式、基于文字文本的自学习模式。本文仅在语句层面，对 HNC 的句类分析和传统的句法分析进行比较研究。

本文所说的传统的句法分析就是指以生成句法树为目的的一种分析方法，它以短语为处理的基本单元，即所谓 S = NP + VP。方法有两种：一种是从上到下；一种是从下到上。HNC 的句类分析，是 HNC 语义块及语句层面理论的工程实现。它是以语义块感知为切入点，形成句类假设，转入句类假设检验，最后进行语义块构成处理。这一处理策略从根本上来说，就是以语义块为处理单元，在句类知识的启发指导下，确定语句的深层结构，并进一步进行解模糊处理，实现语句的理解。

那么，HNC 的句类分析与传统的句法分析到底有哪些具体的不同呢？下面拟从处理思路和处理策略两方面对这一问题进行分析。

2 处理思路的不同

2.1 处理的基本单元：短语—语义块

从乔姆斯基的短语结构语法到美国现代计算语言学家马丁·凯的功能合一语法，都是以句法树为语句表达的基本框架，这种框架的基本单元是短语。一个句子由若干个短语构成，每个短语充当一定的角色，各角色互相配合，构成一棵结构完备的句法树。相应的，以这些

* ① 本文被选入 1998 年 11 月北京召开的“'98 中文信息处理国际学术讨论会”论文集；② 本文受到国家“九五”重点攻关项目“计算机中文信息处理平台及产品开发”之课题“汉语理解系统的核心技术”的资助。

语法理论为背景的所有句法分析法都是以短语为基本单元的。

HNC 的句类分析是以语义块为基本单元的,语句是由一个个语义块组合而成,语义块是句类的函数(不同句类的语义块是不同的)。

以短语为基本单元的句法分析,在形式上对西语是适用的,因为西语有丰富的短语标记,如‘the,a,for,with...’之类的词。另外,西语的词语形态变化比较明确,词在句中的语法功能相对稳定,句法规则比较严密。这使得可以以词类组合这种简洁的方式生成短语。但是,汉语是非形态语言,词在句中的句法功能不稳定,句法规则很灵活。这给汉语的短语生成带来了很大的困难。

“汉语没有像西语那样完备的短语标记,但有‘把、被、对、向、就……’等之类的语义块指示符,……这是汉语与西语之间的基本差异之一”(论题 1)。HNC 的句类分析正是抓住了汉语有语义块指示符这一特点,提出以语义块为处理的基本单元,从语义块感知入手,辨识出句类,进而进行句类分析,以达到理解语句的目的。

句类分析是面向整个自然语言的,它不仅适用于汉语,同时也适用于西语。对西语必须从短语到语义块,才能真正走向理解。

2.2 深层结构:句法树—句类

要想理解语言,必须寻找它的深层结构,这一点是毫无疑义的。

黄昌宁先生曾指出:“从生成语法的观点来看,句法分析的任务是依据一部语法来判断一个任意的输入句子是否是合乎该语法的句子,如果是则为这个句子建立某种数据结构(如树或多重表)以描述其句法结构。”由此可见,生成句法树是句法分析的主要目的。

句法树以词为它的终结点,以短语为非终结节点,非终结节点是在终结点归并过程中动态产生的。句法树就是由许多终结点和非终结节点组成的。但是到底一个句子应该由多少个短语、多少个词构成,这一直是句法分析感到困惑的问题。相应的,一个句子的句法树有多少个,句法分析也给不出明确的回答。句法树以述语动词为中心,但是句法树的述语动词的分类没有和句子的分类挂钩,它们之间不存在映射关系。

HNC 的句类分析和传统的句法分析有本质上的不同,它“提出了以主语义块构成语句的数学和物理表示式的概念,对语义块提出了特征语义块 E、广义对象语义块 JK 和辅语义块 fK 三分类的概念,对句类提出了基本句类、混合句类和复合句类的三分类标准,对基本句类提出了按作用效应链划分的基本标准,从而得到了 7 个基本句类和 36 个混合句类的重要结论,并进而穷尽了对基本句类一级子类和混合句类一级子类的发现,即语句物理表示式的穷尽发现。以语句物理表示式为基点,我得出了广义对象语义块 JK 是句类函数的结论,而句类可由特征语义块唯一确定(但可能存在多个句类代码),这样,我就完成了建立语句全局联想脉络的预定目标。……句类分析不过是这些理论结果所揭示的水到渠成的语句深层分析之路”(论题 14)。

句类分析以语句的物理表示式为基点。文献[2]给出了这些表示式的详细清单。以语

义块构成的语句物理表示式是语句深层结构的描述。一个句子有几个语义块,各语义块的基本特征,句类分析都可以作出明确的预期。句类分析的过程实质上是概念联想的激活和验证过程。如下面的例子,HNC 通过语义块感知和句类分析,给出分析结果如下:

巩固占领阵地时,应　按匆促防御要领　组织　防御。

fK　　QE　　fK　　X　　B

句子的物理表示式是:XJ = X + B,这是一个省略了作用者 A 的作用句。它有两个辅语义块,第一个辅语义块由句蜕块构成,辅块标志"时"在块尾。第二个辅块有块首标志"按",不难通过辅块的自足性检验。句子中的"应"是 E 语义块的"上装"QE(论题 4),它是 E 块的激活因子,和"组织"一起构成 E 语义块。这些结果的获得,在句类分析时只需要运用几条简明的句类知识规则。

HNC 认为,所谓语句的深层结构,就是 HNC 发现的数学表示式和物理表示式。这些表示式的基本类型完全确定,复杂类型可以组装,这表明 HNC 的句类体系是完备的,是在自然语言的无限不确定的表观现象中寻找到的有限确定的本质。数学表示式指明句类有多少语义块,物理表示式说明特定的某个句类所包含的是什么性质的语义块。语句物理表示式是对语句全局联想脉络的具体描述,有了它,句类分析才有可能走向模拟大脑感知语言的模式。HNC 的句类分析对前面所举那个例句的处理正是以数学表示式和物理表示式为基础的。句类分析通过语义块感知处理,依据句类知识提供的简明规则,就可以确定动词"组织"是整句的 E 块,辨识出句类是作用句,从而唯一确定了语句的深层结构,形成了句子唯一的分析结果。

2.3　述语动词—E 语义块

传统句法分析的句法树是以述语动词为中心的,而且一个句子一般只允许有一个述语动词,但是汉语的述语动词没有形态标志,而且汉语句子中还可能出现动词"满天飞"的现象,动词有可能飞到主语、谓语、宾语、补语、定语、状语等各个成分上去。因此,"述语动词的辨识"一直是汉语句法分析的难点。

那么,面对汉语句子中多个 v 概念这一现象,句类分析是如何处理的呢?在上面的例子中,共有 6 个 v 概念,HNC 的句类分析确定"组织"是它的 E 语义块,"巩固"是句蜕块的 E 要素,其他几个 v 概念都是伪动词,它们虽然有 v 的属性,但是在这个句子里,它们不表现 v 属性。

汉语的述语动词只有一个吗?以句蜕块扩的语义块理论和 E 语义块主体构成理论,以及复合句的理论为依托,HNC 打破了"一个句子只有一个述语动词"的约束,它认为基本句类确实是只有一个 E 语义块,而复合句、块扩、转换都可以有两个或两个以上的 E 语义块。如:

复合句　　我上街买菜。

块扩　　我们担心亚洲经济风暴会冲击我国经济。

句类转换　　我们要进行低产田的改造。

块扩的多级嵌套,还会出现两个以上的 E 语义块。

打破了“一个述语动词”的约束,就不会对动词“满天飞”而感到束手无策。句类分析演绎地给出了 E 语义块感知的一整套处理策略,这些策略覆盖了汉语多动词出现的各种可能。利用这一策略可以顺利地得到“组织”是上面语句的 E 语义块。

2.4　词性标注—概念激活

要进行句法分析,必须确定词类,进行词性标注。这是汉语句法分析的常识。但是汉语没有形态变化,词的兼类现象比较严重,词在句子中的句法功能极不稳定。虽然大规模的电子词典详细地表述了词汇的词类、句法、语义特征,给词性标注的实现打下了很好的基础,但是正如朱德熙先生指出的:汉语的两个区别于印欧语系的显著特点之一是,汉语词类与句法成分之间不存在简单的一一对应关系。所以根据标注的词性,句法分析无法确定词在句子中的功能,也就无法给出句子的深层结构。

如果说,西语是以形态变化来表现词的句法功能的话,汉语就是以词义组合方式来实现它的形态变化的。我们必须抓住汉语这一特点,绕过词性这些表层现象,直接进入语义层面,通过概念关联来探求它的句法功能。“人对语言的理解实质上是一个对概念的激活、联想、浓缩、存储的过程,”体现在句子分析中就是通过概念激活语句的深层结构,即句类。HNC 的句类分析是以它在概念层面的知识表述体系为基础,知识表述体系中的概念是按照作用效应链分类的,作用效应链同时也是句子的分类标准,所以概念与句子成分之间有一定的对应关系,通过概念的激活就可以得到句类,从而也就抓住了语句的深层结构。这种概念激活的方式彻底解决了汉语“词无定类,类无定职”的困扰。

2.5　分词:瓶颈—瓶底

汉语的词与词之间没有明确的切分信息,存在词的切分歧义,需要进行分词处理。分词是汉语的特殊需要,也是句法分析的必然要求。当我们面对汉语拼音文本时,分词模糊显得更为严重。中文信息界许多专家在这方面做了很多工作,但是目前还没有一个令人信服的分词结果,分词处理仍然是句法分析的“瓶颈”。

分词实质上是在消解词的切分模糊,由于汉语不仅“词无定类”,而且“类无定职”,所以仅仅依靠句法功能,不可能彻底解决分词模糊,必须进入语义层面,根据“词义”来确定词的功能。HNC 句类分析对汉语的词切分模糊是在语义块构成处理时才作彻底的消解。也就是在语义块感知成功,得到语句全局联想脉络,完成句类分析后。在语义块感知和句类分析时,允许存在词的切分模糊。我们用下面的例子来简要说明 HNC 句类分析对汉语这一特殊负担的对策。

zhong guo d jing ji hui fu dao li shi zui hao shui ping.

(中国的经济恢复到历史最好水平)

注:各个例子中,“d”,“l”分别是系统的指定字“的”和“了”。

首先进行分段层选,将例子中的有词切分模糊的音组成一个音段,音段之间是明显的切分点,肯定不能组词。分段结果如下:

zhong guo | d | jing ji hui fu dao li shi | zui | hao | shui ping.

其中,“jing ji hui fu dao li shi”这个音段中,有“经济、机会、恢复、辅导、道理、历史”等多个高频词。对这些分词歧义,句类分析先采取包容的态度,对各种切分组合都进行语义块感知和句类假设。假设结果是这个句子可以有两种构成,一个是以“恢复”为 E 语义块,一个是以“辅导”为 E 语义块。对这两个假设作优先级判断和句类检验,最后,句类分析将确认“恢复”的假设是合理的。根据这个结果,将这个音段里“hui fu”的模糊消解掉,然后在“恢复”的句类指导下,进行语义块构成处理,确定这个音段为“经济恢复到历史”。至此,这个音段里的词切分模糊才完全消解。由此看出,HNC 的句类分析将“分词”看作水到渠成的“瓶底”,而不是“瓶颈”。

2.6 句法—语义

汉语是意合型语言,所以汉语的句法分析是离不开语义的。

黄昌宁先生说过:“汉语句法分析器应该先句法后语义,所谓的先句法、后语义,主要是指先取得输入句子的句法结构表示,然后再据此转换成某种句义表示(例如格框架)。”

句法结构是以句型为依托的,“句型只是句子的形式表述,它惟有列举,不可演绎,总量不定”(论题 1-1)。以这种句法结构作为理解汉语的第一步,实在是背离了汉语“意合”的特点。从句法结构转换成格框架,这需要建立起句法结构和语义解释之间的映射关系,我们暂且不论这种映射关系是否可以建立,我们首先看汉语的格框架是否完备?格语法有两个关键问题没有解决:一、应该分多少个格?二、如何划分必须格和可选格?格语法用于描写汉语更显示出它的局限性:汉语的一些流水句、无动句、连动、兼语、紧缩、动补、省略等结构,无法用格语法一个动词统率一个句子的模式来描述,其中连动句和兼语句尤为突出。以上这些问题的存在就无法获得能够覆盖自然语言特别是汉语语句全貌的完备的格框架。

实际上,中国工程院资深院士陈力为教授早就指出:“汉语语法还未形成规范化,而且人们习惯于非规范化的语法,于是语义研究的重要性比西方语言重要得多。”陈院士这番话,已经暗示我们必须抛开句法 - 语义的思路,从语义深层进行汉语的理解。HNC 就是抓住了汉语这一特点,建立了语义块和语句的理论模式,而且 HNC 的句类体系(数学表示式和物理表示式)是完备的。

HNC 的语义块的物理表示式和菲尔墨的格框架都是对语义角色的表述,为什么格框架不能实现完备表述而 HNC 的物理表示式则可以实现完备表述呢?

首先,HNC 的语义角色有基元和复合之分,复合角色是由基元角色复合而成的。如反应句的反应引发者 XBC,可以由 XB + XC 复合,也可以由 XC + XB 复合。基元角色中的 C 角色基元具有关键性作用,由它产生了块扩和句蜕的重要思想,而后者在“句类辨识”中,起着

决定性的作用。菲尔墨的‘格’是不可分解的,也不可能由一个句子构成。这就产生了“格”的数量不定的问题。

其次,菲尔墨的‘格’是动词的造句特征,仅涉及名词短语,或者说,仅涉及‘主谓宾’。而HNC的基元角色分两大类:E语义块基元(包括X、Y、P、S、R、T、D、jD),广义对象语义块角色基元(A、B、C)。这两大类基元在句类指导下的复合,就是完备地给出了语言的语义角色。

HNC理论极力追求基元性和完备性。正是有了以上这些语义角色基元和一套完整的概念基元,才保证了HNC的句类体系的‘完备性’和概念体系描述的‘完备性’。

以上从六个方面对HNC的句类分析和传统的句法分析的处理思路作了简单的比较。我们可以推断,以生成句法树为目的的传统句法分析器和以语句物理表示式为依据的HNC句类分析,它们的处理策略肯定是大相径庭的。

3 处理策略的不同

句法分析器一般采用多扫描的确定性算法。确定性算法完全排除了回溯和伪并行,任何句子结构一经产生便不得更改,便是最终输出的一部分,在分析过程中,不允许有任何暂时结构被构造。确定性算法不是严格地以从左到右的方式进行分析,它一般是通过‘向前看’和‘等着瞧’,采取多次扫描的方式来进行的。汉语句法分析器的多次扫描是‘从左到右’和‘从右到左’相结合的。“从左到右”是从扫描,它主要进行预处理,在词汇一级作捆绑操作,以构成部分词组;主扫描是‘从右到左’的,主要是依据词类进行歧义处理或归并操作,最终得到合理的句法结构。

确定性算法在理论上确实是比较吸引人的,它避免回溯,控制简单,给计算机的实现提供了很大的方便。但是我们应该看到,自然语言,特别是汉语这种意合型语言,它结构特别灵活,本身就存在很大的模糊,面对这些模糊,HNC的句类分析是确定性和不确定性相结合的。

具体地说,HNC的句类分析在总体处理策略上是确定性的,它的‘中间切入,先上后下’的处理策略,从语义块感知和句类辨识入手,先句类分析,后语义块构成分析,这个处理顺序是确定的,不管语句的模糊有多大,结构有多么复杂,采用这么一个处理策略,HNC句类分析总可以得到一个句子合理的物理表示式。这是HNC完备的概念描述体系以及语义块和语句表述体系所保证的。传统的句法分析正因为没有这两大理论模式,所以它的确定性算法得到的结果实质上是不确定的,它不能保证结果的正确性。

HNC的句类分析在总体上采用确定性算法的同时,在各个处理层次上,它又是非确定性的。“中间切入,先上后下”的处理策略是以‘假设检验’为基本手段的,“假设检验”贯穿于语义块感知、句类分析和语义块构成的各个层次。而‘假设检验’必然带来不确定性因素。以语义块感知为例,感知的目的是为了抓住E语义块,并进而确定句类。汉语的E语义块

相当复杂,它存在 EQ 和 EH 主体分离现象,同时又有句蜕、块扩、转换等伪 E 的干扰,还有汉语的复合句的构成,所以 E 语义块感知的算法应该也必须是非确定性的。我们试以一个例子来说明这种不确定性。

da ji du pin mai mai d ju cuo ti gao l wo guo d sheng yu.

(打击毒品买卖的举措提高了我国的声誉。)

这个例子中有"打击、嫉妒、买卖、提高"四个"v"概念,那么它的 E 语义块是哪一个呢?随着拼音流的输入,我们首先假设了"打击"是 E 语义块,它构成一个!31 格式的句子,但是,当输入到"提高"时,系统发现"提高"也是"v"概念,这时对这两个"v"概念作 v1 - v2 假设判断,判断结果是,"打击"通过句子自足性检验,它是句蜕的 E 要素,而"提高"是整个句子的 E 要素,从而修正了前面的假设——"打击"从整句的 E 降级为句蜕块的 E。

上面这个例子说明了语义块感知过程中的非确定性算法。那么,句类分析是如何将确定性和非确定性有机地结合在一起的呢?HNC 的句类分析器的对策是智能调度。

智能调度的本质是数据驱动,但这里的数据驱动和句法分析器的数据驱动有本质上的区别。句法分析器的数据驱动是一种自底向上的策略,具体说,就是根据当前的数据(一般是词类信息或者词组标注信息),决定系统应该进行哪一种合一运算,如名词 N 和名词后缀 K 合一成名词短语 NP,合一完成后,NP 就是当前数据,NP 和 VP 又合一成 SP,由此可见,句法分析器的数据驱动主要是判断需要进行那一种合一运算。HNC 的句类分析的数据驱动,是指根据当前数据黑板的信息(主要是句类得到与否,广义对象语义块的数量,单音词感知的 1 概念,(E,1)联合感知的结果等),判断当前处理的层次和要进行的操作。当句类已经得到,就马上从语义块感知转向句类分析,否则继续进行语义块感知。概言之,两个数据驱动不同有两点:一是数据的类型不同。传统的句法分析主要依据词类信息,它的数据是局部的,仅仅可能对相邻的位置产生预测;HNC 的句类分析的数据,都是一些关键激活信息,它们可以决定句子的深层结构,从而引导语句分析生成语句物理表示式,以达到理解的目的。二是驱动的对象不同。传统的句法分析数据驱动的是合一运算,操作对象是复杂特征集;HNC 的句类分析数据驱动的是"假设检验",驱动的依据和目标都是语句的物理表示式。

智能调度是 HNC 的句类分析"中间切入,先上后下"处理策略的必然要求,它完全抛开了传统的句法分析器的自底向上、自顶向下或两者结合的控制机制。客观的说,自底向上和自顶向下的机制,如果面对确定的数据结构,存在完善的推理机制,它们是一种很好的方法,可以使处理的每一步都是有效的、确定的。但是语言中存在着大量的模糊,使用确定的机制去处理模糊,其结果是可想而知的,单就针对汉语句子中"动词满天飞"的现象,它们就显得无所适从。而面对语言的模糊性,智能调度就游刃有余,它不但能决定"中间切入"的时机,也可以有效地"先上后下",而且最关键的是它可以得到句子的物理表示式,从而真正实现语句的理解。像上面的"动词满天飞"的现象,它决不会影响智能调度的决策,影响的仅仅只是调度的工作量。

智能调度不回避"回溯"问题,语言的复杂性决定了对它的处理不可能一次性成功,"回

溯”是必然的。智能调度的子模块“K 调度”主要就是针对“回溯”问题。这里仅以 K 调度的一个功能“单音 E 的处理”为例来说明。汉语中有很多单音动词,智能调度在语义块感知阶段是先回避单音动词的,因为它有可能产生“草木皆兵的困境”。如果一个句子是以单音动词作为它的 E 要素的,智能调度第一次处理肯定是没有发现 E 或者发现一个伪 E,这时必须进入 K 调度,产生回溯,重新寻找单音 E。K 调度的回溯不是完全推翻以前的假设,重新开始。它是依据已有的假设,如 QE、hv 等信息,在可能的位置上作相应的新的假设。从某种意义上说,回溯是智能调度的一个策略,是一个提高效率,重点突破的有效策略。由于建立了适当的回溯机制,HNC 的句类分析也免除了句法分析器采取的多次扫描的策略。

传统的句法分析的多次扫描,不管它是从左到右,还是从右到左,都是为了进行组合,将低一级的单位,如词、词组,组合成高一级的单位,如短语、句子。之所以采用多次扫描,是由传统的句法分析依赖的句法规则系统所决定的。一般每条句法规则由三部分组成:句法短语产生式、约束条件部分、结果传递部分。短语产生式是由词类和短语类构成,约束部分和传递部分是综合利用语法、语义、静态、动态多元化的手段对产生式内部和产生式之间的关系作详细描述。产生式之间必然是递归的,这就决定了句法分析器必须进行多次扫描。而 HNC 的句类分析为什么能取消多次扫描呢?关键在于 HNC 的句类分析依据的句类知识,句子的句类唯一地确定了它的物理表示式,这里没有什么递归可言,也就没有多次扫描的需求。另外,HNC 的句类分析实际上已经用“优先”代替了“规则”;“优先”是软规则,它具有一定的相对性,即使条件完全满足,结论也不一定成立,它上面还有语境的约束;另一方面,如果条件不满足或部分满足,结论也可暂时保留而不完全否定。HNC 的句类分析的整个过程就是优先级的不断调整的过程,正是“优先”的运用,使得 HNC 的句类分析避免了多次扫描,也使得 HNC 的句类分析能从语言巨大的模糊中抓住全局联想脉络,进而达到对语句的理解。

4 小　结

通过 HNC 的句类分析和传统的句法分析的比较,可以看到这两种不同的分析方法代表两条迥然不同的自然语言理解路线。

正如林杏光先生所指出的:“HNC 将‘自然语言理解’的初级阶段正确定位于消解语言的五重模糊(即面对语音流的五重模糊:发音模糊、音词转换模糊、词的多义模糊、语义块构成的分合模糊、指代冗缺模糊,面对文字流的后三重模糊,具有模糊消解能力)。透过自然语言无限和不确定的表观现象,HNC 抓到了沉淀在语句深层的有限和确定的本质,这就是 HNC 在词汇和语句层面的两个‘完备’,即概念描述体系的‘完备’和句类体系的‘完备’。由于有了这两个‘完备’,HNC 的句类分析才冲破了语句理解道路上的重重障碍,达到了消解语句五重模糊的目标,实现了语句理解的第一步。”

致谢　本论文在写作过程中得到黄曾阳先生和林杏光先生的指导和帮助,谨此致谢。

参 考 文 献

[1] 黄曾阳. HNC 理论概要. 中文信息学报，1997，11（4）：11　20.（网址：http://farad.ioa.ac.cn/hzy.html）

[2] 黄曾阳. 自然语言理解处理的 52 个论题. 1998.（网址：http://farad.ioa.ac.cn/hzy.html）

[3] 黄曾阳. HNC 理解处理论文选录. 1996.（网址：http://farad.ioa.ac.cn/hzy.html）

[4] 林杏光. 中文信息界语义研究谈要. 语言文字应用，1998（3）

[5] 林杏光. 正确引导汉语研究和汉语理解——有关人工智能开发的重要前提. 科技导报，1997（4）

[6] 刘志文等. 自然语言语句的 HNC 表示. 语言文字应用，1998（2）

[7] 晋耀红. 基于 HNC 理论的句类分析系统的设计与实现. 中国科学院声学所硕士学位论文. 1998

[8] 张全. 基于 HNC 理论的语义块感知处理. 中国科学院声学所博士学位论文，1996

[9] 苗传江. 自然语言理解的新进展——简评黄曾阳先生创立的 HNC 理论. 科技导报，1998（3）

[10] 何东平. 我国计算机理解语言研究走出新路——黄曾阳创立的概念层次网络理论正在产品化. 光明日报，1998 年 6 月 12 日第一版

[11] 黄昌宁. 研制汉语句法分析器的对策. 计算机开发与应用，1989，5(2)

[12] 孙茂松. 汉语句法分析中的一种多扫描确定性算法及其在篇章理解中的应用. 清华大学硕士学位论文，1988

[13] Fillmore C J. The case for case. In：Bach E，Harms R eds. Universals in Linguistic Theory. New York：Holt，Rinehart，and Winston，1968

[14] Schank R. Conceptual Information Processing. Amsterdam：North Holland，1975

[15] Schank R. Identification of conceptualizations underlying natural language. In：Schank R，Colby K Eds. Computer Models of Thought and Language. San Francisco，CA：W.H. Freeman and Company，1973

[16] Qullian M R. Semantic memory. In：Minsky M ed. Semantic Information Processing. Cambridge，MA：MIT Press，1968

[17] Chomsky N. Aspects of the Theory of Syntax. MIT Press，1965

关于单音节 E 要素感知的处理策略

萧友芙　郝惠宁

（中国科学院声学研究所，北京 100080）

1　问题的提出

本文是应“汉语拼音智能音字转换系统”产品开发组之邀，为语言知识运用策略的指定而写的。该产品预定今年推出 1.0 版本，下文将简称为 1.0 版。

现代汉语从动态上来看，基本上是单音节词和双音节词平分秋色（古汉语以单音节词为主），因此单音节 E 要素的感知显得格外重要。

对于书面语，单音节 E 要素的辨识及其处理过程与从多个双音动词 v 中选定 E 要素类似。对于“口语”（拼音输入是它的简化情况），单音节 E 要素的辨识则要复杂得多。首先需要发现可能充当 E 要素的位置，然后对多个候选者进行验证。

本文的讨论不考虑复合句类中含单音节 E 要素的情况。对于复合句类，两个 E 要素或其中之一属于单音节 E 要素的情况都很难处理，1.0 版基本上不予考虑。

以“，”标记出的拼音串或文字串，不一定是一个句子。因此，HNC 处理提出了“K 调度”[1]的概念。K 调度的本质就是判定该拼音串是不是一个句子。

这项判断离不开句间信息，1.0 版不利用句间信息，因此，从理论上说，它不可能对 K 调度作出完善的处理。

但是，1.0 版必须有所作为。对任何复杂问题，不能陷入“山穷水尽疑无路”的困扰，相反，必须努力寻求“柳暗花明又一村”的出路。

对于单音节 E 要素感知来说，“又一村”的“村”就是现代汉语的下列特点：（1）它常用单音“hv”或单音“QE”作为 E 要素的指示信息；（2）单音 E 要素主要用于口语，在应用文中比较少见。掌握了现代汉语的这两个特点，就不会为单音节 E 感知的巨大复杂性所慑服了。

反过来说，如果对现代汉语最常用的单音 hv，QE，E 麻木不仁，不作出相应的激活响应，那就枉称为模拟人类语言感知过程的 HNC 技术了。因此，应该十分明确：单音 E 感知是同单音 hv，QE 的感知密切相关的。

进入单音 E 感知有两种可能的入口[1]，第一个是在语义块感知阶段未发现双音 E 要

素,第二个是,原来发现的双音 E 在句类检验中被否定,属于句类分析的回溯处理。1.0 版虽然以第一个入口为主,但也必须考虑第二个入口。

2 单音 E 要素的出现分别有下列四种情况 :

(1)(E,hv)联合出现
(2)(QE,E)联合出现
(3)(QE,E,hv)联合出现
(4) E 单独出现

针对这四种情况,应分别采取不同的感知策略,并分别命名为策略 1,策略 2,策略 3,策略 4。这里,应强调指出,决不能把四种策略循环使用一遍。

单音 E 要素感知处理的要点是先选定策略步骤,而这个选定必须由数据驱动完成。这里的数据主要是指以下三个方面的情况。

(1) 音段的分布情况[2]
(2) 语义块的自足性信息
(3) 亮点启示信息

对音段分布状况要优先注意奇音段—奇音段相连和仅有偶段或奇段的情况。

对自足性信息要优先注意 :基本概念语义块,广义 l9 语义块。

对亮点启示信息首先是指定字的运用,其次要优先注意双音“ hv ”特别是它在串尾的情况,双音 QE 之后无双音 E 要素的情况。

特别要指出的是 :语言逻辑概念 l9,相当于传统语言学的指示代词[3],绝大多数情况“ l9 ”充当广义对象语义块 JK 的块首或辅语义块的块首,这个信息对语句违例格式中前面两个广义对象语义块 JK 的分界判定极为宝贵。HNC 理论把传统语言学的人称代词独立定义为“ p400 ”, p400 在语句中或是自足的广义对象语义块 JK 或是语义块的块首,而且主要是广义对象语义块 JK 的块首。同时还应该注意到当 p400 与 l9 紧连时,p400 一定在前,后面的 l9 一般是同位语。

3 单音 E 感知四项基本原则

(1) 先奇音段后偶音段
(2) 先短音段后长音段
(3) 先中间后两头
(4) 先奇音段—奇音段,后其他

4 具体方法

策略 1.(E,hv)联合感知

a. 偶音段内部

b. 奇音段—奇音段连接处

c. 奇音段内部必然产生多音词

策略 2.(QE,E)联合感知

同策略 1,两者差异仅在于 E 位置的前后不同

策略 3.(QE,E,hv)联合感知

a.奇偶音段或偶奇音段连接处

b.奇音段内部

c.偶音段内部必然产生多音词

策略 4.E 单独出现

a.奇音段内部

b.偶音段内部必然产生多音词

5 一般步骤

(1) 如果可疑音段只是偶音段,考虑策略 1 或策略 2。

例如:他 | shuo guo | 这个 | qing kuang |

(2) 如果可疑音段只有一个奇音段,优先考虑策略 4。

例 1:他 | hui hua bei you tian | 了

例 2:他们 | da shou xia | 的 | ren

(3) 如果可疑音段为多个单音段,不考虑策略 4。

例如:我 | ceng | xie | guo shi ge |

(4) 如果可疑音段为奇音段—奇音段相连,应选用策略 1。

我们 | ming tian shang wu da | xia shan tou | 的 | 敌人

(5) 如果可疑音段是奇音段—偶音段或偶音段—奇音段,应考虑策略 3。

总之,单音节 E 要素的感知一定要遵循先易后难,逐步深入的原则,优先处理“hv”和“QE”及“E”联合出现情况。

1998 年 3 月

参考文献

[1] 黄曾阳. 52 个论题之 25:论调度及 K 调度. 见本书

[2] 黄曾阳. 52 个论题之 23:二论音节感知:段接处理. 见本书

[3] 黄曾阳. 论文 1:自然语言语义网络的基本构成及其特性. 见本书

HNC 语言知识库的概念类别符号体系

萧友芙　郝惠宁

（中国科学院声学研究所，北京 100080）

引言

HNC 的符号体系在黄曾阳先生的〈HNC 理解处理论文选录〉【1】【2】【6】中，及其〈HNC 理解处理的 52 个论题系列〉的论题 33　36 中已有系统阐述。本文将从 HNC 知识库建设的角度，对知识库表示项目之一的“概念类别”，进行系统说明。这个项目必须拥有自身的符号体系，因为它在 HNC 知识库的总共 12 个项目中，处于一个很特殊的地位。黄曾阳先生在【21】中，曾将“句类代码”项目称为“HNC 知识表示的纲、统帅和灵魂”。仿照黄先生的比喻，“概念类别”项目可称为“参谋长”。

概念类别对语义块感知、句类格式的假设和句类分析的要素检验提供最简明的信息。为了便于句类分析软件对这些信息的利用，需要从工程应用的角度对 HNC 理论符号体系作相应改造和补充。以上所说，也就是本文的要点。

1　概念类别项目的内涵

作为词知识库的项目之一，概念类别的内涵，要比〈论文系列〉所定义的“概念类别”更为宽泛，前者是工程定义，后者是理论定义。前者包含后者，但引入了一系列新的表示方式。下面对概念类别项目的内涵作具体说明。

（1）概念类别一词的意义本身是很模糊的，它可以是自然语言概念基元意义下的概念类别[1,2]，也可以是语义块和句类意义下的概念类别[3]，还可以是组合结构意义下的概念类别等等。在理论上，对各种概念类别各有自己的范畴约定。本文的概念类别可以说是它们的综合，但作了必要的扩充和补充，以利于软件的应用。当然，在综合时与原来理论上的定义保持完全一致，不作任何修改。

扩充的含义主要是两方面：

第一是打破概念基元类别符号仅用小写英文字母的定义，例如“j”是基本概念的定义，“j1”表示时间。时间是基本概念的重要类别之一，实际上概念层次网络的每一个节点都是

一个概念类别。因此,在“概念类别”项目中,“j1”就是一个类别符号,也就是说它可以包括数字。这样做有利于信息表示的净化,也不违反理论原则。

第二是打破五元组连用或一般概念类别符号连用的种种固有约束,或不确定性,使基元的组装化更加简便和灵活。对任何概念类别分析出基元,然后再对基元进行组装,可以说是HNC方法论的精髓。但如果仅仅利用“HNC符号”项目表达组装,将显得非常繁琐而不堪重负。而“HNC符号”与“概念类别”两个项目配合起来,就能充分发挥“相得益彰”的作用。

例如汉语的“领导”一词,本义是动词,也可以是名词,还可以是“领导者”的简称。这时在“HNC符号”中可以仅写本义“vc441”,而在“概念类别”中,则可以写成“p,v + g + ug”。因为,在语用上“领导”一词主要用于“领导者”义项。这里的组合结构符号“ + ”也有扩充的意义,下文说明。

符号的补充主要是针对动词语用特征作更细致的说明,下文详述。

(2)“概念类别”项目不仅针对传统意义下的词,也针对语素、短语、语义块,甚至语义块的并合,对这些不同类型的情况,“HNC符号”表示将显得烦琐,“概念类别”则可以给出简明表示,以利于软件对这些知识的运用。

(3)传统的词性标注是“概念类别”的重要内容之一,但我们主要采用灵活的组装方式。此文的目的之一是使组装化符合软件设计的要求,优化组装方式或结构。由于作者水平有限,这个目标不是一步可以达到,但本文希望促进这一方面的研究。

2 概念类别符号的定义

为读者阅读的便利,这里也将〈论文系列〉已定义的概念类别一并列出,其中形态分类的“混合类”及工程类的“特殊定义”两部分是本文的发展。

(1)概念基元语义分类

抽象概念

基元符号:

φ —— 基元概念语义网络

j —— 基本概念语义网络

l —— 语言逻辑网络

s —— 以上三类抽象概念的边缘或综合概念

f ——“语法”网络

(注:φ在具体应用中一般省略不写)

复合基元符号:

φj —— 基元概念向基本概念的挂靠

lj —— 逻辑概念向基本概念的挂靠

lφ —— 逻辑概念向基元概念的挂靠

jl —— 基本逻辑概念,不采用挂靠方式,有自身独立定义的层次符号

语法基元：

h ——词语的后语素　　q ——词语的前语素

具体概念

基元符号：

w ——物　　p ——人

复合基元符号：

pw —— 人造物　　gw ——信息物

rw —— 静态效应物　　rvw ——动态效应物（雨、雪、风）

pr —— 有某种职称的人（总统、皇帝等）

rp —— 有某种荣誉的人（英雄、劳模、明星等）

pg —— 非真实的人（神话或小说中的人物）

jw —— 基本物的子集，不采用挂靠方式，有自身独立定义的层次符号

语法复合基元：

gp ——称呼，在这个概念的前后一定有特指概念

h□——无条件概念类别转换

综合概念

基元符号：

x ——物性，修饰具体概念

复合基元符号：

px —— 人造物的物性　　gx —— 信息物的物性

pj —— 人物化的基本概念　　wj —— 物化的基本概念

jx —— 基本物的物性，如：温度、色彩

xj —— 物性的基本特性

pe —— 相当于有法人代表的机构

jgw —— 基本信息符号，语言、文字、各种符号

xjz —— 物性的值，如：体积、长度、重量、密度

xjzz —— 物性值的量词，如：吨、公顷

jzz —— 时间间隔和空间间隔值，如：小时、分、秒、方位角的度、分、秒

（注：p、w、x 概念没有独立定义自身的语义网络，对它们的表述只能借用抽象概念的层次网络符号，即采取挂靠方式表述。）

（2）概念类别形态分类

单元类

基元符号：

v ——动态表示（相当于传统语言学的动词）

g ——静态表示　　u ——属性表示

z ——值表示　　r ——效应表示

复合基元符号：

vv ——构成特征语义块 E 的 EQ

vu ——含 v、u、ug 概念的词

uv ——仅修饰动词的副词

ug ——修饰抽象概念 g、r 和具体概念的修饰词

uu ——可修饰抽象概念 v、u、ug 的修饰词

zz ——量词

zzv ——动量词

hv ——动词后面的时态、势态或性态说明[4]

qv ——动词前面的时态、势态或性态说明[5]

复合类 例如：

v,g ——兼动态和静态概念,优先动态概念

g,v ——兼静态和动态概念,优先静态概念

z,v ——兼值和动态概念,优先值概念

u,g ——兼属性和静态概念,优先属性概念

v,g,u ——兼动态、静态和属性概念,优先动态概念

g,u,z ——兼静态、属性和值概念,优先静态概念

(3) 工程分类

我们把符号体系的字母串和数字串的表示方式扩大,使概念类别项目形成独立的定义。

(a) 一般定义 (jm、jmn、lm、lmn、jwm、jwmn、wjm、wjmn)

例如：

wj2 ——狭义空间

pj01 ——社会空间

wj01 ——广义空间

w9 ——近代物

l0 ——主语义块切分标志符

l02 ——主语义块对象标志符

如：l0200 为作用对象、l0220 为转移对象

l03 ——主语义块内容标志符

如：l0320 为转移内容

l11 ——方式辅块标志符

l15 ——条件方式辅块标志符

l2 ——主语义块搭配标志符,一定要成对出现,分别指示两个语义块,其中至少有一个是主块,另一个可为主块或辅块。

la —— 特征语义块 E 逻辑说明符,主要是修饰动词 v 的副词

lb —— 语句间逻辑说明符,也可充当语义块指示符

(b) 特殊定义

g, h□g —— 兼 g 概念和否定前紧邻 v 的功能。

v, I0 —— 动态表示,可转换成作用效应句。

v, I01 —— 动态表示,可转换成作用句。

v, I02 —— 动态表示,可转换成作用句。

v, I03 —— 动态表示,可转换成作用句。

v, I10 —— 动态表示,可转换成一般承受句。

v, I12 —— 动态表示,可转换成被动承受句。

v, I4 —— 动态表示,经常以‘ ! 310 ’格式表述。

v, I51 —— 动态表示,但是当由它构成语句中的‘ E ’时,广义对象语义块一定扩展为一个句子。

v, I52 —— 动态表示,但是当由它构成语句中的‘ E ’时,广义对象语义块 一定不能扩展为一个句子。

(c) 综合类 例如:

p, v ——兼具体概念 p 和抽象概念 v,优先考虑 p 概念

v, p ——兼抽象概念 v 和具体概念 p,优先考虑 v 概念

v, u, p ——兼抽象概念 v、u 和具体概念 p,优先考虑 v 概念

(d) 义况信息

HNC 知识库把‘ 词频及语境 ’的信息综合成义况信息。将义况的 12 级数字表示分为三档,相应于数字 0 3 级的为一档,4 7 级为二档,8 b 级为三档。对概念类别的多元性表现,若同属一档用逗号‘ , ’隔开,第二档用一个加号‘ + ’标出,第三档用双加号‘ + + ’表示。

举例如下:

v, g + u —— v, g 为一档,u 为二档

v + vu + + ug —— v 为一档,vu 为二档,ug 为三档

p, ug + v —— p, ug 为一档,v 为二档

+ g + v —— g, v 皆为二档

(4) 其他分类

基元和复合基元概念以各种方式的组合构成自然语言的词汇。词汇的自身组合结构可表示词汇在语义块中的角色优先性,以及它可能提供的有关语义块的配置信息,为句类分析提供重要信息。尤其概念类别为 vC、vB、Bv、Cv 组合结构,它们代替了语句的两个语义块或两个语义块的核心部分,已经不是传统意义上的词,而是具有事件表述功能的短语。

vC ——内容组合型。含有内容组合型词汇的句子通常不需要另加内容,而只需补充对象,有时对象也不需补充,就和 JK1 构成一个句子。

vB ——对象组合型。含有对象组合型词汇的句子通常不需要另加对象,而只需补充内容,有时内容也不需补充,就和 JK1 构成一个句子。

Bv ——主谓型组合结构Ⅰ

Cv ——主谓组合型结构Ⅱ

上列四种组合结构都具有反结构。

3 工程应用

下面对一些特殊‘概念类别’符号所提供的特殊信息作具体说明。

vv —— 构成特征语义块 E 的一部分 EQ,对句类不产生影响,但它是寻找决定句类的 EH 的先导。

vu —— 这个复合基元是 S04 句类的重要判据。

hv —— E 要素辨识的重要标志,也是新词辨识的重要标志,属于 E 块的核心部分。

v, g —— 特征语义块 E 优先感知的依据。

g, v —— 在整个拼音流中未发现其他可作为 E 块的‘v’时,就应该考虑这个‘v’作 E 块。

g, u —— 在整个拼音流中未找到 E 块时,应该考虑‘u’概念构成的 S04 句类。

l0 —— 主语义块切分标志符,除了 l00 外,后紧邻的‘v’仅是句蜕中的‘E’应继续往后寻找拼音流的 E 块。

l9 —— 否定后紧邻的‘v’概念。

la —— 它后面的第一个‘v’概念可作为 E 语义块核心部分的优先候选。

jl —— 提供基本判断句信息,是否基本判断句的‘是’一定否定其前后紧邻的‘v’概念。

HNC 知识库‘概念类别’项目与传统语言学的词性标注截然不同。词性只是它的一部分。在表达方式上,除了通常的单项表示外还采用了组装方式。其中心目标是为 HNC 句类分析当好‘参谋长’。为语义块感知特别是 E 团块和多 E 块感知及为句类假设检验的要素检验提供简单明了的信息。

1998 年 7 月

参考文献

[1] 黄曾阳. 论文 1 自然语言语义网络的基本构成及其特性. 见本书

[2] 黄曾阳. 论文 6 概念知识和语言知识. 见本书

[3] 黄曾阳. 论文 2 自然语言的深层结构及句类分析. 见本书

[4] 黄曾阳. HNC 理解处理的 52 个论题(论题 5). 见本书

[5] 黄曾阳. HNC 理解处理的 52 个论题(论题 4). 见本书

基于 HNC 理论的句类分析系统的设计与实现*

晋 耀 红

（中国科学院声学研究所，北京 100080）

概念层次网络理论（HNC）是基于语义的面向整个自然语言的理论框架。它以概念描述体系为基础，以语义块为基本单元，给出了语句深层结构的完备描述——句类。以句类知识为指导的语句分析处理就是句类分析。句类分析既不是基于规则的推理，也不是基于语料库的统计，而是用语句的物理表示式激活语句的全局联想脉络，HNC 认为，这正是人脑感知语言过程的模式。

本文实现了黄曾阳先生提出的“中间切入，先上后下”这一句类分析的处理策略，即以语义块感知和句类辨识为切入点，先进行句类假设检验，然后进入语义块构成处理。依据这一处理策略，本论文设计了相应的分析处理算法。这些算法的全面设计与实现为 HNC 创新的理论思路迈上了产品化和产业化的征程奠定了坚实的技术基础。

语义块感知算法生成语句的结构表示式，句类分析将结构表示式映射到语句的数学和物理表示式，并检验表示式的合理性。句类分析以智能调度为核心，智能调度的本质是数据驱动，数据是由数据黑板组织的。分析处理的知识来源于三个层面的知识库：概念知识库、语言知识库和常识知识库。

句类分析特别适合汉语的分析处理，它实现了语句理解的突破，解决了长期困扰中文信息处理的一系列难题。

1 句类分析系统的总体设计

1.1 基本策略

HNC 将自然语言理解定位于消除语言中的五重模糊（即语音模糊、音词转换模糊、词的多义性模糊、语义块切分组合模糊、指代冗缺模糊），模糊的五重性可以进一步概括归纳为多

* 摘自晋耀红的硕士学位论文（1998）。该论文被声学所推荐申请“中国科学院院长奖学金”特别奖。

义性模糊。因此,解模糊处理实质上是多义(包括歧义)选一处理。多义选一处理的一般原则是依靠上下文的联想处理(即概念的激活、扩展、浓缩、储存)。上下文联想处理包括词语层面联想处理、语句层面联想处理、句群层面联想处理、篇章层面联想处理。第 1 节所说的联想处理仅限于词语和语句层面的联想处理。

HNC 句类分析系统是如何进行词语和语句层面的联想处理的呢?简单地说,就是进行概念之间的关联性处理,我们将称之为概念之间的语义距离计算,这是多义选一的根本手段。HNC 设计的概念层次网络符号以及基于这一符号体系建立的概念知识库和语言知识库为语义距离计算提供了必要的知识。这些知识的纲领是句类知识。"句类知识包括四个方面:句类格式知识、语义块构成知识、语义块之间的概念关联知识、语义块和句类的转换知识。"句类知识是语句的全局联想脉络,是句子深层结构的完备描述,也是语义距离计算的基本依据。

根据以上的基本知识,我们设计 HNC 处理器的基本策略为"中间切入,先上后下"。所谓"中间切入",就是以语义块感知为切入点,排除一切干扰,辨识句类。所谓"先上"就是指,在得到句类后,进行句类检验,在语句一级对处理做合理性检验。所谓"后下",是指在句类的指导下,进行语义块构成分析,消解语句中存在的各种模糊(包括分词、词的多义选一等)。HNC 处理策略我们简称为"句类分析"。

就处理一个具体的语句来说,HNC 首先不是分词,而是分段。随后在音段内部感知特征语义块(相当于"述语动词")及其引导信息和广义对象语义块的引导信息,根据特征语义块假设整个语句的类别(即句类)。这就是语义块感知和句类辨识。假设出句类后,必须在句类知识的引导下,进行语义块关联性检验,并对语句合理性作出判断。这就是句类检验。句类假设检验是 HNC 处理的核心。句类假设检验是对上述四方面句类知识的综合应用,如果句类检验成功,则句子理解正确,根据语义块构成知识消解语句的各种模糊,也就是进行语义块构成分析;如果句类检验失败,则重新进行其他的假设检验。在这一过程中,句类知识起着控制全局的指导作用,是"消解模糊"的最有力的武器。有了句类知识的指导,句类分析就可以"高屋建瓴"地进行语句理解。

HNC 的句类分析策略免除或缓解了传统句法——语义分析遇到的一系列困扰,如:对词性的过分依赖,自动分词成为必须先解决的"瓶颈",多个动词出现时,述语动词的辨识等。

"中间切入,先上后下"的句类分析策略完全抛开了传统的分析控制机制。传统的句法分析过程可以分为两大类:1. 自底向上(bottom-up),它是由数据驱动的;2. 自顶向下(top-down),它是由预期驱动的。它们都不是分析控制的理想机制。语言是人类高级智能活动的产物,对她的理解处理也应该是智能的,不能完全模仿数据信息处理的方法,必须把握人类语言感知过程的基本特征。人类对语言的感知过程就是对概念的激活、整理、重组的过程,这个过程绝不是盲目的、遍历的,它只能是"中间切入,先上后下"的。

"中间切入,先上后下"是以句类的假设检验为基本手段的。它以语义块感知为切入点,去寻找句类,但是在感知过程中又必须以句类为指导,这好像是一对"鸡生蛋"、"蛋生鸡"的

矛盾。要打破这个“鸡"蛋”的怪圈,必须使用假设检验。假设检验贯穿于 HNC 分析处理的整个过程。

HNC 的假设检验,决不是“一网打尽”式的全面搜索,而在于智能调度。何时进行假设,何时进行检验,这些都必须“相机行事”。这个“机”就是句类。当句类没找到时,运用假设寻找句类,一旦得到句类,就必须转入检验,对假设的句类的合理性作出判断。

智能调度的本质是数据驱动,HNC 数据驱动的本质是区分概念激活信息的强弱,对强激活予以优先响应。这里,我们用“优先”代替了传统的“规则”。传统的产生式规则是计算机特别偏爱的硬性规则,如果产生式左边的条件全部满足,则右边的结论一定成立,如果条件只是部分满足,则结论就被否定。而 HNC 的“优先”是软性规则,是相对性的,即使条件全部满足,结论也不一定必然成立,如果条件部分满足,结论也可以暂时保留而不完全否定。“优先”的智能调度将在后面作详细的阐述。

1.2 系统构成

根据以上的处理策略和控制机制我们设计了句类分析系统的框图,如图 1.1 所示。

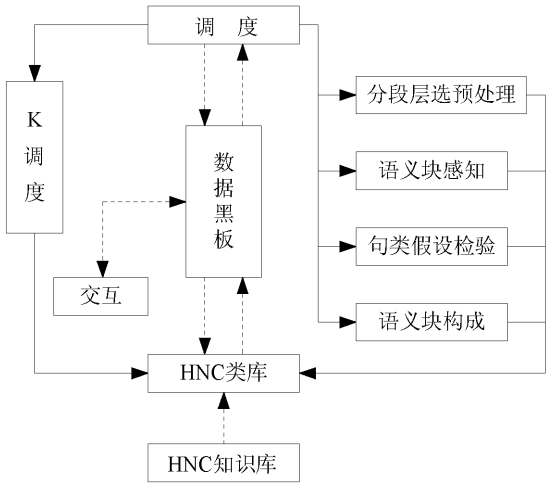


图 1.1 HNC 句类分析系统框图

注:实线表示控制线,虚线表示数据线

从图 1.1 可以看出,HNC 句类分析系统有五大组成部分:1.知识库;2.HNC 类库;3.数据黑板;4.调度;5.处理子系统。

自然语言处理(NLP)必须以语言知识为依托,这是基本常识,HNC 也不例外。HNC 知识库是语言知识的浓缩,它将目标定位在自然语言五重或三重模糊的消解上,是直接为 HNC 处理服务的。

HNC 类库是 HNC 处理的基础,它提供处理需要的各种基本操作,包括各个层次的语义

距离计算,对各层面知识库的调用,还有对数据黑板的操作等一切数据操作。HNC 类库的建立,标志着 HNC 向工程化的方向迈进了一大步,为 HNC 的发展打下了坚实的基础。

数据黑板是 HNC 句类分析系统数据的汇总,它将 HNC 处理的中间数据、结果数据有条不紊地组织到一起,为实现 HNC 智能调度提供有效的依据。数据黑板是一个共享数据集,它记录处理的现场信息,为 HNC 处理策略的实施提供有效的判据。

智能调度是 HNC 处理的指导思想,也是 HNC 追求的目标,只有实现智能调度,才能模拟人的大脑对语言进行处理。HNC 智能调度以数据黑板为原料,根据现场信息和句类知识,作出各种判断,协调各处理子系统,完成句子分析。

HNC 处理子系统完成 HNC 的各项功能,它包括:分段层选处理、语义块感知、句类假设检验、语义块构成、K 调度 5 个子系统。

2 语义块感知

语义块感知是 HNC 处理的切入点,它的基本任务是辨识句类,感知句子中语义块的个数、语义块的类型及构成。

张全博士的博士论文《基于 HNC 理论的语义块感知处理》,已经说明了"中间切入,先上后下"的处理策略是可行的。本论文根据汉语的特点,充分考虑了汉语的块扩、句蜕、语义块分离变换等复杂现象,进行了更深入、更全面的研究,提出了具体可行的语义块感知算法。

本章首先讨论语义块感知算法的设计,然后对感知的核心技术——E 语义块感知作详细的介绍。

2.1 感知算法设计

HNC 认为语句是由语义块构成的,语义块分为主语义块和辅语义块两大类,主语义块是句子语义表达"必不可少"的部分,辅语义块是句义的"可有可无"的成分。主语义块又有四种:特征语义块 E、作用者 A、对象 B、内容 C,四种主语义块在句子中的功能各不相同,下面我们从功能上对语义块作一个划分。

特征语义块 E 是句子的核心,功能上比较特殊,我们将它单独分成一类,仍然记作 E;

作用者 A、对象 B、内容 C 在句子中的作用相对 E 来说比较弱一些,它们的出现是受 E 的控制的,将它们统称为广义对象语义块,记作 JK;

辅语义块的作用相对于四种主语义块来说又低了一级,将它们记作 fK;

另外,为了行文方便,我们将一个句子在形式上记作 FJ。

利用以上这些形式化的定义,我们就可以得到自然语言语句的结构表示式。

简单句的结构表示式可以形式化地描述如下:

$$FJ = \sum_{i=0}^{k}(JK_i) + E + \sum_{i=0}^{m}(JK_i) \qquad (2.1)$$

其中，$m = 0,\cdots,n-1$；$k = 0,\cdots,n-1$；$m+k = n-1$。

这个表示式表明一个基本语句的主语义块数量及其可能的排列顺序。n 表示一个句子的主语义块的数目，n - 1 就是句子中 JK 的个数。（结构表示式中的 JK 加括号，表示它仅仅是一个 JK 的范围，JK 的具体角色并没有确定，这一点是结构表示式和第 3 节将介绍的数学表示式的重要差异。）

汉语的复合句类是由两个句类复合而成，句子中会出现两个 E 语义块，这时，语句的结构表示式可扩展如下：

$$FJ = \sum_{i=0}^{k}(JK_i) + E1 + \sum_{i=0}^{l}(JK_i) + E2 + \sum_{i=0}^{m}(JK_i) \qquad (2.2)$$

其中，$k = 0,\cdots,n1-1$；$l = 0,1$；$m = 0,\cdots,n2-1$；$k+l = n1-1$；$l+m = n2-1$。表示式中，n1 是 E1 的语义块个数，n2 是 E2 的语义块个数。

HNC 用上面的结构表示式描述自然语言的语句，这给计算机对语言的处理带来了很大的方便。语义块感知的过程实质上就是将自然语言的语句映射成它的结构表示式的过程。理论上说，这一映射过程是一一映射的，一个确定的自然语言的文本流唯一地对应一个结构表示式。

为了说明方便，下面我们只对简单句的表示式作详细的介绍。

简单句的结构表示式中，有三个变量数 k，m，n，这三个变量决定了表示式的最后形式，三个变量的各种组合，涵盖了自然语言简单语句的所有出现可能。所以，语义块感知在某种程度上说，就是对这三个变量的确定的过程。

HNC 指出，(k、m、n)是 E 语义块的函数，一个确定的 E 语义块可以得到明确的句类，句类知识又指出各个语义块的出现顺序，也就同时确定了(k，m，n)。所以(k，m，n)可以数学表示如下：

$$(k,m,n) = f(E) \qquad (2.3)$$

对简单句的结构表示式我们可以作简单的数学变换：

$$\begin{aligned} FJ &= \sum_{i=0}^{k}(JK_i) + E + \sum_{i=0}^{m}(JK_i) \\ &= f(E,(k,m,n)) \\ &= f(E) \end{aligned} \qquad (2.4)$$

这样，我们可以得到，一个句子是它的 E 语义块的函数。所以，语义块感知，归根结底，就是 E 语义块感知的问题。

式(2.1)中，两个相临的 JK 之间是怎样划分的呢？汉语的语言逻辑概念给出了明确的回答。语言逻辑概念 l 的设置和具体设计就是围绕着语义块切分和组合这一中心目标而展开的。一般情况下，汉语的两个 JK 之间都有明确的语义块指示符如“把、被、向、对、由……”这类的单音词。这是汉语的一个非常可爱的特点，也是汉语优于西语的一个表现。这样从式(2.1)我们就可以得到：汉语任意两个广义对象语义块之间都有语言逻辑概念或 E 语义块

分隔。又由于 E 语义块一般由 v 类概念构成,所以我们又可以推论出:

推论(2.1). l 类概念和 v 类概念可以将句子切分成多个语义块,也就是说,lv 序列对句子产生语义块的划分。

对简单句来说,一个 lv 序列中只能有一个 v 概念。但是对复杂句,因为它可以有多个 E 语义块,所以它的 lv 序列中可能有多个 v 概念。

实质上,式(2.1)是一个简化的表示式,它没有将辅块映射进去。这是因为辅块是句类的弱函数,弱依赖于句类,一个句子有几个辅块是由语用知识决定的,句类知识无法对它作出确定性的描述,而且辅块的位置可以在任意两个主块之间。但是辅块最大的特点是它的出现必须带 l 指示或 l 标记,l 指示指狭义 l 类概念,l 标记包括综合类概念。所以辅块的出现也可以用 l 概念来界定,推论(2.1)对辅块依然成立。这样一个 lv 序列就是含有 l0、l1、l9、v 类概念组成的激活点的随机序列。

以上的讨论都是针对无模糊文本的,它是基于汉语的词界限已经确定,词义唯一的一种净化讨论。但是,汉语的分词界限本身就是模糊的,而且汉语的词汇一般都有多个义项,所以必须对以上的结论作相应的扩充,才能用它们来指导汉语的语义块感知。

由于以上的结论最后都归结到 l 概念和 v 概念,所以这里首先针对汉语语句中的这两类概念的出现情况作详细讨论,然后根据讨论的结果,对以上的结论作相应的扩展。

一般情况下,汉语的 l 概念都是比较纯净的,它的情况相对来说比较简单,这里不作详细的讨论。以下仅针对 v 概念作讨论。

首先,汉语的词与词之间没有明确的切分标志,存在着词切分模糊,如果面对汉语的拼音文本,这种模糊就显得更为突出,如:shu fu zhu 这个音串中就有(束缚、辅助)两个 v 概念,它们之间存在切分模糊。由于 HNC 句类分析是在语义块构成时才最后解决分词的问题,所以句类分析必须容忍这种模糊。

其次,汉语的词兼类现象比较严重,如"组织防御",这两个都有动词词性,但是在具体的语句中,如"我军正在积极组织防御",只有"组织"表现动词词性,而"防御"仅仅是"组织"的内容。对这种动词连续出现的现象,只有在句类假设检验后,才能最终确定,语义块感知阶段对它们只能采取模糊的态度。

再次,面对拼音文本,一个双音词下可能会有多个动词,比如"tong zhi",就有"统治、通知、统制"三个动词,对这三个动词的确定,显然不应该是语义块感知阶段的任务,而是句类检验的任务。

最后,汉语的词一般都是多义的,一个 v 可能有多个义项,对它的义项的选定,也必须通过句类检验。而且一个 v 可能有多个句类代码,即多个物理表示式,对句类代码的选定,在语义块感知阶段是不能作出判断的。

以上四种 v 概念的模糊性,给传统的句法分析带来了"述语动词发现难"的问题,它一直困惑着中文信息处理界。实际上,这些模糊的最后消解,必须在句类检验后,在确定了语句的物理表示式后,才能完成。语义块感知阶段我们认为这些 v 概念是以"团块"的形式出现。

而这些‘v 团块’同样可以和 l 概念一起将语句切分成多个语义块,所以我们将推论(2.1)扩展为:

推论(2.2). l 概念和‘v 团块”可以将句子切分成多个语义块,也就是说,f‘v’序列对句子产生语义块的划分。

这里的‘v’之所以加引号,是指它是‘v 团块’意思。

由于式(2.1)和式(2.4)都是针对无模糊(无分词、多词性等模糊)文本,要使它们能针对汉语的真实文本(不管是汉字文本,还是拼音文本,都存在词的切分模糊、多词性模糊等),我们也必须对结构表示式作相应的扩展。

上面已经说过,对 v 概念的四种模糊,只能在句类检验后才能消解,所以我们的结构表示式也必须对它们采取模糊的态度。我们将式(2.4)修正如下:

$$\begin{aligned} FJ &= \sum_{i=0}^{k}(JK_i) + \text{“E”} + \sum_{i=0}^{m}(JK_i) \\ &= f(\text{“E”}, (k,m,n)) \\ &= f(\text{“E”}) \end{aligned} \tag{2.5}$$

式(2.5)中,“E’表示它可能是一个多 v 概念组成的‘团块”,里面可以含有多个 v 概念,也可以含有多个句类代码。当然不同的 E 对应于不同的(k,m,n),也就对应不同的结构表示式,所以式(2.5)实质上是一个结构表示式的集合。

f‘v’序列对句子的划分仅仅是一个可能性,序列内可以有多个“v 团块’和多个 l 概念,那么,到底哪一种 f‘v’的组合可以生成如式(2.5)的合理的结构表示式集合呢?这就需要通过‘E 团块’感知来确定。

所谓‘E 团块’感知就是对 f‘v’序列中的各个“v 团块’内部以及‘v 团块”之间的关系作出判断,将 f‘v’序列转换成合理的结构表示式集合。

“E 团块’感知必须在整句的 f‘v’序列生成后才能进行,所以“E 团块’感知一定是音串结束后的动作。当我们面对一个不断输入的拼音流时,在拼音流结束标志出现之前,只能生成 f‘v’序列,当音串结束后,再进行‘E 团块’感知。

从上面的讨论我们可以看出,语义块感知就是要生成语句的结构表示式集合,这个集合是用‘E 团块’来表示的,基于这一点,我们设计语义块感知的算法如下:

1. 将输入的音串分成一个个音段,包容它们的分词模糊;

2. 在音段内,对语义块指示符 l 概念和可能构成 E 语义块的 v 类概念进行激活,形成一个个有效的激活点;

3. 根据这些激活点生成 f‘v’序列;

4. 当音串结束后,进行‘E 团块’感知,将 f‘v’序列转换成结构表示式集合。

依据语义块感知算法,我们设计语义块感知处理框图如图 2.1。从框图中可以看出,感知子系统由 9 个模块组成,每个模块相对独立地完成一个功能。下面对这 9 个模块的功能作简单的说明。

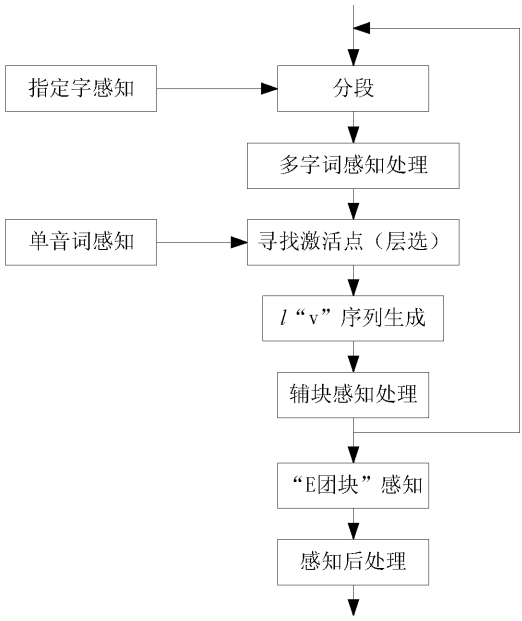


图 2.1　语义块感知处理框图

分段模块：所谓“分段”,就是对输入文本进行切分,将一个词与它前后边界模糊的词划到一个段内。对汉字文本如:制定 新年度 计划,这里,“新年度”即为一个段,因为“新年”“年度”都是词,他们之间存在切分歧义。对这句话的拼音文本:zhi ding xin nian du ji hua,“新年”“年度”“计划”都是词,所以“xin nian du ji hua”将是一个音段。HNC 处理器对拼音文本是以音段为单位进行处理的。分段处理是汉语分析的预处理阶段,也是汉语分析的特殊要求。

指定字感知模块：本系统设置的 7 个指定字“的、了、和、是、有、不、在”,都是汉语的高频字,而且它们对 E 语义块感知影响很大,对它们采取指定的方式,可以大大减轻 E 语义块感知的模糊性。本模块的主要功能是加强或取消指定字附近的 v 概念的 E 假设。指定字一般都单独组成一个音段,所以它还影响分段的结果。

多字词感知模块：本系统对多字词的处理是滞后的,先按照无多字词的情况处理,当在一个音段内,发现多字词的词尾时,才开始对多字词影响到的音段重新处理。多字词处理需要回溯到分段处理,因为多字词的出现影响了原来音段的划分,它将重新生成一种音段划分(多音词单独成音段)。本系统采取的是多字词优先的处理策略,但是对多字词影响到的音段以及多字词前后音段的情况都有所记录,留待以后回溯处理。

单音词感知模块：语义块指示符 l 概念一般都是单音词,所以对 l 概念的感知要求首先对单音词进行感知。单音词感知将在下面详细说明。

寻找激活点(层选)模块：本模块在一个音段内部完成感知算法中的步骤 2,即对 l 概

念和 v 概念进行激活,形成有效的激活点。所谓‘激活’,就是根据知识库中的概念类别和层次网络符号,对一个词的概念内涵进行义类分类,保留 l 类概念和 v 类概念,其他概念一概不作处理。由于本系统一开始没有进行分词处理,而是作了分段,所以这里的激活实质上是在音段内,进行一次简单的层选处理,仅仅关注含有激活点的层选。如在下面的句子中:

bu he li d gui zhang zhi du shu fu zhu l wo men d si xiang .

(不合理的规章制度束缚住了我们的思想)

注:例子中,“d”,“l”分别是系统的指定字‘的’和‘了’,下同。

这个例子中,“zhi du shu fu zhu”是一个音段,它有“制度、读书、束缚、辅助”四个词,其中,“读书、束缚、辅助”都有 v 概念,本模块对这几个 v 概念都激活,给下面的处理提供原料。对这三个 v 概念的激活就考虑了这个音段的两个层选,“制度 束缚 zhu”和‘zhi 读书 辅助”。

f‘v’序列生成模块:根据 l、v 的激活点生成 f‘v’序列。本模块将在 2.3 节详细阐述。

辅块感知处理模块:辅块虽然是句类的弱函数,但它强依赖于辅块类型,它的类型由辅块指示符决定,所以辅块的内涵信息与标志类型有很强的关联性。因此,辅块感知最基本的一条原则就是‘标志信息与内涵信息并用’,由此,产生辅块自足性检验的问题。辅块的标志可以是括号式的,也可以是单向的,对带有单向标志的辅块,如果满足它的自足性的概念出现,就认为辅块已经自足,辅块的感知就此完成。这就是辅块的自足性检验。本模块的主要功能是确定辅块的上下界,利用的信息是辅块标志和由标志所决定的辅块自足性概念。自足性检验的主要手段是语义距离计算。另外,时间语义块是辅块中比较特殊的一类,它的出现可以不带标志,所以本模块对时间语义块进行了特殊处理。

“E 团块”感知模块:本模块是在音串结束后才动作的。它对 f‘v’序列中的‘v’团块,进行 E 语义块感知处理,并进一步将 f‘v’序列转换成合理的结构表示式集合。2.4 节将对本模块作详细介绍。

感知后处理模块:本模块是在句子的标点出现时才动作的。它主要完成语义块感知处理的一些善后操作。首先是对 uE 感知,E 语义块的属性修饰语有两类:第一类,uv 类概念,这类概念的同行性比较充分,是发现 E 块的充分条件,可以作为发现 E 语义块的激活点;第二类,uu 或 u 类概念,他们一般插入到 E 块中间,不能作为激活点,容易和 JK 发生混淆。uE 感知就是对 E 块中间的修饰语进行感知,主要操作是同行优先。其次是 HE 感知处理,HE 一般由基本概念构成,但是基本概念不是映射的关键点,所以 HE 容易和 JK 发生混淆,必须对 E 后的 JK 作 HE 判断。另外,本模块还对句类辨识进行一次修正,对有双重角色的激活点作唯一性操作。

下面我们将对语义块感知的几个主要模块作详细说明。2.2 节将讨论汉语单音词的处理。2.3 节介绍 f‘v’序列生成,2.4 节介绍句类辨识模块的设计。

2.2 单音词感知

单音词感知是汉语的特殊需要。汉语是音、形、义三位一体的单音节语言,一个音下有

多个形,一个形下有多个义。当一个音节作为单音词使用时,在一个具体的语句中,仅取其义项之一,如何去完成这一复杂的多义选一处理呢?HNC 的回答就是单音词感知。

汉语单音词使用虽然灵活,但它也不是无迹可寻的。汉语的单音词主要出现在语言逻辑概念和基本概念中,如"把、被、向、对"等的语义块切分指示符,基本概念中量的多少大小,质的好坏优劣,度的很最,性的正反真假等。也就是说,单音的出现只限制在有限的几个义类中,只有抓住这一特点,才能使单音词的处理循规而行。

单音词的处理分两步:第一步,音—义类;第二步,义类—义项,对每个汉字的义项进行分别描述。HNC 根据这两个需要,设计了相应的知识库,即音节感知库和字知识库。

音节感知库对汉语 400 个音逐个描述,每个音下的汉字都按照 8 类概念分类。这 8 类概念是:

(1) JK 和 fK 感知的逻辑概念 l
(2) E 感知的逻辑概念 E
(3) 构造新词的活跃语素 Y
(4) 基本命名概念 M
(5) 数词 U
(6) 量词 N
(7) 基本概念 j
(8) 动词 v

根据这 8 个义类对单音词进行音节感知,是 HNC 的创新,也是汉语分析处理的必由之路。由于作为语义块切分组合的逻辑概念大多都是单音节,所以音节感知是语义块感知的关键,而语义块感知是 HNC 句类分析的关键。音节感知完成后,对某些概念,如第八类概念 v,还必须进入概念感知,以实现义类—义项的对应,这就是单音词感知的第二步。单字概念感知是音节感知的继续,它是在音节感知的基础上,也是在音节感知的指导下进行的,只有完成这一步,才真正实现汉语单音词"音—形—义"的转化,单音感知才算结束。

单音词感知的结果有两个:一、概念义类分类,指导语义块感知的后续处理。二、概念激活点,为以后的假设检验作准备。

2.3 f"v"序列生成

f"v"序列生成,就是根据句类格式宏观知识,对激活的 l 概念和"v"团块的排列顺序作一次检验,不通过检验的,将暂时不进入 f"v"序列,只记作疑点,留待以后回溯。

f"v"序列生成的基本准则为 lv 准则。

lv 准则包含下列 6 条基本规则:

规则 1:l0 必须在 E 或 EH 之前,不能在 E 或 EH 之后。

规则 2:如果 E 之前只有一个 JK,则一定是标准格式。

规则 3:对广义作用句,如果 E 在句尾,必定是规范或违例格式。

规则 4：对广义作用句，如果 E 在句首，必定是 ‡310 格式。

规则 5：规则 1 可以推广到 l1。

规则 6：如果 l0 之后出现多个 v，则优先选取殿后者为 E。

以下对 lv 准则在序列生成时的使用作简要说明。

规则 1 在序列生成时，是指一般“v”团块后，不允许出现其他的 l0 概念，如果出现，它们将不能组成合理的序列，这个 l0 暂时先不进入 f“v”序列。目前对 l 概念采取宁缺毋滥的策略，排除了 l 概念“满天飞”的可能。

规则 2 是指当一个“v”团块前面没有 l0 概念出现，而且这个“v”团块在句子中间时，则它的后面也暂时不允许 l 概念出现。

规则 3 是指当一个“v”团块在句子末尾时，它的前面一般应该有一个或者两个 l 概念，如果没有，必须回溯，到疑点中间寻找一个可能丢失的 l 概念。

规则 4 是指当一个“v”团块在句首时，它的后面一般情况下只有一个语义块，没有其他的 l 概念。

规则 5 的使用同规则 1。

规则 6 实质上是对“v”团块的优选，如果序列中出现一个 l 概念，那么优先选择最后出现的“v”团块。规则 6 的全面使用是在“E 团块”感知时，这里对它的使用是适可而止的，能优选的，就作出判断，否则，先不作处理。

以上准则的使用，使 f“v”序列的生成有一定的指导，同时，汉语的很多特点，也给我们对“v”团块的排除提供了很大的方便。汉语虽然没有词性标记，兼类现象特别严重，给述语动词的发现带来了很多困难。但是汉语对它的述语动词也作了很多限制，在某些情况下，一个动词是绝对不能作述语动词的，由这些动词构成的“v”团块就可以先行排除。如在“这、那”之后的动词不可能是述语动词等。对汉语这一特点，我们必须在 f“v”序列的过程优先使用。这就是下面的排除准则。

准则一：紧靠“的”前面的动词一定可以排除，紧靠“的”后面的动词，除“是”和句尾（包括以“，”标记的小句）动词外，也都可以先行排除。

准则二：单字词“这、那、哪、某、任”和由它们构成的双字词“这个、哪个”等 l9 类概念之后的动词可以排除。

准则三：单字词“性”，双字词“问题、方式”等“h□g”类概念之前的动词可以排除。这一类概念的作用就是将前面的概念 g 化。

准则四：紧靠单字词“是”的动词，不论前后，一律先行排除，将单字词“是”作为基本判断句 jD 的绝对激活因子。

通过 lv 准则和以上的排除规则的使用，我们可以得到一个比较精炼的 f“v”序列，这就为后面的“E 团块”感知作好了准备。

2.4 “E 团块”感知

在讨论“E 团块”感知之前,首先对 E 语义块的构成作简单介绍。

E 语义块的构成比较复杂,从形式上,它可以表达为:

$$E = QE + EQ + EH + HE \quad (2.6)$$

其中,QE 是 E 语义块的上装,EQ + EH 是 E 语义块的主体,HE 是 E 语义块的下装。下面对这几个构成成分分别作说明。

QE 是 E 语义块的上装,它在句子中是“可有可无”的,但是如果它在句子中出现,它就是 E 语义块的激活因子。QE 一般是语言逻辑概念和基本逻辑概念,它包括:1.势态逻辑判断概念 jlvu,如“应该、必须、可能、也许、必定……”之类。词类上相当于传统的情态动词。2.时态修饰概念 luv6、luv7,如“正在、正、已、将……立刻、马上、即将、尽快……”之类。3.语言逻辑修饰概念 luva,如“再、又……”等。在词类上相当于传统的副词。QE 经常偏离 E 语义块的主体。如:

我们必须对经济秩序进行整顿。

HE 是 E 语义块的下装,它一般是由基本概念构成的一个短语。如:

张三狠狠地踢了李四两脚。

其中,“两脚”就是 HE。它们是对 E 语义块的补充说明。HE 一般都在 E 语义块后,并且在句子结尾。

EQ + EH 是 E 语义块的主体,我们在下面对它作详细的说明。

EH 和 HE 之间,还可以加入我们称之为 hv 的,如“着、了、过、到、来、一下、起来……”等。他们是对 E 语义块的时间或者空间特性的说明。

一个句子的 E 语义块可以由一个,也可以由两个或两个以上的动词构成,一个句子也可以有多个 E 语义块。下面对这两点作简要说明。

1. E 块主体构成,在形式上可以描述为 E = EQ + EH = EQ + E = E + EH。其中,EQ + EH 是为描述如“建立和完善”这种情况的,两个 v 概念之间交式关联,联合起来共同组成一个 E 块;EQ + E 是描述如“开展工作”这种多动词情况的,它虽然有两个动词,但第一个动词是 vv 类概念,要求必须补充另外一个 v 概念,才能形成意义完备的 E 语义块;E + EH 是一种动—静结合的方式,如“做手术,搞对象”之类,这种组合方式是汉语所特有的,汉语有大量的如“作、做、搞”之类的高层 v 概念,他们必须带内容才能形成完整的 E 语义块。

这里,需要强调的是,E 块的主体构成经常会出现分离现象,在它们中间插入一个语义块。如“开展政治体制改革的研究”,其中,“开展研究”构成 E 块,但它们之间插入了“研究”的内容“政治体制改革”。E 语义块分离现象是汉语的特色之一,也是汉语处理的难点。

2. 作用效应型的 E,如“迫使,命令”等,它们要求在对象之后,必须加上第二个 E 要素,形成“E + 对象 + E1”的构成。如:

这件事迫使我们马上采取行动。

如果没有后面的动词“采取”,就会使“迫使”在语义上变得不完整。这就打破了一个中心动词的传统句法约束。作用效应句实质上是传统的兼语句的一种。但是 HNC 认为,“采取”只是对“迫使”的一个补充,它不能和“迫使”并驾齐驱。

3. 句类转换,如句子“周总理受到人民的爱戴”中,虽然有两个动词,但是其中的“受到”只是将句子转换成被动承受句,句子的概念关联性是由“爱戴”决定的。汉语像“受到”这一类高层 v 概念比较多,如“给以、予以、使得”等,它们在句子中往往只起到引导句类转换的作用。

4. 块扩语句和句蜕语句。汉语由于没有关系代词充当从句接口,不得不以行云流水的自然方式把语义块扩展为语句(块扩),或者把语句蜕化为语义块(句蜕)。如下面的例句:

我们担心亚洲经济风暴会冲击我国的经济。(块扩)

我们难以表达全国人民对周总理的爱戴之情。(句蜕)

例句 1 中,有两个动词“担心”、“冲击”,但是它的特征语义块是“担心”;“我们担心”的内容是“亚洲经济风暴冲击我国的经济”这件事,而不是“亚洲经济风暴”,所以必须认为“亚洲经济风暴冲击我国的经济”是语义块扩展的句子(相当于英语的从句,只是这里没有从句指示代词),整个句子结构才合理,如果把它当作兼语句,就显得有点牵强附会。

例句 2 中,虽然“爱戴”前有一个“的”否定它的整句 E 的身份,但是从语义上说,“全国人民对周总理的爱戴”表达一个完整的意思,只有将“爱戴”当作这个子句的 E 块,才能反应他们之间的语义关联性。这个子句就构成了一个句蜕块。

5. 复合句类,它才是典型的两个述语的情况,汉语的连动句和兼语句大都属于这一类。如:

这个报告将提交大会讨论。

我上街买菜。

“E 团块”感知就是对 f“v”序列中的各个“v 团块”内部以及“v 团块”之间的关系作出判断,判断它们之间是否有以上这些复杂构成,并将 f“v”序列转换成合理的结构表示式集合。

这里,之所以叫“E 团块”感知,是因为 f“v”序列中的 v 是“v 团块”,“v 团块”对应的 E 语义块可能不止是一个,它也是一个“团块”,由它转换成的是一个结构表示式集合。

“E 团块”感知可以分为两大部分,一部分是对“v 团块”内部,一部分是对“v 团块”之间。这两部分实质上都是对“v 团块”包含的 v 概念两两之间的关系作出判断。

这里先给出 HNC 对两个 v 概念之间关系的论断。

两 v 概念按出现先后分别记为 v1 和 v2。

规则一:如果 v1 为 Et,则 v2 为 E;(Et 表示引导转换的 E)

规则二:如果 v1 为 EQ,则 v2 为 EH;

规则三:如果 v1 为 E,则 v2 实现块扩或表现句蜕;

规则四:如果 v2 为 E,则 v1 表现句蜕;

规则五:如果 v1 与 v2 共用一个语义块,则 v1 和 v2 分别为 E1 和 E2,它们共同构成复合句。

这些演绎规则覆盖了 v1、v2 可扮演角色的各种可能性,这一点 HNC 理论是可以保证的。

这些规则都适用于"v 团块"这间的判断,如果"v 团块 1"中的 v1 和"v 团块 2"中的 v2,满足以上的某一个规则,则我们认为这两个团块之间也满足这条规则。但是如果这两个"v 团块"之间选择这种关系,则在团块内部必须选择 v1 和 v2。

规则一和规则二也适用于"v 团块"内部的处理,但规则一只能应用于在句子末尾的"v 团块",这一点是 HNC 理论的推断。

这些规则完备地给出了两个 v 概念之间的关系,多个 v 概念无非是两个的延伸。通过这些规则,我们可以给出任意两个 v 概念之间的关系,但这些规则的使用关键在于规则前提的确认。前提的确认主要手段是语句和语义块的自足性检验。那么如何进行语句自足性检验呢?HNC 的回答是 E 语义块"假设、检验"。

E 语义块"假设、检验"的主要思想是:假设一个 v 概念激活点是 E 语义块的构成成分,则结合当前的语义块切分情况,映射出它的结构表达式 FJ;根据假设的 E 语义块的句类知识,判断表达式 FJ 是否是这个 E 语义块的合理映射,如果是,则检验通过,确认这个 E 语义块的构成,否则,否定假设。

这里,E 语义块"假设、检验"判断的依据是句类知识。句类知识主要有四个方面:一是句类格式知识,二是语义块构成知识,三是语义块之间的概念关联知识,四是语义块和句类的转换知识。这里应用的主要是第一和第三类知识。

"假设、检验"对句类知识的运用策略可以概括如下:

第一,判断句类决定的语义块个数是否满足。一个句类所需要的 JK 个数是完全确定的,如果有多种句类代码,选取符合预期数的代码。如果 JK 个数多于预期数,则考虑 JK 分离;如果少于预期数,则考虑省略。

第二,(E,l)联合感知。汉语的 l 类概念作语义块指示符是很有规律的,如"把"只能作作用句的对象指示符,或转移句的内容指示符;"被"只能作作用句的作用者 A 指示符等。(E,l)联合感知就是根据句子中出现的 l 概念,判断这个 E 假设是否满足 l 概念对句类的限制,如果不满足,则否定假设。

第三,基本句类知识的应用。汉语的作用句的 A、B 语义块,反应句的 X2B 语义块,转移句的 TB、TBk、T2C、T3C 语义块等,都各有特定的概念层面知识,这些知识统称为基本句类知识。这些知识具有一发千钧之力,运用它们可以在概念层面对 E 假设作检验。

第四,词知识库中的句类知识如果有可能出现块扩、转换、作用效应句之类的指示,则可以对多个 v 概念的假设进行检验。

根据以上的讨论,我们设计"E 团块"感知的处理框图如图 2.2。

图 2.2 共有 5 个模块,分别是 E 语义块感知模块、块扩感知模块、句蜕感知模块、句类转换感知模块、复合句类感知模块。这里,没有对作用效应句的感知作处理。

E 语义块感知模块就是上述 E 语义块"假设检验"思想的具体实现。它的主要任务就是

利用“假设检验”对以上 v 概念演绎规则的运用。通过演绎规则的使用,将给出“E 团块”的构成情况。如果“E 团块”中没有块扩、句蜕、E 语义块分离以及句类转换等复杂的构成,将直接将 f“v”序列转换成语句的结构表示式集合。如果团块中,出现 E 语义块分离的现象,本模块也对它们作相应的转换。如果感知出“E 团块”中有块扩、句蜕以及句类转换等复杂的构成,本模块将作出相应的指示,引导后面的块扩、句蜕等处理。

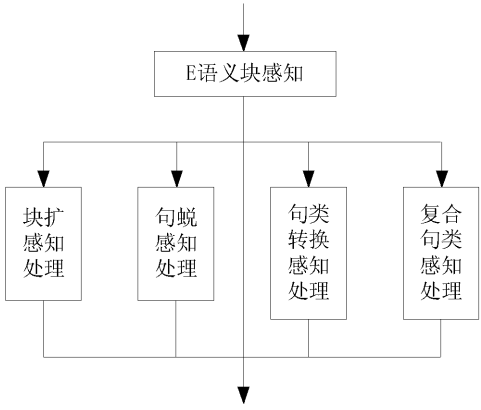


图 2.2 “E 团块”感知框图

下面我们将用具体的例子来对 E 语义块分离、块扩感知处理、句蜕感知处理、句类转换以及复合句类处理这 4 个模块分别作详细讨论。

以下的讨论应该都是针对“E 团块”来说明的,但是为了行文方便,我们还是使用“E 语义块”来作讨论,每个“E 语义块”都可以是一个“E 团块”的一部分。

2.4.1 E 语义块分离处理

E 语义块分离是汉语灵活性的一个表现,它主要是 E 语义块主体的分离。E 语义块主体有三种构成,E = EQ + EH = EQ + E = E + EH,现在仅以 EQ + E 为例对分离加以说明。

EQ + E 有两种分离,一种是 EQ 分离出去,E 在原来位置,另一种是 E 分离出去,EQ 在原来位置上,如:

我们要开始研究政治体制的改革。

我们要开始政治体制改革的研究。

我们要开始对政治体制改革的研究。

这三个例句中,“开始研究”是 E 语义块,“开始”是 EQ,“研究”是 E,但后两个例句是 E 语义块的分离,它们属于不同的分离。

例句 2 是第一种分离,它是将 E 向后分离。因为分离前句子是 !0 格式,句子中不带 l 标记。例句 3 属于第二种分离,将 EQ 向前分离,分离前它的句子格式是 !11 格式,句中含有 l 语义块指示符。

从上面的例子,我们可以看出,E 语义块分离是跟句类格式密切相关的,不同的句类格

式将产生不同的分离情形。同时可知,EQ + E 的分离,不管是那一类型的分离,它的 E 总在句子末尾。这就是分离的两个线索。E 语义块分离的处理就是根据这两个线索对分离情形作出判断。

2.4.2 块扩感知处理

块扩是汉语特有的现象,汉语经常是把第二个语义块即 JK2 扩展成一个句子。如果 E 语义块感知得到两个 E 块,其中第一个 E 块是可以块扩的,这时就需要进行块扩感知处理。块扩处理就是对这两个 E 块分别生成两个映射,并在两个映射之间建立起块扩的关系。

这里结合上面的 E 语义块感知对块扩处理举例说明。

Wo men dan xin ya zhou jing ji feng bao hui chong ji wo guo d jing ji .

(我们担心亚洲经济风暴会冲击我国的经济。)

上面这个例子中,有两个 v 概念"担心""冲击",系统先对"担心"作出句类假设,同时,词知识库中的信息指示"担心"的第二个语义块 JK2 可以扩展,所以将"担心"的状态记为可以块扩;当处理到"冲击"时,对这两个 E 假设作检验,发现"担心"是可以块扩的,又由于"冲击"的 E 语义块假设在子句"亚洲经济风暴会冲击我国的经济"中满足语句自足性检验,所以将这一子句记为块扩子句,"冲击"是子句的 E 语义块,整个子句是"担心"的一个扩展语义块。

这里,语句自足性检验包含两个意思:第一,它的 E 要素所要求的语义块的个数("冲击"的句类是作用句,它要求有 3 个语义块)得到满足。第二,它的 E 语义块需要的要素概念得到满足。语句自足性检验一般只在概念层面进行,不深入到词汇层面。

2.4.3 句蜕感知处理

"句蜕"是汉语中最常见的语言现象,从形式上看,它很像是西语中的一个个"从句"经过包装处理,以"压缩块"的形式嵌在主句中,继续扮演自己本来的角色。"的"的出现经常是句蜕的标志。

句蜕块的一般表示式为:

$$JK = \langle J \rangle \text{ 或 } JK = \langle (J) \rangle$$

这里,JK 代表广义对象语义块,J 代表一个完整的句子。〈 〉是句蜕的表示。

从表示式可以看出,句蜕块在形式上可分为两种类型。

类型 1 JK = 〈J〉

表示一个句子内部,在一个语义块之前加上"的"将句子蜕化。

其中,以下形式的句蜕最为典型

$$\langle J \rangle = \langle \text{!3jEJ} + \text{的} + JK_k \rangle$$

上式可以这样理解:语句 J 中某广义对象语义块 JK_k(全部或部分)在形成"句蜕"形式"〈J〉"时偏离了标准格式而被移至"的"后,该 JK_k 可以是原标准格式下语句的 JK1,JK2 或 JK3,"!3jEJ"表示"未被移走的部分"。

如:

〈102 + JK2 + E + 的 + JK1〉　〈把李四打了的张三〉

〈101 + JK1 + E + 的 + JK2〉　〈被张三打了的李四〉

〈E + JK2 + 的 + JK1〉　〈打了李四的张三〉

当〈J〉中 v 被移至句蜕块尾，JK_k之间有“102”标志，则也可能形成如下形式的句蜕块：

$$〈J〉 = 〈\Sigma JK_k + 的 + 〈E〉〉$$

上式中的〈J〉一定是由非标准格式语句蜕化而来。

如：

〈JK1 + 102 + JK2 + 的 + E〉　〈张三对李四的伤害〉

类型 2　JK = 〈(J)〉

表示在一个整句后面加“的”(或“19”)和一个高层概念构成的块素，共同构成一个完整的语义块。

如：

〈(张三打了李四)这件事〉引起了领导的重视。

〈(李四被张三打了)这件事〉引起了领导的重视。

领导很重视〈(李四被张三打了)这件事〉。

〈(张三把李四打了)的事〉你们知道吗？

〈(打击假冒伪劣商品)的工作〉是一项紧迫的任务。

〈开创新形势下精神文明建设的新局面〉成为〈全党和全国各族人民极其关注的大事〉。

发现句蜕块的“钥匙”有两个：句类代码和“的”字。“的”字是发现句蜕的最直观、最基本的条件。在有“的”的前提下，“的”前或后存在 v，是发生句蜕的必要条件。

句类知识的运用是句蜕处理的关键，只有充分利用它，才能对句蜕进行有效的感知。如：对动宾结构 vB 或 vC 型，出现在“的”前，就有两种判断，例如下面的句子：

1. 这篇讲话回答了〈经常困扰和束缚我们思想的许多重大认识问题〉

2. 防止和消除文化垃圾的传播，防止和遏制腐朽思想和丑恶现象的滋长蔓延，是在社会主义现代化进程中必须认真解决的历史性课题。

在例 1 中，“困扰和束缚”的只能是“思想”，不可能是“问题”，这在束缚的句类知识里有所表示，应判定为句蜕。而在例 2 中，“防止……”的应是“传播”或“滋长蔓延”，而不应是“垃圾”和“思想”、“现象”。这些通过在概念层面对 v 作 BC 检验，进行“同行优先”计算，都可以作出正确判断。因此，尽管有“的”，它也不应作句蜕处理。

根据以上的讨论，我们可以设计句蜕的处理策略：

首先，如果 v 后有“的”出现，并且句子中有其它 v 概念可以作 E 语义块假设，则必须作句蜕判断；

其次，这个 v 作 E 语义块假设，并对假设进行检验，如果检验通过，则判为句蜕，否则，否定这个 v。

下面用例子对句蜕的处理加以说明。

liao jie l zheng shi qing kuang d chang zhang cai qu l jin ji cuo shi.

(了解了真实情况的厂长采取了紧急措施。)

例子是 JK1 的句蜕,它的意思是"厂长了解了真实情况,厂长采取了紧急措施"。这个例子中有两个 v 类概念"了解""采取",系统开始处理时,先对"了解"作出句类假设,当处理到"的"时,将"了解"的状态设为可能句蜕(因为,句蜕经常伴随着"的"的出现),当发现"采取",要对它进行句类假设前,先对"了解"的假设进行自足性检验,检验通过,就将"了解"设为句蜕的 E,将"了解了真实情况的厂长"记为句蜕块,然后再对"采取"作句类假设,并进行后续处理。

2.4.4 句类转换、复合句类感知处理

句类转换是汉语的特色之一,汉语的句类转换形形色色,从句类的观点大概可以分为以下几类:

1. 作用句与效应句的相互转换
2. 作用句与被动承受句的相互转换
3. 一般句类向一般承受句的转换
4. 一般句类向作用句的转换
5. 一般句类向基本判断句的转换
6. 一般句类向效应句的转换
7. 一般句类向状态句的转换

每类转换都有转换规则,我们建立了相应的转换规则库。下面仅以作用句向被动承受句的转换为例来说明。

作用句的句类格式是:XJ = A + X + B

被动承受句的句类格式是:X12J = X1B + X12 + XAC

作用句向被动承受句的转换规则:

$$A + X + B \longrightarrow B + X12 + A + X$$

如:

作用句"亚洲经济风暴冲击日本经济"可以转换成

被动承受句"日本经济受到亚洲经济风暴的冲击"

从以上例子可以看出,转换后,句子中增加了"受到"这个 X12 型概念,其他各个语义块的语义角色都没变,这就是转换必须坚持的原则。我们将"受到"这样的词称为转换引导,将"冲击"称为被转换的 E。转换规则库就是以这两类成分的句类代码为入口的,两个句类代码共同决定转换规则的选择。

句类转换感知的主要任务就是发现转换引导和被转换的 E,根据它们选择转换规则,得到转换后的句类格式,并进一步得到它的结构表达式 FJ,利用这个结构表达式进行语义块

感知。

复合句类的感知是典型的双述语发现的问题。对它的感知处理与块扩的处理有点相似,这里不作详细介绍。

2.5 小结

第 2 节详细介绍了语义块感知的设计与实现。语义块感知以映射结构表达式为目标,以“E 团块”感知为核心,充分进行概念的激活、联想、假设。重点讨论了 f“v”序列生成以及“E 团块”感知。

语义块感知是 HNC 处理器“中间切入,先上后下”处理策略的关键,它的实现,为这一策略的最后实现打下了坚实的基础。

3 句类假设检验

句类假设检验是 HNC 处理的核心,也是 HNC 处理的精髓所在。只有进入句类假设检验,才有可能模拟人类语言感知,有效地进行语义距离计算,立足于概念之间的关联性作出各种各样的判断,并从初级联想处理跨入中级联想处理,并进一步迈入篇章理解的殿堂。

本节详细讨论句类假设检验的算法设计,并对假设过程和检验过程作详细的介绍,最后讨论语义块构成。

3.1 假设检验算法设计

我们知道,语句的物理表示式是语句的深层结构,而只有得到语句的深层结构,才能真正进入理解,句类假设检验的任务就是得到语句的合理的物理表示式。所谓“句类假设检验”,就是在语句一级,根据语义块感知的结构——语句的结构表示式,假设出句类,在句类知识引导下,对语句合理性作出判断,确定各个语义块的角色,并得到各个语义块的要素概念,以此建立语句的全局联想脉络,为后续的篇章处理作准备。

假设检验的第一步当然就是要假设出语句的全局联想脉络,即它的物理表示式。这就是句类假设。句类假设是语义块感知的后续处理,它的处理对象是语义块感知的结构表示式 FJ,所以句类假设实质上也是要建立一种映射,语句的结构表示式到它的物理表示式的映射。如果我们用 EJ 代表它的物理表示式,我们就有以下的映射关系:

$$FJ \longrightarrow EJ$$

一个确定的结构表示式只可能对应一个物理表示式,它们之间是一一对应的。

但是语义块感知的结果并不是确定的结构表示式,感知进行的是“E 团块”感知,“E 团块”内部可能有多种构成情况,由它们得到的结构表示式“FJ”实质上可以对应到多个物理表示式。所以本系统中进行的是一对多的映射,句类假设可能假设出多个物理表示式。建立这种一对多的映射的过程,就是下面要介绍的句类假设的过程。

不同的‘E团块’将映射出不同的物理表示式,也就需要不同的假设过程,所以在假设之前,必须根据‘E团块’的构成,对假设类型作出判断。有了假设类型,才可以引导句类假设进行不同的操作。因此在假设之前必须有一个预处理模块,预处理从结构表示式的‘E团块’分析出假设类型,然后根据假设类型,调用不同的假设操作。

假设出的多个物理表示式,哪一个(或哪几个)是语句的合理的物理表示式呢?这就必须通过句类检验来判断。句类检验的任务就是在句类知识的引导下,确定语句的物理表示式。检验是从两个层次来进行的。

首先,是语义块要素关联性检验。上面说过,物理表示式描述了一个语句语义上的概念关联性,包括E语义块与广义对象语义块之间、两个广义对象语义块之间的概念关联。句类假设主要是从它的E语义块出发,建立语句的物理表示式,但是各个语义块之间的关联性是否满足,这就必须检验语义块要素之间的关联性。如果我们面对的是模糊文本,要素检验的过程也是一个解模糊的过程。

其次,句类检验要进行合理性检验。合理性检验就是给出各种评价函数,包括句类检验的评价函数、语义块评价函数、语句评价函数。这三类评价函数主要从合理和不合理两方面来评价,一般来说,合理性联系于漏报,不合理性联系于虚警,两方面应该相辅相成。先进行两极化处理,然后再给出优先级排序,这显然符合效率优先的原则。

根据以上对句类假设检验的简单讨论,可以设计算法如下:

1. 从结构表示式分析假设类型,形成调度策略。
2. 根据假设类型进行句类假设,将语句的结构表示式映射成多个的物理表示式。
3. 对语句物理表示式进行语义块要素检验。
4. 对语句作合理性检验,给出它的评价函数。
5. 根据评价函数,确定语句的实际的物理表示式。

图3.1的句类假设检验框图是这个算法的具体实现。下面对这个框图作简要说明。

假设检验框图由三部分构成,假设检验预处理、句类假设、句类检验。

假设检验预处理:根据语义块感知的结果,分析出句类假设的类型,并依据这个类型作出假设检验的调度。假设类型中不仅包含“E团块”的现场信息,还包括块扩、句蜕、分离等各种复杂类型的组合信息。如果有这些复杂构成信息,本模块将调用相应模块进行不同的句类假设。

句类假设:完成结构表示式到物理表示式的映射,假设出句类代码、句类格式以及各个语义块的语义角色。不同的句类假设类型将有不同的假设操作,而假设类型中不同的复杂构成也将进行不同的句类假设,所以在框图中,各种复杂构成的分析处理是和句类假设在同一个大框架中。

句类检验:进行语义块要素检验,并作出合理性评价,生成语句评价函数。根据评价函数,给出语句的物理表示式。

下面对假设检验的各个环节分别讨论。

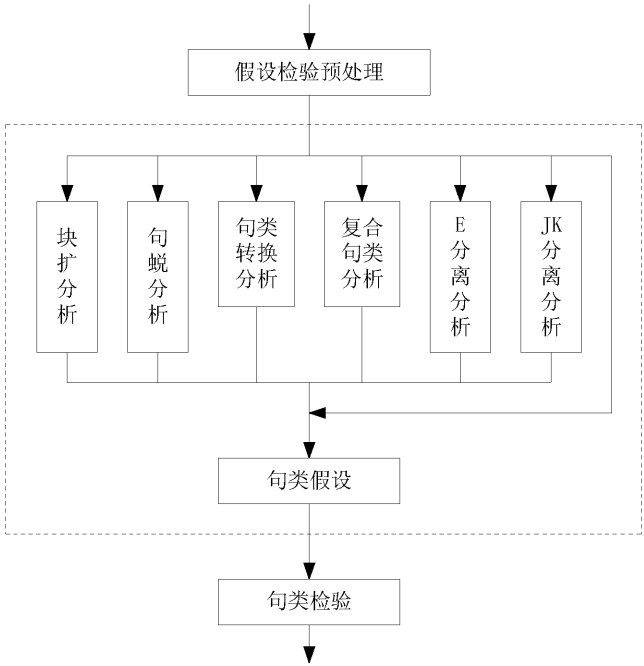


图 3.1　句类假设检验处理框图

3.2　假设检验预处理

假设检验预处理的主要任务就是进行形势分析,根据‘E 团块’的信息分析出假设类型,并作相应的准备工作。

“预处理主要是形势分析,而形势要点决定于:一个 E 团块?两个 E 团块?多个 E 团块?”,这是假设检验预处理的纲领,也是生成假设类型的出发点。“E 团块”的多少,将直接影响到假设检验的处理策略,如果只有一个 E 团块,将主要根据 E 团块位置的首中尾作格式判断;如果有两个 E 团块,将首先进行句类转换、块扩、句蜕以及复合句类等判断操作,并根据不同情况进行假设检验;如果存在多个 E 团块,则需按多重两分处理。所以假设类型必须在 E 团块的个数判定这个大前提下来生成。

依据以上的讨论,我们将假设类型分三级表示,即:

(1)“E 团块”之间的相互关系;

(2)“E 团块”的位置信息(包括它与 l0,l1 的关系);

(3)“E 团块”内部特征。

对每一级假设类型,我们都具体分类如下:

1.“E 团块”之间的相互关系(这里只给出两个团块之间的关系,多个之间的关系无非是两个之间关系的扩展):

(1) 两 E 团块构成 E 块主体,并发生分离;

(2) 两 E 团块构成作用效应型的双 E;

(3) 两 E 团块构成句类转换;

(4) 两 E 团块形成语义块扩展(块扩)和语句蜕化(句蜕);

(5) 两 E 团块构成复合句类。

对这一级假设类型中的各种构成,实质上在语义块感知结果中已经有所指示,通过检查感知结果,可以得到这个指示,我们也就可以形成第一级的假设类型。对一个"E 团块"的情况,在这一级里没有表示,因为它不涉及团块之间的关系。

2."E 团块"的位置信息(包括它与 l0,l1 的关系):

(1) 句尾的"E 团块",并且前面有 l0 概念;

(2) 句尾的"E 团块",并且前面没有 l0 概念;

(3) 句首的"E 团块";

(4) 句中的"E 团块",存在广义对象语义块分离;

(5) 句中的"E 团块",没有广义对象语义块分离。

这一级假设类型主要描述"E 团块"的位置以及它与 l0, l1 的关系,这一信息对句类格式的假设至关重要。

3."E 团块"内部特征:

(1) 仅有一个 E 语义块,而且句类代码唯一的"E 团块";

(2) 仅有一个 E 语义块,但是句类代码不唯一的"E 团块";

(3) 多个 E 语义块在句子同一位置上,如"tong zhi"(统治、通知、统制);

(4) 两个以上 E 语义块,它们之间是串行关系,如"组织防御";

(5) 两个以上 E 语义块,它们在一个音段内有交叉,如"hui fu dao"(恢复、辅导);

(6)"是、有、比"构成的"E 团块"。

这三级假设类型,大部分都可以从语义块感知的结果中得到,只有对第二级中的广义对象语义块分离的分析必须根据感知结果,由预处理来自行完成。广义对象语义块 JK 分离表现在语义块感知的结构表示式

$$FJ = \sum_{i=0}^{k}(JK_i) + E + \sum_{i=0}^{m}(JK_i)$$

中,就是 JK 的个数多出一个,即 k + m = n (n 为 E 要求的 JK 的个数,包括 E 在内)。比如,对以下的例句:

例 3.1 张三被李四打断了腿。

例句的结构表示式是:

FJ = (JK) + (JK) + E + (JK)　　(即"张三 + 李四 + 打断 + 腿")

它的 E 语义块是"打断";"打断"决定的句类是作用句,作用句的句类知识指示它只能有三个主语义块(包括 E 语义块),只能有两个 JK,但是感知到的结构表示式有 4 个主语义

块,3 个广义对象语义块 JK,由此可以判断它的广义对象语义块出现分离。

汉语的广义对象语义块分离一般很有规律。对三主块句,一般是 JK2 发生分离,即它的物理表示式标准格式中的第二个 JK,比如作用句 XJ = A + X + B,它一般只能是作用对象 B 发生分离,作用者 A 经常都比较老实。对四主块句,一般是 JK3 分离,对信息转移句 T3J = TA + T3 + TB + T3C,总是 T3C 发生分离。这可能是汉语在潇洒中的一点严谨吧。汉语的这一规律性很强的特点,给广义对象语义块分离的分析处理带来了极大的方便。对上面的例 3.1,我们就可以映射成如下的物理表示式:

$$XJ = B + A + X + B$$

本系统对广义对象语义块分离的处理思路就是依据这一点而设计的。

预处理模块在分析出这三级假设类型后,针对不同的假设类型,还要为后面的句类假设作相应的准备工作。比如,对第二级的第二类,即句尾的"E 团块",并且前面没有 l0 概念,预处理模块必须进行回溯处理,在句子前面寻找可能遗漏的 l0 概念,不管找到与否,都要将这个信息带给句类假设去处理。句类假设模块将根据这个 l0 概念是否找到,来判断应该假设的句类格式。

假设检验预处理模块在完成假设类型的分析,并作完相应的准备操作后,就进入句类假设模块进行假设。一种假设类型一般对应于"E 团块"中一个或多个具体的 E 语义块,本模块在分析出假设类型的同时,也就确定了每个类型所对应的 E 语义块,这样,经过预处理,3.3 节的句类假设面对的将是一个比较纯净的同类型的"E 团块"。

3.3 句类假设

句类假设就是要完成从结构表示式到物理表示式的一对多的映射,在介绍假设处理之前,先简单介绍一下句类假设依据的理论知识。

我们知道,依据句类知识,每个语句都有一个物理表示式。语句的物理表示式用语义块的物理表示式来表达,例如,

作用句的物理表示式是:

$$XJ = A + X + B \qquad (3.1)$$

式中的 A 表示作用者(施事),X 表示作用,B 表示作用的对象(受事)。

信息转移句的物理表示式是:

$$T3J = TA + T3 + TB + T3C$$

式中的 TA 表示信息发出者,T3 表示信息转移,TB 表示信息接收者,T3C 表示信息的内容。

上面的两个物理表示式中,X、T3 是句类代码,也是它的 E 语义块的名称。其他都是各个语义块的名称,也就是广义对象语义块的语义角色。物理表示式以及它的语义块的名称,精确地描述了一个语句语义上的概念关联性,它是语句联想脉络的形式表示。

物理表示式中各个语义块的顺序可能会发生变化,比如,作用句就有 A + B + X,B + A + X,B + X + A,X + A + B,X + B + A 这几种变化,这些变化构成了一个句类的物理表示式的集

合。所有句类的物理表示式的集合就构成了自然语言语句的物理表示式集合。这个集合涵盖了所有语言的语句可能出现的形式,是自然语言深层结构的完备描述。

针对于一种语言来说,一个句类的物理表示式有一种天然的排列顺序,HNC 把按照这种顺序排列的物理表示式称为这个句类的标准格式。比如,汉语中作用句的物理表示式 XJ = A + X + B,就是作用句的天然顺序,它就是作用句的标准格式。句类的标准格式是语义块排序的一种本征节律,这就是 HNC 提出的‘对象 + 内容’或‘内容 + 对象’的基本节律。标准格式下,语义块之间不加语义块指示符。如下面的例子:

张三 打断 了 李四的腿。

A X hv B

这个例子是一个作用句,它采用的是标准格式,它的语义块 A、B 之间就不加‘把、对、向、被’之类的语义块指示符(其中,hv 是特征语义块构成的一部分)。句类的标准格式可以由句类代码唯一确定,以此我们建立了句类代码知识库。

所有的物理表示式可以概括成下面的数学表示式:

$$S = JK1 + E + \sum_{k=2}^{n} JK_k \qquad (3.2)$$

式(3.2)中,n 是 E 的句类所要求的语义块的总个数,E 表示物理表示式中的特征语义块,JK1 表示句类标准格式中的第一个语义块,JK2 表示句类标准格式中的第二个语义块,依次类推。对作用句,JK1 就代表 A 语义块,JK2 代表 B 语义块。

HNC 把所有的物理表示式抽象成数学表示式,用数学表示式来研究语义块的排序现象。HNC 把语义块的不同排序叫做句类格式。数学表示式为我们脱离具体的句类,抽象地研究句类格式变化,提供了方便。比如,对所有的三主块句(要求有三个主语义块的句类),我们就可以通过数学表示式的数学变换,给出它的可能的句类格式。三主块句的句类格式有:

S = JK1 + E + JK2

S = JK2 + E + JK1

S = JK1 + JK2 + E

S = JK2 + JK1 + E

S = E + JK1 + JK2

S = E + JK2 + JK1

S = E + JK2

S = JK1 + E

利用这些格式,以及格式中两个 JK 之间是否有语义块指示符,HNC 定义了四种类型的格式:标准格式、规范格式、违例格式、省略格式。

这里,只给出部分三主块句的格式,格式中“ ”代表语义块指示符:

标准格式:!0J = JK1 + E + JK2

规范格式：！11J = JK1 + JK2 + E

！12J = JK2 + JK1 + E

违例格式：！21J = JK1 + JK2 + E

省略格式：！310J = E + JK2

给每个数学表示式一个确定的格式代码（！0，！11 等），我们建立了句类格式代码表，格式代码表给出了语句格式的完备性描述。

对一种语言来说，每个句类的标准格式是唯一的，通过标准格式可以对应出句类格式中各个 JK 代表的角色，然后将这些角色代入句类格式，就将一个抽象的数学表示式映射到一个具体的物理表示式上。比如：

张三 把 李四的腿 打断 了。

A l0 B X hv

由‘打断’我们知道这个例子是作用句，它的句类格式是 !11J = JK1 + JK2 + E，而作用句的标准格式为 A + X + B，所以我们就得到这个例子的物理表示式是 A + B + X。

由于 HNC 发现的 7 个基本句类和 57 个一级子类，完备地描述了自然语言语句的句类标准格式，而句类格式代码表又从数学上抽象地给出了句类格式的所有可能形式，所以，自然语言语句的任意一个物理表示式都可以用句类标准格式加句类格式代码来描述。也就是说，用句类代码加句类格式代码，可以给出语句联想脉络形式表示的完备性描述。这个结论就是句类假设的理论依据。

句类假设的任务就是要从结构表示式中得到句类代码和格式代码。它要依据假设类型对‘E 团块’内所有可能都作出假设，假设的合理性将由句类检验来判定。下面我们只对在一个句类代码下，语句的格式代码的假设作简要说明。

从数学表示式中可以看出，语义块的排列顺序主要是 E 语义块和广义对象语义块的顺序，而汉语的广义对象语义块又有带语义块指示符和不带语义块指示符两种，所以，语义块指示符和 E 语义块在格式中起着关键性的作用。根据语义块指示符和 E 语义块的相对位置以及语义块指示符的概念类别，我们就可以确定语句的格式。比如，对一个三主块句的结构表示式 FJ = (JK) + (JK) + E，如果第二个语义块 JK 的指示符是 l02（对象语义块指示符），又由于它的 E 语义块在最后，所以它的格式代码就是 !11 是规范格式。

语义块指示符在感知过程中就已经得到，所以简单句的格式我们不难分析。复杂类型的语句格式分析相对来说不容易把握，它们的个性化都比较强，必须针对不同的复杂类型，进行不同的格式分析。从图 3.1，可以看出，对不同的复杂类型的分析是和句类假设同时完成的。

下面仅以广义对象语义块分离为例作简单说明。

3.2 节已经指出，广义对象语义块分离在结构表示式上的表现就是 JK 个数多出一个，对三主块句，一般是 JK2 发生分离，四主块句，是 JK3 分离。根据这一结论，和简单句格式分析的理论，我们可以对广义对象语义块分离的格式作以下分析。

三主块句的结构表示式是 FJ = (JK) + (JK) + E + (JK),第二个 JK 前有语义块指示符,则认为它是广义对象语义块分离。如果这个语义块指示符是 l02,则认为是 !11 格式,如果是 l01,则认为是 !12 格式。它们的数学表示式如下:

JK1 + l02 + JK2Q + E + JK2H !11

JK2Q + l01 + JK1 + E + JK2H !12

四主块句广义对象语义块分离的结构表示式是 FJ = (JK) + (JK) + (JK) + E + (JK),它的数学表示式有下面几种可能:

JK3Q + l01 + JK1 + l02 + JK2 + E + JK3H !135

JK1 + l02 + JK2 + l03 + JK3Q + E + JK3H !115

JK2 + l01 + JK1 + l03 + JK3Q + E + JK3H !125

JK1 + l03 + JK3Q + l02 + JK2 + E + JK3H !116

上面我们主要讨论了语句格式代码的分析,必须强调的是,这些格式的变化都是针对广义作用句而言的,广义效应句一般都只采用标准格式。广义效应句包括效应句、状态句、过程句、基本判断句。广义作用句包括作用句、转移句、关系句等。

句类代码和格式代码的确定,就可以唯一的确定语句的实际物理表示式,从而对句类的假设也就完成。假设的结果是多个可能的语句物理表示式,到底哪一个合理呢?这就是句类检验的任务。

3.4 句类检验

句类检验就是运用句类知识对物理表示式的合理性作出判断,并根据判断结果,得到合理的语句的实际物理表示式。下面先介绍一下对一个假设结果的检验,多个检验以此类推。

“句类知识有四方面的内容:一是句类格式知识,二是语义块构成知识,三是语义块之间的关联性知识,四是语义块和句类的转换知识”。其中,第四项知识属于语言更深层次的理解,这里暂时不涉及。第二项语义块构成知识,将在下一节作详细的讨论。

语义块之间的关联性知识,主要是语义块核心之间的关联性。语义块核心我们又称其为要素。所以对语义块关联性知识的运用就是对物理表示式的各个语义块进行要素关联性的计算,也就是要素检验。

HNC 的主语义块有四种:特征语义块 E、作用者语义块 A、对象语义块 B 和内容语义块 C,对某些句类如关系句又有 RB1 和 RB2。其中,特征语义块是核心,它决定其他语义块的概念优先性,所以检验是以 E 为主体进行的。同时,有的句类如状态句,它的 BC 之间也有较强的关联性,也必须对它们作检验。由此,我们得到要素检验的内容:

E-A 检验

E-B,B1-B2 检验

E-C 检验

B-C 检验

这几项检验的成功与否,将直接影响语句合理性的判定。

在一个语义块内部,它的要素位置如何确定呢?汉语有一个极为可爱的特点,它的广义对象语义块 JK 的核心,一定在语义块的最后。这就给要素检验带来了极大的方便。我们一般只要在语义块末尾进行要素检验即可。

HNC 的要素检验是分层次进行的,这是 HNC 的知识表示和处理策略的必然要求。首先,必须区分具体和抽象概念,不同句类的不同语义块所要求的概念大类也可能不同,所以先根据对两大类概念作检验显然符合效率优先的原则。其次,应该区分人、物、事。最后,再对层次网络符号进行全面的检验。这三个层次的检验结果的加权,就可以给出要素检验的评分。

要素检验的基本手段是语义距离计算。语义距离是对概念之间关联性强弱的定量表述。概念关联性或语义距离的概念,"在某种意义上是对传统的词性约束规则的扩展和深化。扩展表现在它试图表述语义块之间或语句要素之间的约束,深化表现在它试图尽可能给出条件"。语义距离计算就是计算概念之间的相关系数。

HNC 的概念表述具有层次网络的特性,有网络符号,主要用网络字母表达,还有层次符号,主要用数字表达。这些符号逐层比较,相同为 1,相异为 0,将每层的比较结果相加就是概念的相关性。语义距离计算就是对这一比较过程的实现。

概念之间的关联性需要通过多重层面予以表达,有概念层面的关联性,有词汇层面的关联性,有语法层面的关联性,有语义块内部的关联性,有语义块之间的关联性。不同层面相关系数的量化和计算方法都应该有所不同。这是语义距离计算的基本特点。

相关函数是一个条件概率,语义距离的条件性更为突出,在某种意义上,条件的把握是计算语义距离的关键。条件则通过交式关联、"同行优先"准则和句类知识三条途径来表述,前两条实际上就是词性匹配的具体条件,第三条是运用链式关联知识的条件。

从上面的说明可知,语义距离的计算首先要区分语义块内部和语义块之间两种情况。要素检验主要是对语义块之间的关联性作语义距离计算。下一节的语义块构成将进行语义块内部的语义距离计算。

要素检验主要是运用:句类知识:概念关联性知识库中的上列各种链式关联知识。

语义距离计算可以得到各语义块之间的关联性评分,但评分不是目的而是手段,目的是作出语句合理性判断,也就是下面的语句合理性分析。要作出合理性判断,需要评价函数,而评价函数并不等同于评分,这里就需要利用关联性评分对语句的物理表示式给出评价,得到句类检验评价函数。这就是句类检验的主要任务。

句类检验评价函数,分简单句类和复合句类两种。这里我们只讨论简单句类的评价函数。

首先,简单句类可能有各种复杂构成,比如块扩、句类转换、句蜕、广义对象语义块分离等。句类检验应该对它们作出评价。这些复杂构成与句类密切相关,某些句类或者某些 v 概念,就不允许出现复杂构成。比如,一般承受句

X101J = X1B + X10C + XBC

它的 X10C 语义块一定不分离,如果在当前语句的物理表示式中,出现 X10C 的分离现象,则必须给出否定的判断。

其次,对物理表示式的格式给出判断。各个句类所允许的语句格式也有所不同。这在知识库中都有明确的指示,比如广义效应句只允许标准格式,不可能出现规范格式。这些必须在评价函数中得到体现。

最后,根据上面的语义块要素检验的评分,对语句的关联性作综合评价。这个综合评价,不是语义块要素检验评分的加权平均,它是依据句类知识,根据各个语义块对句类的依赖程度,作出的综合判断。比如,转移句的 TB,它要求必须是具体的地点,如果 TB 的要素不是地点,则应该马上给出否定的判断。

有了句类检验评价函数,我们就可以进行语句合理性分析。合理性需要从正反两方面来考察,即‘合理’和‘不合理”。“合理’联系于漏报,“不合理’联系于虚警。合理性分析主要是根据句类检验评价函数和下一节要介绍的语义块构成评价函数,对语句合理性作整体性的判断,必要时,需要作回溯处理。依据合理性判断结果,我们就可以对假设的多个物理表示式作出选择,得到合理的表示式。这也就完成了句类检验的任务。

通过句类检验,就可以将一个“E 团块’的模糊消解,从而将 E 语义块精确定位。至此,就将多个 v 概念的困扰彻底消除了。

3.5 语义块构成

语义块构成分析实质上是上面句类分析的一部分,但它主要是对广义对象语义块内部的局部处理。

语义块在形式上可以是一个词、一个短语或者一个句子。这里,对它的一般构成作简要说明。语义块各构成成分的形式化定义为:

语义块	K
核心部分	KH
说明部分	KQ
前缀部分	QK
后缀部分	HK
语义块各部分	FK(KH,KQ,QK,HK 的代表)
语义块函数	K(jmn),K(fmn)
必须扩展成语句的语义块	【K】
不能扩展成语句的语义块	□k□

这里,语义块函数的第一个变量 j 和 f 分别表示基本概念和语法概念。例如,K(j0) K(j1) 分别表示序描述和时间描述语义块,其他类推。K(f30) 表示名称语义块,K(f42) 表示疑问语义块,其他类推。

语义块的形式描述如下：

$$K = KQ + KH \qquad (K.01)$$

$$K = QK + (KQ + KH) + HK \qquad (K.02)$$

$$< HNC > = K(\cdots)$$

$$FK = (HNC)$$

式中，< HNC > 表示语义块的 HNC 命名，即语义块语义符号。K(…)表示语义块函数，(HNC)表示语义的 HNC 表示，即层次网络符号表示。最后的表示式 FK = (HNC)就是语义块构成的 HNC 表示。

下面对基本概念的时间语义块作形式化描述。

宏观特定时间表示：

$$(fpj1 ; pj1) + \sum(\sum j308 + wj10-)$$

如：康熙三十一年，公元 1996 年 10 月 29 日，10 月 29 日

微观特定时间表示

$$wj10\text{-}00c + (\sum j308 + wj10\text{-}000) + (\sum j308 + jzz12\text{-}0)\ ;$$

$$wj10\text{-}00c + (\sum j308 + wj10\text{-}000) + jzu41$$

如：下午 3 点 23 分

上午 8 点左右

像时间描述语义块这样，可以写出明确表示式的语义块叫作 WD(well-defined 的简写)语义块。WD 语义块的构成分析相对比较简单，我们对它采用规则的形式进行处理，规则就是它的语义块表示式。

对一般语义块，如果它没有明确的表示式，怎样进行构成分析呢？回答是从“对象、表现”的二分法着手。一个语义块不管它形式上有多么复杂，它总是由对象和表现两个方面构成的。HNC 认为语义块内部的对象和表现可以无限制的嵌套。如对语义块 XBC，它的构成可形式化描述为：

$$\begin{aligned} XBC &= XBCB + XBCC \\ &= XBCBB + XBCBC + XBCC \\ &= XBCB + XBCCB + XBCCC \\ &= XBCB + XBCCB + XBCCCB + XBCCCC \\ &\cdots\cdots \end{aligned}$$

根据“对象内容”的划分，广义对象语义块的构成有良性和非良性之分，良性语义块的对象 B 和内容 C 可以明显地分开，且排列顺序固定，如作用句的对象语义块 B = XB + YB + YC，XB 为作用对象，YB 为效应对象，YC 为效应内容，这三项顺序不容颠倒。非良性语义块的对象和内容无确定排列顺序，例如反应句的 XBC 的反应引发者 XB 及其表现 XC 在多数情况下排列不能事先确定。它可以是 XBC = XB + XC，也可以是 XBC = XC + XB；XBCB + XBCC；XBCC + XBCB。

语义块构成的主要操作是语义距离计算。语义块内部语义距离的计算主要是运用“同行优先”准则,概念关联性知识库中“交式关联”知识。

所谓“同行优先”准则,是对层次网络符号天然属性的一种简明陈述,正式的陈述是:同行的五元组概念及挂靠的(w,p)类概念优先相互搭配。从应用的角度来看,这不过是用数字符号表达概念关联性的一个简单技巧。在具体应用这一准则于语义距离计算时,要区分四种不同的搭配方式,因为每种搭配方式各有自己的约束准则。四种搭配方式是:修饰型搭配,补充型搭配,并合型搭配,对象内容型搭配。这四种方式涵盖了一般语义块所有可能的搭配形式,对它们的区别对待,是语义块构成分析的指导原则。

面对模糊文本,语义块构成分析也是解模糊的过程,在本系统中,就是将语义块内部的拼音流转换成确定的汉字流。

3.6 小结

第 3 节详细介绍了句类假设检验的理论和处理策略。句类假设检验是建立在语句的物理表示式之上的。对语句物理表示式的完备描述,是 HNC 理论最大的贡献,也是自然语言理解的重大突破。

句类假设检验以语句物理表示式为出发点,以句类知识为依托,从语义深层,对语句的概念关联性作出合理性判断,并对模糊文本进行解模糊处理。

语义块感知以及句类假设的实现解决了长期困扰自然语言处理的语义层面的句子理解处理,给语句理解一个最终解决方案。它对长期困扰汉语的单音节处理、述语动词的发现、分词等问题都作出了肯定的回答。

4 智能调度

前面我们介绍了 HNC 的语义块感知处理和句类假设检验处理,它们在系统中不是线性的关系,它们之间是互相渗透,互相支持的。对它们的处理必须“相机行事”,如何“相机行事”?答案就是智能调度。

4.1 智能调度

第 1 节已经指出,智能调度的本质是数据驱动,HNC 数据驱动的本质是区分概念激活信息的强弱,对强激活予以优先响应。优先响应并不是作出肯定的选择,而否定其他的激活,如果随着分析进程的不断深入,激活信息的强弱发生变化,这时也可能否定以前的优先响应,而重新作出选择。可以说,HNC 的假设检验的过程就是对优先信息的不断选择、排序的过程。分段层选处理、lv 序列处理、语义块感知处理,都只是优先级的判断过程,它们只是把优先考虑的组合及其有关的信息提供给句类分析,最终的确认则有待于句类分析之后。“优先”贯穿于 HNC 处理的全过程,但优先只是处理的策略和手段,我们的目的是对句子进

行多义选一处理,如何使‘优先’的发展朝着我们的目标前进?这就是智能调度的中心任务。

由于 HNC 句类分析系统是在语句一级的理解处理,智能调度首先必须对音串的结束符作出响应。音串的结束符包括‘,。!?’等,它们一般表达语句的语气,是陈述句、疑问句、祈使句等。HNC 的语句物理表示式和语句数学表示式都是针对陈述句的,如果要对疑问句或祈使句处理,必须对这些表示式作相应的语义块变换。所以智能调度在得到音串结束符,语义块感知过程结束后,要根据结束符的类型,进行相应的变换处理,然后,再进入句类假设检验。

智能调度的关键是句类知识的运用。句类知识的得到与否是智能调度的分水岭,没有得到句类知识之前,调度的主要任务是辨识句类,分段层选和语义块切分组合的优先级都取决于它是否有利于句类的辨识。当辨识出句类后,调度的任务就转移到利用句类知识,作优先级的调整,并对后面的假设检验作出预测。下面我们就对智能调度的这两个阶段分别作详细的阐述。

首先,对没有得到句类知识,进行句类辨识时的调度作详细介绍。分段层选和语义块切分组合的优先级主要是由它们所含的激活信息的强弱决定的。我们以一个例子来说明智能调度如何使优先级向着有利于句类辨识的方向发展。

zhong guo | d | jing ji yi jing | hui fu dao li shi | zui hao | shui ping .

(中国的经济已经恢复到历史最好水平。)

(注 这里‘的’是本系统的一个指定输入,……|……|……是音段的划分。)

以上这个例子中,有一个五音段‘hui fu dao li shi”,音段里有‘恢复、辅导、道理、历史’四个双音词。这个五音段有三个层选结果:

层选 1　“hui | fu dao | li shi”(hui 辅导 历史)

层选 2　“hui fu | dao | li shi”(恢复 dao 历史)

层选 3　“hui fu | dao li | shi”(恢复 道理 shi)

层选 1 中,有两个激活信息,“hui’可以作 E 的情态修饰 QE,“辅导’可以假设 E。层选 2 也有两个激活信息,“恢复’可进行 E 假设,“dao’可以是 E 的趋向后缀 hv。层选 3 只有一个激活信息,“恢复’可进行 E 假设。层选 3 的优先级最容易判断,它的激活信息最弱,而层选 1 和层选 2,如果作 E 假设,都有相应的 QE 或者 hv 支持。它们的激活信息强弱相当,这时判断就比较复杂,智能调度必须进行跨段处理。在前一个音段“jing ji yi jing”内,激活信息“已经’是 E 的时态修饰 QE,它对后面的‘辅导’和‘恢复’都是加强的,所以这里单纯利用激活信息还是无法判断层选 1 和层选 2 的优先级,智能调度有什么高招吗?这时必须依靠句类知识,进行简单的句类检验。

这里,简单的句类检验,就是对两个 E 假设的广义对象语义块作高层的要素检验,即具体概念和抽象概念的检验。层选 1 和层选 2 的两个 E 假设‘恢复’‘辅导”,它们的 JK1 的边界都是‘jing ji”,汉语的广义对象语义块的要素一定在末尾,也就是两个 E 假设的要素都是“jing ji”。“辅导’的作用者语义块优先于具体概念,而‘恢复’则是具体概念和抽象概念都可

以“jing ji”(经济)是一个抽象概念,根据语义距离计算的结果,我们可以知道“恢复”和“经济”的关联性更强一些。由此可得,层选 2 的优先级比层选 1 高。这样智能调度就完成了层选的选择,同时也辨识出句类,完成了语义块感知的主要工作,同时也进行了初步的句类检验。

另外,句类的优选还要依靠句类及其格式宏观先验知识的运用和(E,l)联合感知。句类格式是句类知识的纲,对它们的应用,将大大提高优选的效率。HNC 定义的四种句类格式都有它自己的特点。对汉语来说,标准格式的 E 语义块都在第二位置,三主块句规范格式的 E 块一定在句子末尾,省略格式 !310 的 E 块在句首。汉语的 l0 语义块指示符使用比较频繁,使用 l0 的句类只能是广义作用型句类,l0 的出现必然意味着规范格式。这些都是进行优先级判定时的重要依据。

感知到句类后的智能调度如何动作呢?由于句类已经感知到,也就是抓住了句子的全局联想脉络,这时,调度的中心是利用句类知识对后面的感知进行预测,并判断各种复杂构成。一般情况下,E 语义块后不允许出现语义块标志符,如果出现,则它的优先级将很低,仅仅留待以后回溯处理时使用。如果 E 语义块后又出现另外的 v 概念,这时在假设时必须依据句类知识判断是否可以形成块扩、句蜕、复合等复杂构成。比如下面的例子:

wo men bu neng wang ji zhi min zhe tu sha wo guo tong bao d zui xing .

(我们不能忘记殖民者屠杀我国同胞的罪行。)

这个例子的拼音流输入时,我们首先对“忘记”作出 E 假设,当“屠杀”出现时,由于它是一个 v 概念,也需要作 E 假设,智能调度这时采取观望的态度,保留这两个 E 假设。当拼音流输入结束时(就是遇到“。”时),智能调度对这两个 E 假设作句类检验,“忘记”是反应效应句 X2Y * 1,它引导的句类的物理表示式是 EJ = YB + X2Y + XBC,它的第二个语义块 XBC 天然可以块扩或句蜕,所以我们优先考虑对第二个 E 假设“屠杀”作块扩、句蜕检验,检验结果表示,“殖民者屠杀我国同胞”这部分拼音流满足“屠杀”这个 E 假设句类知识的要求,而在拼音流中,“的”的出现又将它们作为整句 E 的可能给否定了,由此,智能调度作出“殖民者屠杀我国同胞的罪行”是句蜕的判断,它们作为“忘记”的一个语义块而存在,整句的 E 是“忘记”。这样,智能调度就依靠“屠杀”和“忘记”的句类知识完成了多个 E 假设的判定问题。

句类知识是智能调度的主要依据,但它不是智能调度的全部。智能调度还需要利用各种“亮点”信息。本系统设置的七个指定字就是“亮点”的一种。设置七个指定字的根本目标就是实现智能调度。

七个指定字“的、了、和、是、有、不、在”都对 E 语义块感知起着关键性的作用。它们的这一特点,给智能调度的实现带来了很大的方便。

我们对七个指定字对 E 块感知的影响作简要说明。

“的”的作用在于先取消其先后 v 概念的 E 资格。

“了”是 E 已经发现的最强的标志。

“在”是条件辅块 Cn 常用的指示符,同时也是很强的 hv 和 E 假设。

“是、有”是基本判断句的 E,它们一般没有规范格式,也没有 E-B ,E-C 的关联性。

“不”作为最常用的否定符号,是 E 要素即将出现的可靠标志。

“和”对其前后的概念有对仗性要求,有助于对并合型 E 块的发现。

总之,句类知识和“亮点”信息是智能调度的关键,也是智能调度的主要数据,智能调度的数据驱动,就是指这些知识的利用,并根据它们作出有效的判断,以完成句类分析的综合调度。

4.2 K 调度

4.1 节已经指出,智能调度的关键是句类知识的运用,也就是说,必须假设 E 的存在,这里的 E 一般是双音词。因为,在语义块感知阶段,我们必须先暂时回避单音动词,否则将陷入“草木皆兵”的困境。那么,如果句子中确实采用了单音动词,或者句子中没有动词时,系统如何动作呢?HNC 的策略是 K 调度。

一个没有双音 E 要素的音串(串以逗号为终止标记),它可以有两种情况,一种是有单音动词,一种是没有单音动词。

如果语句确实采用了单音动词,则可能有两种情况,1.未发现 E 块;2.伪双音动词干扰。对情况 1,可以直接转入 K 调度来寻找单音 E 块;对情况 2,则需要通过句类检验排除伪词干扰后,才能转入 K 调度。这就是 K 调度的两个入口。

如果句子中没有单音 E 要素,它可以是一个插入语 fFK,一个辅块 fK,一个广义对象语义块 JK,或者是一个无 E 块的 S04 句类。因此,K 调度的操作要在排除这四种可能之后,才能进入单音 E 要素的感知。

K 调度实质上是智能调度的一部分,它也坚持调度的原则“数据驱动”。进入 K 调度后,首先必须要根据数据黑板的数据生成一个调度策略。具体来说,如果黑板内有 fFK、fK、JK 的记录,则直接进入相应的检验,这里的 JK 一般是句蜕块;如果句子末尾是 vu、u 类概念,则优先判定它为 S04,并做句类检验;最后,才进入单音 E 的感知。

对单音 E 的感知,首先必须确定单音的位置。汉语中单音词的模糊都很大,确定位置应该根据其他的旁侧信息。一般,单音 E 出现时,它的后面都有 E 块的后缀 hv 或者 E 前缀 QE,它们是单音 E 发现的重要依据。汉语的基本概念组成的单音 E 一般是单独出现,如“占、加……”等。对出现这些单音的位置都必须敏感。

单音 E 的位置假设出后,还必须对这个单音词作多义选一处理,从多个汉字中选择出一个合适的单字词。这时,我们主要利用单字的句类知识,如它的要素关联性知识,基本句类知识等。对每个字作句类检验,得到一个检验合理的结果。

4.3 数据组织

智能调度的本质是数据驱动,那么,如何组织数据,才能使智能调度顺利进行呢?本系

统采用数据黑板的形式组织数据。数据黑板就是一个共享的数据集,它记录语义块感知和句类分析过程所有的数据,这些数据随着处理的不断深入而不断更新,智能调度主要依靠黑板的当前数据状态决定执行那一部分功能。

图 4.1 给出数据黑板的主要结构图。

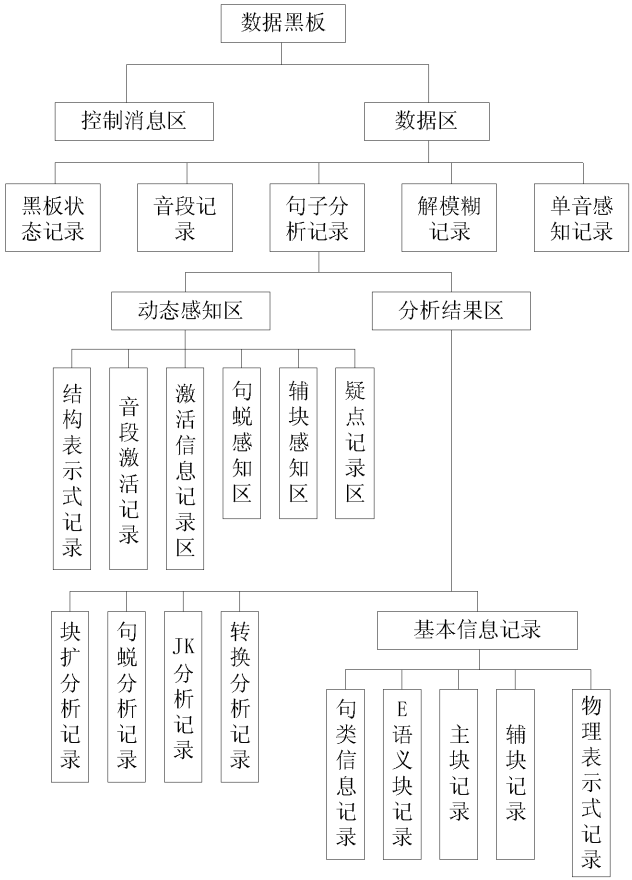


图 4.1

数据黑板首先分成数据区和控制区。控制区负责记录输入消息和输出消息,这里的输入、输出是针对数据黑板而言的。输入消息是当前执行的动作返回给黑板的,它提示当前动作完成,并对数据黑板刷新。输出的消息是根据输入消息和当前黑板状态综合而来的,它提示智能调度下一步要执行的动作。整个系统就是由这些输入、输出消息来驱动的,消息是数据的综合。

数据区的设计策略是分层组织。

首先,按照数据类型的不同来组织。我们设计了单音词感知记录区、解模糊记录区、状态记录区、音段记录区和句子分析记录区。这几个区记录的数据对象各不相同,这样设计体

现了数据封装的原则,与一种数据对象无关的操作将无权存取这种数据对象的记录区。

其次,按照分析过程进行分层组织。在句子分析区内,我们设计了动态感知区和分析结果区。这两个区分别对应于语义块感知和句类分析。动态感知区是在拼音流输入过程中,不断更新的,它的核心是结构表示式,同时它还记录了疑点,为回溯处理作好准备。当拼音流结束时(以‘,’或‘。’结束),动态感知区最后一次刷新,给出语义块感知的结果,提供给句类分析。分析结果区是在拼音流输入结束后,根据语义块感知的结果,进行句类分析和语义块构成分析,得到语句的物理表示式和解模糊结果。

我们可以看出,数据黑板的设计充分考虑了智能调度的使用,它的分层组织的设计策略是智能调度的要求,也是 HNC“中间切入,先上后下”处理策略的要求。

4.4 小结

第 4 节详细说明了 HNC 处理器的智能调度策略以及调度的特殊处理 K 调度。智能调度的本质是数据驱动,它主要的依据是句类知识和‘亮点’信息。K 调度的关键是作出单音 E 假设,单音 E 的假设要依靠 E 语义块的前后缀 QE 和 hv 的信息。数据组织必须为智能调度服务,本系统采用数据黑板的组织形式,设计时充分体现了分层组织的思想。

5 结 束 语

本文依据 HNC 理论,设计了 HNC 的语义块感知和句类辨识算法以及句类假设检验算法,并以此实现了 HNC 句类分析系统。

HNC 句类分析系统第一个目标是针对汉语设计,它实现了以下九项技术突破:

1. 语义块扩展成语句的处理(块扩处理);
2. 语句蜕化成语义块处理(句蜕处理);
3. 复合句类处理(多述语动词处理);
4. 语句合理性和自足性检验;
5. 句类转换处理;
6. 语义块构成处理;
7. 语义块分离变换处理;
8. 语义块构成合理性和自足性检验;
9. 语义距离计算及快速局部处理。

这九项突破解决了汉语理解中的以下核心难题:1. 汉语述语动词的辨识;2. 分词‘瓶颈’;3. 词性兼类;4. 单音词辨识及其模糊消解;5. 新词辨识及伪词分解。由此,HNC 句类分析系统已经取得了语句理解的突破。

HNC 句类分析系统已显示出令人惊喜的理解能力和巨大的潜力,它将随着以下两方面的发展而不断成长:一是知识库的规模和水平;二是 HNC 核心技术的软件实现的进程。

HNC 理解技术有三大法宝：一是语句层面的句类知识；二是句群层面的句间知识；三是篇章层面的语境知识。目前我们只利用了第一项法宝，我们相信，随着后两项知识的运用，HNC 句类分析系统的理解能力必将达到一个新台阶，将在消解模糊方面接近或达到人的智能。

HNC 的分析处理现在处在初级阶段，我们还有很多工作要做。首先，要完善语句层面的句类分析，使语义块感知和句类分析的能力更加强大。这需要 HNC 联合攻关组的共同努力。其次，要进入句群和篇章层面，使理解向深层次方向发展。随着技术的不断完善，HNC 技术将使机器翻译、电话翻译、人机对话、智能检索、自动文摘等语言处理的各个领域获得实质性的重大进展，为我国创新语言信息产业带来曙光。

参 考 文 献

[1] 黄曾阳. HNC 理论概要. 中文信息学报. 1998，11（4）

[2] 黄曾阳. 自然语言理解处理的 52 个论题. 1998. 6

[3] 黄曾阳. HNC 理解处理论文选录. 中国科学院声学研究所声场声信息国家重点实验室自然语言理解课题组，1996.3

[4] 黄曾阳. 语义学日记. 1994

[5] 刘志文等. 自然语言语句的 HNC 表示. 语言文字应用，1998（2）

[6] 林杏光. 正确引导汉语理解与汉语研究——事关人工智能开发的一个重要前提. 科技导报，1997（4）

[7] 林杏光. 中文信息界语义研究谈要. 语言文字应用，1998（3）

[8] 林杏光. 张志公先生 90 年代汉语语法观. 张志公先生逝世周年追思学术讨论会论文集. 北京：人民教育出版社，1998

[9] 林杏光等. 现代汉语动词大词典. 北京：语言学院出版社，1994

[10] 张全. 基于 HNC 理论的语义块感知处理. 中国科学院声学所博士学位论文，1996. 7

[11] 申凌. 服务于语音识别的汉语理解系统. 中国科学院声学所硕士学位论文，1996. 7

[12] 苗传江. HNC 理论的基本内容. 中科院声学所‘HNC 知识库培训班’教材，1997，3

[13] 季宏. 汉语自然语言理解系统——在语言识别和文语转换中的应用. 中国科学院声学所硕士学位论文，1995，7

[14] 张普. 汉语语言信息处理与研究. 见：（日）西稹光正编. 语境研究论文集. 北京：语言学院出版社，1992

[15] 黄昌宁，夏莹. 语言信息处理专论. 北京：清华大学出版社，1996

[16] 陈力为，袁琦等. 中国中文信息处理平台 905 工程，1995

[17] 黄昌宁. 研制汉语句法分析器的对策. 计算机开发与应用，1989，5（2）

[18] 孙茂松. 汉语句法分析中的一种多扫描确定性算法及其在篇章理解中的应用. 清华大学硕士论文. 1998. 12

[19] 姚天顺. 自然语言理解——一种让机器懂得人类语言的研究. 北京：清华大学出版社，1995

[20] 王苏，汪安圣. 认知心理学. 北京：北京大学出版社，1992

[21] 吴蔚天，罗建林. 汉语计算语言学——汉语形式语法和形式分析. 北京：电子工业出版社，1994

[22] 初敏. 高清晰度高自然度汉语文语转换系统的研究. 中国科学院声学所博士学位论文, 1995.9
[23] 朱德熙. 语法问答. 语文出版社, 1985
[24] 傅雨贤. 现代汉语语法学. 广东高等教育出版社,1996
[25] 蔡自兴,徐光佑. 人工智能及其应用. 北京 :清华大学出版社, 1996
[26] 石纯一,黄昌宁等. 人工智能原理. 北京 :清华大学出版社, 1993
[27] 陆汝钤. 人工智能. 北京 :科学出版社, 1987
[28] Chomsky N. Syntactic Structures. Hague : Mouton,1957
[29] Chomsky N. Aspects of the Theory of Syntax. MIT Press,1965
[30] Fillmore C J. The case for case. In : Bach E and Harms R eds. Universals in Linguistic Theory. New York : Holt, Rinehart and Winston,1968
[31] Qullian M R. Semantic Memory. In : Minsky M ed. Semantic Information Processing. Cambridge, MA : MIT Press,1968
[32] Schank R. Identification of conceptualizations underlying natural language. In : Schank R and Colby K Eds. Computer Models of Thought and Language. San Francisco, CA : W H Freeman and Company,1973
[33] Schank R. Conceptual Information Processing. Amsterdam : North Holland,1975
[34] Schank R. The structure of episodes in memory. In : Bobrow D, Collins A eds. Representation and Understanding. New York : Academic Press, 1975
[35] Schank R. Dynamic Memory. New York : Cambridge University Press, 1982
[36] Schank R, Abelson R. Scripts, Plans, Goals and Understanding. Hillsdale, NJ : Erlbaum, 1977
[37] Bookman L A. Trajectories through Knowledge Space : A Dynamic Framework for Machine Comprehension. Boston : Kluwer Academic Publ. 1994
[38] Alvarado S J. Understanding Editorial Text : A Computer Model of Argument Comprehension, 1990
[39] Lenat D B. CYC : A Large Scale Investment in Knowledge Infrastructure. Communications of the ACM, 1995,38 (11)
[40] Miller G A. WordNet : A Lexical Database for English. Communications of the ACM,1995,38 (11)

HNC 理论的句类*

苗传江

（北京语言文化大学语言信息处理研究所，北京 100083）

1 句类在 HNC 理论中的重要地位

首先需要特别说明，HNC 理论的句类是语句的语义分类，与语言学中一般所说的句类是同名异实的，后者是指陈述句、疑问句、祈使句和感叹句等，基本上是语句的语用分类。

HNC 理论是中国科学院声学研究所研究员黄曾阳先生创立的面向整个自然语言理解的理论框架，它以概念化、层次化、网络化的语义表达为基础，所以简称为 HNC（Hierarchical Network of Concepts，概念层次网络）理论。该理论赢得了自然语言理解的新进展[5,9]。

对“自然语言理解”的“理解”，多年来人工智能领域内的认识有所不同，但大多数人基本上都接受图灵的计算机智能标准。图灵标准是对大脑的全面模拟，是个终极目标，不可能一步实现，这是共识。黄曾阳先生在这一认识的基础上认为，自然语言理解必须分阶段进行，现阶段的首要目标应当是“消解模糊”。自然语言中存在着“五重模糊”，即发音模糊、音词转换模糊、词的多义模糊、语义块构成的分合模糊、指代冗缺模糊（这是对语音流而言的，文字流只有后三重模糊），对这些模糊的消解是人脑和计算机理解自然语言的首要任务，是自然语言理解万里征程的第一步。

为达到消解模糊的目的，HNC 理论创立了词汇的概念表述模式和语句的语义表述模式，开辟了一条全新的语句理解的技术路线，那就是“中间切入，先上后下”，即，从语义块感知和句类辨识入手，靠句类分析消解模糊。例如，当输入的是无声调的拼音串时，面对一个存在着大量模糊的语句，首先不是分词，而只是分段，也就是根据不能成词的两个音之间的断点把拼音串切分成若干段。随后，开始发现“v”（动态）概念并把它假设为特征语义块，据此判定整个语句的类别，即句类（这就是语义块感知和句类辨识）。再后，在句类知识的指导下进行语句合理性检验（这就是句类分析）。如果检验成功，则语句理解正确，语句模糊即可

* ① 本文被选入 1998 年 11 月北京召开的“'98 中文信息处理国际会议”论文集；② 本文承国家“九五”重点攻关项目“计算机中文信息处理技术及产品开发”之专题“HNC 汉语理解系统的核心技术”的资助。

消解 如果检验失败,则再做另外的假设和检验,直至检验成功。

句类分析在消解模糊方面理论上能够达到甚至超过常人的水准,使计算机迈上了语言理解的第一个台阶。在句类分析过程中,句类知识起着全局性的指导作用,是消解模糊的最有力武器。句类知识包括四个方面的内容:句类格式知识;语义块构成知识;语义块之间的概念关联知识;语义块和句类的转换知识。有了句类知识的指导,句类分析就是高屋建瓴的语言处理策略了。

由上可见,句类在 HNC 理论中占有极其重要的地位。

2 HNC 理论的句类划分标准

HNC 理论把语句的核心叫做特征语义块,如"张三打断了李四的腿"中,"打断"就是特征语义块。HNC 理论根据特征语义块的概念来划分句类,即什么概念的特征语义块就决定是什么样的句类。对特征语义块的概念分类和对语句的分类采用同一个标准,这是 HNC 理论句类划分的一大特色。那么,对特征语义块的概念进行分类的标准是什么呢?是黄曾阳先生独创的作用效应链加上作为人类思维活动的基本内容的"判断"。

什么是作用效应链?"作用效应链反映一切事物的最大共性,作用存在于一切事物的内部和相互之间,作用必然产生某种效应,在达到最终效应之前,必然伴随着某种过程或转移,在达到最终效应之后,必然出现新的关系或状态。过程、转移、关系和状态也是效应的一种表现形式。新的效应又会引发新的作用,如此循环往复,以至无穷,这就是宇宙间一切事物存在和发展的基本法则,也是语言表达和概念推理的基本法则。"[1] 作用、过程、转移、效应、关系和状态构成作用效应链的 6 个环节。其中,作用是源头,效应是结果。作用效应链是一个语义网络,每个环节下都有其子网络。

语句的语义由"v"概念表示,这与美国计算语言学家山克的概念从属理论(Conceptual Dependency Theory)是一致的。可惜山克只主要考虑了"转移"类概念,他没有找到描述自然语言中"v"概念的完备集合,而 HNC 理论的作用效应链形成了这样的完备集合,完整地提出了"作用—过程—转移—效应—关系—状态"等 6 个环节。自然语言的主要内容就是对作用效应链的 6 个环节进行局部和总体的具体表述,作用效应链揭示了语言表达的深层要素,形成了对自然语言进行总体表述的完整体系。

用作用效应链加上"判断"来给特征语义块分类,也就是对语句进行分类,形成 HNC 理论的 7 大基本句类:作用句、过程句、转移句、效应句、关系句、状态句、判断句。

3 HNC 理论的句类系统

HNC 理论的句类系统是有层次的。第一个层次就是上述 7 个基本句类,分别举例如下:

作用句:张三打断了李四的腿。|奥委会取消了他的参赛资格。

过程句:李四的腿伤大有好转。|会议刚刚开始。|他父亲去世了。

转移句：张三转交给李四一封信。|张三告诉了李四这个好消息。

效应句：李四养好了腿伤。|李四的腿伤养好了。|我校的学术地位大大提高了。

关系句：张三失去了他多年的女朋友。|理论和实践要很好地结合起来。

状态句：张三穿着皮大衣。|李四正在睡觉。|主席团坐在台上。|张小姐很漂亮。

判断句：张三是学生。|张三的水平不如李四。|总理分析了当前的国际形势。

句类系统的第二个层次是基本句类下面的子类，下面是作用句和转移句的子类：

作用句子类

承受句：张三负责这项工作。|青海地区遭受了特大风雪灾害。

反应句：张先生喜欢李小姐的个性。|海峡两岸人民都非常爱戴孙中山先生。

免除句：张三摆脱了李四的纠缠。|伊拉克人民躲过了一场空难。

约束句：世界各国都禁止毒品买卖。|不合理的制度束缚了群众的思想。

转移句子类

接收句：李小姐收到一封情书。|我们要密切注视敌人的动向。

物转移句：张三转交给李四一封信。|中国赠送给菲律宾五十辆公共汽车。

物自身转移句：张三去上海了。|代表团已经抵达莫斯科。

信息转移句：张三告诉了李四这个好消息。|我国政府重申了这一立场。

交换和替代句：双方交换了场地。|张三代替了李四的职位。

|张三买了一台电脑。|本店出售优质大米。

有作用就有对作用的“承受”和“反应”；“免除”是对作用的免除；“约束”是一种特殊的作用。这些是“作用”这一概念的本质的内容和特点，也正是 HNC 给作用句划分子类的依据。有转移就有“接收”，转移有“物转移”和“信息转移”；“交换和替代”是一种具有双向性的特殊的转移。这些是“转移”这一概念的本质的内容和特点，也正是 HNC 给转移句划分子类的依据。

上面列举的是基本句类的一级子类，一级子类下面还有子类，是基本句类的二级子类。下面是反应句的子类。

反应句子类

后续反应句：张先生吓得不敢说话。

主动反应句：张三期求李四的帮助。

被动反应句：张先生害怕李小姐的脾气。

反应者会有某种行为或表现，称为“后续反应”，反应有主动、被动之分，在理解中对主动反应要追究动机，对被动反应要追究原因。这是 HNC 给反应句划分子类的依据。

基本句类可以形成混合句类和复合句类。混合句类是特征语义块的概念关涉到作用效应链的两个或多个环节的句类，例如下面的反应状态句中，“生气”和“冷静”既是反应也是状态。复合句类是一个语句中有两个或多个特征语义块的句类，例如下面的复合句中分别有“去”和“参加”、“提交”和“讨论”两个特征语义块。

混合句

反应状态句：张先生非常生气。|他总是很冷静。

作用关系句：任何国家都不能干涉别国内政。|学校开除了考试作弊的学生。

关系作用句：张三经常帮助李四。|我们支持中国加入世界贸易组织。

信息转移作用句：张三批评了李四的错误思想。|父母要多鼓励孩子。

复合句：张先生去上海参加国际会议了。|我们将把这份文件提交大会讨论。

基本句类及其子类、混合句类和复合句类构成 HNC 理论的句类系统。（受篇幅所限，上面未列出基本句类的全部子类，混合句类和复合句类也只是举例说明。）

4 HNC 的语义块和句类理论

HNC 理论在句类的基础上建立了语句的语义表述模式，这个模式的建立还需要语义块的概念。语义块和句类理论是 HNC 理论的基本内容之一。

语义块是句子的语义构成成分和单位，它不同于传统语言学的短语。语义块是语义，即语言深层的定义，而短语是语法，即语言表层的定义。语义块这一概念的提出便于描述句子的构成。用词或短语描述句子，无法清楚地界定一个句子是否完备，如果问一个句子应该或者可能有多少个词或短语，便难以回答。有了语义块的概念，就可以明确地回答一个句子有多少语义块以及每个语义块的类型等问题。

语义块分为主语义块和辅语义块两大类，前者是句义的‘必不可少’的成分，后者是句义的‘可有可无’的成分。主语义块有 4 种：特征 E、作用者 A、对象 B 和内容 C，其中特征语义块 E 决定句类。辅语义块有 7 种：方式 Ms(Means)、工具 In(Instrument)、途径 Wy(Way)、比照 Re(Refer)、条件 Cn(Condition)、因 Pr(Premise)、果 Rt(Result)。

EABC 四大主语义块划分的理论依据是：一个语句的内容无非是两个方面，一是表达对象，二是对象的表现，前者是‘什么’，后者是‘怎么样’。作用者 A、对象 B 是表达对象，特征 E、内容 C 是表现。A 是表达对象中的特殊对象，大致相当于严格意义上的施事。E 是语句的核心。对象 B 和内容 C 的划分源于对宾语的类型和构成的分析。宾语跟动词的语义关系十分复杂，过去对宾语类型的划分缺少层次性，难以说清到底有多少类，HNC 则首先把宾语区分为对象 B 和内容 C。例如，‘杀人’和‘放火’、‘练兵’和‘练武’、‘立法’和‘违法’中，前者的宾语‘人、兵、法’是对象 B，后者的宾语‘火、武、法’是内容 C。从宾语的构成来分析，比如‘增强人民体质’、‘加快国有企业的改革进程’、‘提高高校的学术水平’中的宾语都由两部分构成，这两部分之间存在着这样的差别：可以说‘增强体质’、‘加快改革进程’、‘提高学术水平’，但不能说‘增强人民’、‘加快国有企业’、‘提高高校’。具有这样的构成的宾语的前一部分是对象 B，后一部分是内容 C。

七大辅语义块是对汉语的全部语言逻辑概念（基本对应于传统的介词和连词）进行对比分析和综合归纳后得出的。

四大主语义块和七大辅语义块都是高度抽象和概括的结果，它们的具体语义内涵要根

据句类来确定,HNC 对此的精确表述是‘语义块是句类的函数’,这是 HNC 语义块和句类理论的基本论点。例如作用者 A 语义块,它在作用句中是‘产生影响者’,类似于一般所说的施事,在转移句中则是‘转移的发出者’,过程句、关系句和状态句中则不涉及 A 语义块。再如对象 B 语义块,它在作用句和效应句中是‘被影响者’或‘接受者’,类似于一般所说的受事,在过程句、关系句、状态句中则是过程、关系、状态的‘体现者或承受者’,而在转移句中则是‘转移的接收者’。

句类的主语义块按照一定的顺序排列,构成句类格式。句类格式有标准格式和非标准格式之分。标准格式与语种有关,可以定义为某种语言中常用的格式。而实际上,定义哪种格式为标准格式并没有理论意义,因为句类格式是语义层面(即语言深层)的定义,不同的格式之间是可以互相转换的。例如,下面是一般作用句和信息转移句的标准格式与非标准格式。

一般作用句

标准格式:张三打断了李四的腿。

非标准格式:张三把李四的腿打断了。|李四的腿被张三打断了。

信息转移句

标准格式:张三告诉了李四这个消息。|张三透露给了李四这个消息。

非标准格式:张三把这个消息告诉了李四。|这个消息被张三透露给了李四。

句类格式就是语句的语义构成模式。HNC 理论穷尽地发现了自然语言语句的语义模式,这就是前文所述的基本句类及其子类、混合句类和复合句类。基本句类及其子类共有 57 个,他们两两混合构成的混合句类理论上有 $57 \times 56 = 3192$ 个,而实际语言中常见的并没有这么多。

在语义块和句类的基础上 HNC 理论建立了自然语言语句的语义表述模式。基元性和完备性是 HNC 理论在对自然语言的表述中极力追求的目标,这里充分体现了这一点。四大主语义块和七大辅语义块是句子语义成分的基元,七大句类是句子语义类型的基元。语义块是句类的函数,句类及其语义块构成句类格式,它是语句的语义构成模式。57 个基本句类及其子类是自然语言语句的语义构成模式的基元,它们和混合句类、复合句类构成完备的句类系统。

5 HNC 理论的句类划分的重大意义

HNC 理论的句类划分,不论在汉语信息处理方面还是在汉语研究方面,都具有重大的意义。

对句子类型的研究是现代汉语研究的一项重要内容。《马氏文通》问世以来出版的语法著作一般都着重于句子的各个组成部分的解剖,对句子整体结构的研究大多散见于不同的章节,使读者无由纵观全局。近些年来,对现代汉语句子类型的研究已经逐步开展起来。

根据‘三个平面’的语法理论,不同平面上有不同的句子类型:根据句法平面的句法结构

的格局分出来的是‘句型’,根据语义平面的语义结构模式分出来的是‘句模’,根据语用平面的语用价值或表达用途分出来的是‘句类’(不同于 HNC 的句类!)。胡裕树先生指出:目前对句子类型的研究还只着重于句法平面的‘句型’研究;而语用平面的‘句类’研究既不全面、也不深入;至于语义平面‘句模’的研究可以说还是空白。[14]

许多汉语研究学者认识到,汉语是意合型语言,缺少形态变化,对汉语的语句应该着重在语义平面上进行分类。但如何分类,用什么标准来进行分类,一直找不到满意的途径。

美国语言学家菲尔墨的格语法问世以来,被广泛应用于自然语言处理和语言研究。句子的述语动词和名词性成分的格组合在一起构成格框架。林杏光和张庆旭在《现代汉语动词大词典》和《现代汉语述语动词机器词典》的基础上总结出了汉语常用述语动词的 58 种格框架类型,笔者也在此基础上对某些格框架进行了比较细致的研究。格框架就是句子的语义结构模式。我们在研究中认识到,研究格框架对语言信息处理具有重要的使用价值。但是,我们也发现了困难和问题,其根本点在于,从用格系统对动词的描写中归纳出的格框架总是不够系统和完备,这是由格语法不能建立一个完备的格系统决定的。格语法存在的一个大难题是语言中到底总共有多少格,不同的格如何区分和确定,这是包括菲尔墨本人都未能回答的问题。那么,怎么解决这一根本问题,以使具有重要意义的格框架研究获得突破呢?一直苦无良策。当我们认真学习了 HNC 理论以后,终于找到了彻底解决问题的答案。

HNC 理论的四大主语义块和七大辅语义块,加上“语义块是句类的函数”这一基本论点,彻底解决了格语法的难题。四大主语义块和七大辅语义块在不同的句类中就有不同的语义内涵,这样,除 E 语义块外的一个语义块与一个具体的句类相联系,就产生一种格,这便形成了一个完备的格系统,从根本上解决了自然语言到底有多少格的大难题。这个格系统的完备性是由 HNC 句类系统的完备性所保证的。对格语法根本难题的解决也说明,HNC 理论的语义块并不等同于格。

HNC 理论的句类格式就是句子的语义构成模式,也就是乔姆斯基提出的深层结构。前文说到,HNC 理论穷尽地发现了自然语言语句的语义模式。所谓‘穷尽’,一方面是指数量上是有限的,而不是无限的;另一方面是指功能上是完备的,即可以描述自然语言的任何语句的语义结构。这是一个非常大的突破和贡献,它对自然语言理解和语言研究的重大意义是不言而喻的。

我们认为,HNC 理论的句类是句子类型研究的纲领,有了这个纲领,句法、语义、语用平面的句子类型研究都会形成系统。而汉语句子类型研究过去存在的根本难点和弱点正是缺少纲领和系统。

HNC 理论是面向自然语言理解的理论,但它的思想和方法对语言研究,特别是对汉语研究也具有开创性的指导意义。HNC 理论是在充分挖掘汉语特点的基础上创立的,在它的指导下系统、深入地开展汉语研究,将会开辟汉语研究的崭新局面。这使我不由得想起著名语言学家许嘉璐教授说过的一段话:“汉语理解是中文信息处理的高级阶段。在这一阶段的大规模真实文本处理中,不但需要计算机的硬件、软件研究成果,而且需要汉语的研究成果。

语言研究和计算机技术一结合,所带来的不仅是中文信息处理事业的顺利发展,而且有可能引发语言学研究的一场革命。”

致谢 本文写作过程中得到导师黄曾阳先生和林杏光先生的精心指导,雷良颖、刘志文、庄咏□等老师提了许多宝贵的修改意见,谨致谢忱。

参考文献

[1] 黄曾阳. HNC 理论概要. 中文信息学报, 1997(4)

[2] 黄曾阳. HNC 理解处理论文选录. 中国科学院声学研究所声场声信息国家重点实验室自然语言理解课题组. 1996

[3] 黄曾阳. HNC 自然语言理解处理的 52 个基本论题. 1998. http://farad.ioa.ac.cn/hzy.html

[4] 刘志文等. 自然语言语句的 HNC 表示. 语言文字应用, 1998(2)

[5] 林杏光. 正确引导汉语理解与汉语研究——事关人工智能研究的一个重要前提. 科技导报, 1997(3)

[6] 何东平. 我国计算机理解语言研究走出新路——黄曾阳创立的‘概念层次网络’理论正在产品化. 光明日报, 1998 年 6 月 12 日

[7] 张全. 基于 HNC 理论的语义块感知处理. 中国科学院声学所博士学位论文. 1996

[8] 晋耀红. 基于 HNC 理论的句类分析系统的设计与实现. 中国科学院声学所硕士学位论文. 1998

[9] 苗传江. “自然语言理解”的新进展——简评黄曾阳先生创立的 HNC 理论. 科技导报, 1998 (3)

[10] 林杏光等主编. 现代汉语动词大词典. 北京:北京语言学院出版社. 1994

[11] 林杏光. 基于格关系的现代汉语述语动词分类系统. 见:陈力为,袁琦主编. 计算语言学进展与应用. 北京:清华大学出版社. 1995

[12] 张庆旭. 现代汉语述语动词框架研究. 同上

[13] 苗传江. 现代汉语自动词句模研究. 见:陈力为,袁琦主编. 语言工程. 北京:清华大学出版社. 1997

[14] 胡裕树. 试论句子类型的研究. 汉语学习, 1995(5)

[15] Chomsky N. Aspects of the Theory of Syntax. MIT Press, 1965

[16] Fillmore C J. The case for case. In: Bach E, Harms R Eds. Universals in Linguistic Theory. New York. Holt: Rinehart and Winston, 1968

[17] Quilian M R. Semantic memory. In: Minsky M Ed. Semantic Information Processing. Cambridge, MA: MIT Press, 1968

[18] Schank R. Conceptual Information Processing. Amsterdam: North Holland, 1975

自然语言理解的新进展*

——简评黄曾阳先生创立的 HNC 理论

苗 传 江

(北京语言文化大学语言信息处理研究所,北京 100083)

《中文信息学报》1997 年第 4 期发表了中国科学院声学研究所研究员黄曾阳先生的论文“HNC 理论概要”。HNC 是“Hierarchical Network of Concepts(概念层次网络)”的简称,它以概念化、层次化、网络化的语义表达为基础,所以称它为概念层次网络理论。

1 HNC 理论包含极其丰富的内容

HNC 理论把人脑认知结构分为局部和全局两类联想脉络,认为对联想脉络的表述是语言深层(即语言的语义层面)的根本问题。什么是局部联想和全局联想呢?简单地说,局部联想是指词汇层面的联想,全局联想是指语句及篇章层面的联想。HNC 理论的出发点就是运用两类联想脉络来“帮助”计算机理解自然语言。

自然语言的词汇是用来表达概念的,因此,HNC 建立的词汇层面的局部联想脉络体现为一个概念表述体系。概念分为抽象概念与具体概念。HNC 理论的概念表述体系侧重于抽象概念的表达。对具体概念采取挂靠近似表达方法。HNC 理论认为应该从多元性表现和内涵两个方面来描述概念。它创立了五元组用来表达抽象概念的多元性表现,对抽象概念的内涵采用网络层次符号来表达。其网络层次符号包含三大语义网络:基元概念语义网络、基本概念语义网络和逻辑概念语义网络。HNC 的五元组符号和三大语义网络的层次符号组合起来就可完成对抽象概念的完整表达,从而为计算机理解自然语言的语义提供了有力的手段。

全局联想脉络是语句及篇章层面的联想。语句联想的主要内容是语义块和句类理论。语义块是句子的语义构成单位。主语义块 4 种,辅语义块 7 种。句类是句子的语义类别。有 7 个基本句类,它可构成 36 个混合句类。语义块和句类理论的基本论点是:语义块是句

---

* 本文发表于《科技导报》1998 年第 3 期。

类的函数。语义块和句类的这种函数关系具体体现为句类格式。句类格式是指一个句子的主语义块的排列顺序。以句类格式为基点的语句分析叫做句类分析。

以上介绍的两个联想脉络是 HNC 理论的基础部分,它的另一部分内容是自然语言理解的框架。以句类分析为基础,HNC 设计了自然语言处理系统的基本框架,这个框架由 9 个模块组成:1.单音词感知模块;2.语义块感知模块;3.句类分析模块;4.合理性分析模块;5.短时记忆知识模块;6.语境生成模块;7.隐藏知识揭示模块;8.要点主题分析模块;9.短时记忆向长时记忆扩展的模块。

自然语言处理离不开知识库,对知识库的设计和建立也是 HNC 理论的重要组成部分。已经建立了比较完备的概念知识库,目前正在紧张地进行汉语语言知识库的建立。

2 HNC 理论在自然语言表述和处理模式上的进展

HNC 在许多方面都在前人研究的基础上有所前进,这里述说它在自然语言表述和处理模式上所赢得的突破性进展。几十年来,自然语言理解的发展主要围绕着三个方面:1.自然语言的表述和处理模式;2.自然语言知识的表示、获取和学习;3.研制开发自然语言的应用系统。其中,自然语言的表述和处理模式是根本,它决定着整个自然语言理解的方向和进程。若干年来,自然语言理解的各个应用领域,比如机器翻译,都无重大进展,其主要原因正是由于缺少科学完备的自然语言表述和处理模式。黄曾阳先生认识到,自然语言传统分析模式(含统计模式)的根本弱点在于:它们不是描述语言感知过程的适当模式。他通过八年的艰苦探索,终于形成了三大理论要点:1. 要把自然语言所表述的知识划分为概念、语言和常识三个独立的层面,对不同层面采取不同的知识表示策略和学习方式,形成各自的知识库系统。知识库建设的首要目标应定位于自然语言模糊消解,这是 HNC 理论对迄今为止的知识库建设进行总结后得出的论断。2. 建立网络式概念基元符号体系,即概念表述的数学表示式。这个符号体系或表示式应具有语义完备性,能够与自然语言的词语建立起语义影射关系,同时,它必须是高度数字化的,每一个符号基元(每个字母或数字)都具有确定的意义,可充当概念联想的激活因子。这个符号体系就是 HNC 理论设计的三大语义网络及五元组和概念组合结构等,它是计算机把握并理解语言概念的基本前提,称为局部联想脉络,是 HNC 理论的基本内容之一。局部联想脉络的基本思路和做法是:把概念分为抽象概念和具体概念,对抽象概念用语义网络和五元组来表达,对具体概念采取挂靠展开近似表达的方法。3. 建立语句的语义表述模式,即语句表述的数学表示式。这一模式的完备性应表现为可表述自然语言任何语句的语义结构,即乔姆斯基所提出的语言深层结构。为表述自然语言语句的语义结构,HNC 理论提出了语义块和句类的概念,在此基础上形成的句类格式就是语言的深层结构,它是语句分析的基点,称为全局联想脉络,是 HNC 理论的另一基本内容。以上三大理论要点,正是 HNC 理论在自然语言表述和处理模式上赢得突破性进展的表现。

3 HNC 理论在中文信息处理技术上的进展

HNC 理论是面向整个自然语言理解的理论框架,但它首先关注的目标是中文信息处理。中文信息处理包括汉字信息处理和汉语理解。国家语委主任许嘉璐教授指出:"汉语理解是中文信息处理的高级阶段。在这一阶段的大规模真实文本处理中,不但需要计算机的硬件、软件研究成果,而且需要汉语的研究成果。语言研究和计算机技术一结合,必然引起语言学的一场革命。"从一定意义上说,汉语研究是汉语理解的前提和基础。几千年来,汉语语言学的传统研究主要集中在'字'的形、音、义上,相应建立了文字学、音韵学、训诂学。从 1898 年马建忠的《马氏文通》出版开始,汉语语法学出现以西方语言学理论研究汉语的景况,并成为汉语语法研究的主流派。应该说,100 年来的汉语语法研究是有成绩的。但随着汉语语法研究的不断深入,愈来愈多的学者认识到,西方语言学理论总的来说是在形态语言的基础上建立起来的,汉语是无形态语言,用形态语言的理论去描写无形态的汉语,这显然是不对路的。不少学者都想另辟蹊径而又找不到切实可行的道路。这种状况给中文信息处理设置了不可逾越的障碍。HNC 理论开辟了以语义表达为基础的自然语言理解的新路子,因而避开了当前中文信息处理所面临的一系列难题,诸如分词问题、词性标注问题、词的兼类问题、义项标注问题、句法分析问题、句子述语动词的识别问题,等等。由此可见,HNC 理论在中文信息处理技术上获得了突破性的进展。

4 HNC 理论的应用潜力和前景

HNC 理论走向应用的第一步是语义块感知和句类辨识。语义块感知就是找出一个句子中的各个语义块,句类辨识就是通过感知得到一个句子的 E 语义块(述语动词),进而确定这个句子所属的句类。计算机能否感知到语义块关系到 HNC 能否指导实践、是否有应用价值的问题,几年来的工程实践已对此作出了肯定的回答。感知到语义块、辨识出句类以后,就可以运用句类知识对句子进行理解处理,这称为句类分析。句类分析是对大脑语言感知过程的初步模拟。在模糊消解方面,理论上,句类分析应能接近甚至超过常人的水准,这一点已在汉语无声调拼音—汉字转换方面得到了验证。这使计算机向真正的理解迈出了坚实的第一步。在这第一步的基础上,HNC 理论设计了由 9 个模块组成的自然语言处理系统的基本框架。目前,部分模块已在计算机上得到实现。

HNC 理论的创立为我国开创自己的语言信息产业创造了良机。有人说,中国当前的信息产业面临的是八国联军入侵的局势,外国有关的大公司早已看到了中文信息处理的巨大市场,他们在向中国进军,凭着雄厚的经济实力,大肆'收买'中国的人才、技术和成果,如此长久下去,中国人哪还有自己的信息产业。不久前,美国的 IBM 公司推出了汉语语音输入系统,他们有一个不错的语音模型,但是,他们还没有一个好的语言模型。HNC 建立的语言表述和处理模型目前是无人可比的,它应该成为中国人的财富,应该以它为基础开创中国的信息产业。

令人可喜的是,国家计委已把“基于 HNC 理论的研究和开发”列入国家“九五”重点项目。在中国工程院院士陈力为教授等学术界前辈的推动下,为实现 HNC 理论,近一年来组成了“HNC 联合攻关队伍”。这一联合攻关队伍包括中国科学院声学研究所、中国人民大学对外语言文化学院和北京语言文化大学语言信息处理研究所等三家单位。他们在资金严重短缺的境况下紧张地工作,取得了显著的成绩。“HNC 联合攻关队伍”在过去一年里取得的一个重大成绩是,使 HNC 理论体系的完善从个人思考模式转向集体创立模式,这表明 HNC 理论的发展和应用存在着巨大的潜力和广阔的前景。

参考文献

[1] 黄曾阳. HNC 理论概要. 中文信息学报, 1997, (4)

[2] 林杏光. 正确引导汉语理解与汉语研究——事关人工智能研究的一个重要前提. 科技导报, 1997, (4)

简论黄曾阳先生创立的 HNC 理论*

苗 传 江

（北京语言文化大学语言信息处理研究所，北京 100083）

《中文信息学报》1997 年第 4 期发表了中国科学院声学研究所研究员黄曾阳先生的论文“HNC 理论概要”，这是一篇具有开创性的力作，它展示了自然语言理解的突破性进展，读后令人自豪和振奋。

HNC 理论是黄曾阳先生用长达八年的时间潜心探索、精心架构的创新成果，包含极其丰富恢弘的内容，在概念的表述系统、语句的表述模式、知识库的建设、自然语言理解系统框架的设计等方面，都有独到的建树和精到的见解。HNC 理论的精深内容和卓越的贡献远非一篇文章所能尽所欲言的，本文仅根据笔者初步的学习所得简论其两方面的内容和贡献。

1 HNC 理论创立了自然语言表述和处理的合理模式

自然语言处理作为人工智能的一个分支，已有 40 年的发展历程，形成了计算语言学这一跨接语言、信息、认知科学和计算机技术的边缘学科，它的发展主要围绕三个方面：1. 自然语言的表述和处理模式；2. 自然语言知识的表示、获取和学习；3. 研制开发自然语言的应用系统。其中，自然语言的表述和处理模式是根本，决定着整个自然语言理解的方向和进程。若干年来，自然语言理解的各个应用领域都无重大进展（比如机器翻译，特别是汉语与印欧语之间的翻译，搞了几十年，至今仍与实际应用水平相去甚远），其主要原因正是由于缺少科学完备的自然语言表述和处理模式。

纵观语言研究和自然语言处理的历史，在自然语言的表述和处理模式方面，源于印欧语系的语法学和句法分析一直居于主导地位。八大词类、六种句子成分、短语结构和句法树成为语言分析的基本概念和依托。对于这一传统分析模式，仅在 70 年代，就曾一度受到菲尔墨（Fillmore）和山克（Schank）的质疑和挑战。80 年代以来，语料库语言学的兴起使人们对统计模式产生了过高的期望，以致忽视了菲-山挑战的实质意义。

黄曾阳先生认识到，自然语言传统分析模式（含统计模式）的根本弱点在于：它们不是描

* 本文已投《中文信息学报》，待发表。

述语言感知过程的适当模式。

面对语音流的五重模糊(发音模糊、音词转换模糊、词的多义模糊、语义块构成的分合模糊、指代冗缺模糊),面对文字流的后三重模糊,大脑的语言感知应付裕如,表现了强大的解模糊能力,自然语言处理技术当前无从望其项背。

近 20 年来,自然语言处理囿于传统模式,不图突破。但是,它所面临的所有重大课题,从音词转换到机器翻译,从全文检索、信息抽取到智能阅读助手,都在呼唤语言表述及处理新模式的诞生:呼唤上下文联想处理向"知其所以然"的语义理解前进;呼唤向语言感知的方向靠拢。随着网络时代的来临,这一呼唤的迫切性和严峻性在与日俱增。

响应这一呼唤才意味着真正的突破,但突破的契机何在?悲观论者认为:语言感知过程密切依附于大脑中万亿神经元的神经网络,依附于浩瀚无垠的世界知识海洋,在对这个"网络"和"海洋"的奥秘未作充分揭示之前,模拟语言感知过程是不现实的。

事情果真如此悲观么?不。黄曾阳先生对此进行了八年的艰苦探索后,形成了以下三大理论要点,这三大要点集中体现了 HNC 理论在自然语言表述和处理模式上的突破。

1. 要把自然语言所表述的知识划分为概念、语言和常识三个独立的层面,对不同层面采取不同的知识表示策略和学习方式,形成各自的知识库系统。知识库建设的首要目标应定位于自然语言模糊消解。

这是 HNC 理论对迄今为止的知识库建设进行总结后得出的论断,具有极其重要的指导意义。

人工智能必须以知识为依托,自然语言理解必须以语言知识为依托。这是常识,没有人对此提出过疑义。但是人工智能和自然语言理解最需要什么样的知识?这些知识如何表达,又如何获得?这是知识库建设的基本问题。对这个问题的认识自人工智能诞生以来,已有了巨大的进步,但从自然语言理解的需要来看,这个进步是远远不够的。

人工智能的早期发起者几乎将知识混同于规则,这是不奇怪的,因为规则易于为计算机所把握。利用规则进行推理的过程,可利用产生式给以形式描述。这样,计算机的程序就可以模拟大脑的思考。如果大脑的思考过程仅仅是逻辑推理,那么,知识等同于规则的认识就是正确的。当然,大脑的运作过程不仅仅是推理,但推理终究是大脑运作的基本表现之一。因此,规则的运用仍然可以取得显著的效果。20 世纪 70 年代崭露头角的专家系统就是规则运用的巨大成果。不久前,IBM 的"深蓝"计算机在与国际象棋世界冠军卡斯帕洛夫的人机大战中赢得了胜利,应该说体现了这一运用的顶峰成就。

逻辑推理对自然语言处理、语言学和知识库建设都有重大影响。在语言学上的近期突出表现是蒙塔古语言学的兴起,在知识库建设上的集中表现是美国的 CYC 计划。至于自然语言理解,应该说,到目前为止,所有的自然语言理解系统,从早期的 LUNAR 和 HEARSAY 到最近的 LeMICON 都是规则系统。尽管后者的知识获得是自学习的,但知识的运用仍然是规则的。

以产生式形式表现的规则就是逻辑学的蕴涵关系,它是推理的基本形式。按照逻辑学

的观点,知识就是一系列的命题,命题之间存在推理关系。规模空前、推理规则达 100 多万条的 CYC 知识库就是基于这一思路花了 10 年时间(1985 1995)建立起来的,当初其主建者曾宣称,到世纪之交,CYC 知识库将成为计算机的基本配置之一。但是,到 10 年届满时,这个梦想完全落空,CYC 被一些人视为失败的典型。

CYC 建设的 10 年期间,正是语料库语言学大发展的 10 年,但主建者对此似乎置若罔闻,这成了批判者的基本论点,但主建者心里明白,他所追求的知识不是简单的统计可以得到的。那么,CYC 的根本问题何在?

根本问题在于该知识库的目标和知识表示方式。

CYC 知识库主建者将目标定位在建立一个万能的‘常人’自然语言理解系统,以弥补领域专家系统的不足。例如,一个心血管疾病的诊断专家系统并不能辨认患者年龄与体重的填写错误,CYC 系统可以帮助它解决这类问题。显然这涉及浩瀚无边的常识性知识,如果对这类知识采用一阶谓词加自然语言的方式加以描述,数以百万计甚至千万计的规则也难以包容,因此,CYC 含有 160 万条规则是不奇怪的。但是,问题的要害不在于一阶谓词,而在于以自然语言充当命题的概念表述符号,这是规则膨胀的根本原因。

上述 CYC 的目标应该说是自然语言理解的天职。主建者在语料库的呼声压倒一切时不逐时流,按既定方针坚持到底,值得钦敬。问题在于 CYC 的目标不可能一蹴而就,主建者犯了 70 年代山克先生的同样的错误,在沼泽地上建立高楼大厦。

自然语言理解的基础是语言知识,在语言知识里既包含与语言形式无关的概念知识,又包含与语言形式有关的纯语言知识,HNC 理论把前者称为概念知识,把后者称为(纯)语言知识,把语言知识以外的世界知识称为常识性知识,并且认为,把知识划分为概念知识、(纯)语言知识和常识性知识,并分别建库,这应该是知识库建设的第一条根本原则,CYC 及迄今为止的所有知识库都没有遵循这一原则。与 CYC 同时进行建设的大规模语言知识库还有美国的 WordNet 和日本的 EDR,这两类知识库存在的根本问题与 CYC 相同,但主建者的总体思路还不如 CYC。

语言知识库建设应将服务目标首先定位于自然语言五重或三重模糊的消解,HNC 理论把这一点作为知识库建设的第二条根本原则。口语五重模糊和书面语三重模糊的消解是理解的前提,但模糊消解的具体办法多种多样;消解的过程与理解的过程既有同步性又有异步性;模糊消解的深度是可测定的,而理解的深度是不可测定的(至少在目前);对模糊消解进行假设检验可以是无条件的,而理解是有条件的。如果说,自然语言理解的最终目标——如同大脑一样理解自然语言——过于遥远,那么能否把模糊消解作为近期目标,集中兵力予以突破呢?

计算语言学界对此并未形成共识。从理论上这个问题很难阐述明白,但从语言信息产业的角度来看,则可以说是一目了然。语音识别、文字识别、全文检索、机器翻译、文字校对等方面都已有应用软件投放市场,这些软件的共同弱点何在?就是在模糊面前无能为力,而用户对此又十分敏感。因此提高语言信息产品的市场信誉,从而提高市场占有率的根本出

路在于提高消解模糊的能力。这一点,不应存在任何疑义。明确语言知识库建设的这一中心目标十分重要,因为它关系到知识项的选择,关系到人工方式与语料库运用方式的分工等重大决策。

在知识表示方式上,HNC 知识库不再像 CYC 等一样用自然语言充当表述符号,而是创立了两套描述自然语言的数学表示式,即概念表述的数学表示式和语句表示的数学表示式。

2. 要建立网络式概念基元符号体系,即概念表述的数学表示式。这个符号体系或表示式应具有语义完备性,能够与自然语言的词语建立起语义映射关系,同时,它必须是高度数字化的,每一个符号基元(每个字母或数字)都应具有确定的意义,可充当概念联想的激活因子。

HNC 理论建立了这样的体系,该体系由五元组、语义网络和概念组合结构组成,它是计算机把握并理解语言概念的基本前提,称为局部联想脉络,是 HNC 理论的基本内容之一。局部联想脉络的基本思路和做法是:把概念分为抽象概念和具体概念,对抽象概念用语义网络和五元组来表达,对具体概念采取挂靠展开近似表达的方法。

概念有抽象与具体之分。在一般人看来,抽象概念总是比具体概念难于把握,中文信息处理界已做的汉语语义分类工作,对抽象概念总有力不从心之感。HNC 理论认为,实际上,抽象概念比具体概念更具有基元性、系统性,因而更容易表达;具体概念是客观存在物在人的思维中的直接反映,它里面包含了许多世界知识,而对世界知识是很难进行详尽表达的,所幸的是,人对具体概念理解和认识的深度可以比抽象概念浅,所以可以采取实用原则,"不求甚解"。HNC 理论侧重于抽象概念的表达。

HNC 理论设计了五元组、语义网络和概念组合结构来表达抽象概念。五元组是指{动态、静态、属性、值、效应}五大特性,它们是词性的基元,用以表达概念的外在表现。任何概念都具有五元组特性,比如英语中词根相同、词性不同的词就体现了同一概念内涵的不同的五元组特性,而汉语中的兼类词只不过是用一个词表达了同一概念内涵的几个五元组特性。语义网络用以表达概念的内涵。语义网络是树状的分层结构,每一层有若干个节点,每个节点代表一个概念基元(而不是词),每一层的若干节点分别用连续的数字标记,网络中的任一节点都可以通过从最高层开始到该节点结束的一串数字唯一地确定和表示,这种数字串称为层次符号。节点代表的概念基元通过不同方式的组合就可以表达各种各样的、无数的概念,而不受语种限制。概念组合结构用以表达概念基元的组合方式。五元组符号、层次符号和概念组合结构符号组合起来,就构成 HNC 的概念表示式。

HNC 用五元组和语义网络分别表达抽象概念的外在表现和内涵,这种表达方式便于描述概念之间的关联性。有的语义系统中分了事物类、运动类、时空类和属性类等几大类,这种分类割裂了概念之间的天然联系,因为按照这种分类,"总攻"和"进攻","航速"和"航行"、"斗志"和"昂扬"等大量概念上有天然联系的词会被划到不同的大类,表达它们之间的关联成为一个大问题,这一点是设计者已认识到的。在 HNC 的表述体系中,网络中的任何节点都具有五元组特性,上列词义只不过是同一概念节点的不同五元组表现而已,这样,它们之

间的关联就显式地体现出来了。

HNC 设计了抽象概念的三大语义网络:基本概念语义网络、基元概念语义网络和逻辑概念语义网络。三大语义网络是“概念基元”的聚类和系统,而绝非“词”的分类。语义网络的设计思想有两个主要来源:一是奎廉(Quillian)的语义网络、菲尔墨的格语法和山克的概念从属理论;二是汉语的“字义基元化,词义组合化”现象。第一个来源提出了“语义基元”的杰出思想并暗含着“总体表述”的宏伟目标,第二个来源则提供了语义基元的宝贵原料。汉语字少词多,仅用几千个汉字加以组合就构成许多的词。几千年来,汉语随着社会的发展而发展,新词不断增加,但组成词语的汉字却很少变化。汉字字义的基元化和汉语词义的组合化是一个伟大的宝藏,HNC 语义网络的建立深深发掘了这一宝藏。

HNC 用语义网络表达概念,其首要目标和价值在于给出概念关联性知识和联想脉络的线索,而不是给出概念的精确表示。自然语言理解的中心任务是解模糊,如同音模糊消解、一词多义模糊消解等,这些模糊的消解统称为多义选一处理。对自然语言词汇的多义选一处理是人类理解自然语言过程中最频繁、最基本的操作。对这一操作过程的形式模拟不在于并行处理或快速计算,而在于以什么巧妙的方式完成大量语义距离的计算。语义网络层次符号的构造方式把最频繁、最基本的语义距离计算变成了对层次符号的简单逐层比较。这是 HNC 用语义网络层次符号表达概念的基本出发点。层次符号是一种灵活的分层结构,它到任一层都代表一个概念,至于这个(些)概念与相应的语言概念之间,究竟谁是谁的近似,已无关紧要。重要的是,层次网络符号对概念的局部联想脉络给出了明确的表示。

三大语义网络是 HNC 理论的核心,是精心构造和设计的结果,每一个节点的设置都颇费思虑。这一设计的完成是一项伟大的创造。

语义网络层次符号的设计为计算机理解自然语言提供了有力的手段、奠定了坚实的基础。当然,在工程实现上首先要完成用层次符号描写自然语言词汇语义的工作,这是一项浩大而艰巨的工程,但这个瓶颈问题跟过去相比已有了本质的不同,过去缺乏语义描写的完备手段,现在手段已具备,剩下的只是工作量的问题了。

语言理解的基础是把握概念,而如何把握自然语言表达的纷繁万千的概念,语言学和自然语言理解长期以来都没有重大进展。传统语言学对词义有相当深入的研究,但缺乏系统性和宏观理论指导。现代语义学的义素分析法和语义场理论都富有启发意义,前者把词义分析成更小的单位,蕴涵着概念基元的思想;后者着眼于词义之间的关联性,蕴涵着系统网络的思想。但是,它们还难以应用于自然语言理解系统,因为它们还远不够完善,还没有解决表述自然语言概念的根本问题。义素分析法没有解决“自然语言到底有多少义素”的问题,语义场理论没有解决“自然语言到底有多少语义场”、“语义场该怎样划分”、“语义场之间和内部有怎样的关系”等问题。这些问题的根源在于缺少对自然语言概念的宏观把握。HNC 理论设计的基元化、层次化、网络化的三大语义网络从根本上解决了这些问题。语义网络的各个节点,即概念基元,相当于义素。网络高中层节点的完备设计,加上可扩充的分层结构,使它具有了描述任何概念的能力。语义网络是一个整体的设计,是一个完整的系

统,它各个节点下的网络都形成相关联的概念的聚类,这些聚类就相当于语义场。更重要的是,通过语义网络,语义场内部、语义场之间都建立了联系。

3. 要建立语句的语义表述模式,即语句表述的数学表示式。这一模式的完备性应表现为可表述自然语言任何语句的语义结构,即乔姆斯基(Chomsky)所提出的语言深层结构。

HNC 理论建立了这样的表述模式,这个模式是在句类和语义块基础上形成的句类格式,它是语句分析的基点,称为全局联想脉络,是 HNC 理论的另一基本内容。

HNC 理论的句类是对语句的语义分类。自然语言的语句千变万化,如何进行语义分类呢?这一直是个无从下手的难题。HNC 理论成功地解决了这一难题。三大语义网络中的基元概念语义网络有六个一级节点:作用、过程、转移、效应、关系、状态,这六个节点形成作用效应链。"作用效应链反映一切事物的最大共性,作用存在于一切事物的内部和相互之间,作用必然产生某种效应,在达到最终效应之前,必然伴随着某种过程或转移,在达到最终效应之后,必然出现新的关系或状态。过程、转移、关系和状态也是效应的一种表现形式。新的效应又会引发新的作用,如此循环往复,以至无穷,这就是宇宙间一切事物存在和发展的基本法则,也是语言表达和概念推理的基本法则。"HNC 理论根据作用效应链的六个环节对语句进行分类,加上作为人类思维活动基本内容的判断,共形成 7 大基本句类。各基本句类在语句的语义构成上各有鲜明特点。基本句类下面有不同层次的子类,子类的定义有总体设计。基本句类可以构成混合句类。自然语言的语句虽然丰富而复杂,但它们表达的信息总是由 7 个基本句类组成的。基本句类、子类和混合句类构成 HNC 理论的句类系统,从已经对语言材料做过的大量分析来看,这个系统是完备的。

HNC 理论的语义块是语句的语义构成单位。语义块概念的提出便于从语言深层描述语句。用传统语言学的词或短语无法清楚地界定一个句子是否完备,如果问一个句子应该或者可能有多少个词或短语,便难以回答。语义块是语义,即语言深层的定义,它不依赖于形式,可以明确地根据句类描述语句的构成。经过高度抽象和概括,HNC 理论确定了四大主语义块(特征、作用者、对象、内容)和七大辅语义块(条件、手段、工具、途径、参照、因、果)。四大主语义块中,特征语义块决定句类。

HNC 理论关于句类和语义块的基本论点是:语义块是句类的函数。这是该理论在建立语句语义表述模式上的精华之所在。这一论点包含着丰富的内容,有两点是基本的:不同的句类需要不同的语义块配置;语义块的具体内涵要根据句类来确定。语义块和句类之间的函数关系是概念层面的固有知识,与语种无关。句类和语义块配置构成句类格式,这就是语言的深层结构。有了句类和语义块的合理设计及它们之间的函数关系,HNC 理论就可以完备地表述自然语言语句的语义结构了。

乔姆斯基提出语言的深层结构,被称为一场革命,但是他没有解决如何描述语言深层结构的问题。格语法理论的创立者菲尔墨是对宾语和主语进行语义分类的第一位先行者,最早想到了"语义块是句类函数"的概念,可惜他的理论匆忙出台,在理论的总体性和层次性方面都比较欠缺。HNC 理论在他们的基础上创立了完备的语言深层结构表述模式,具有突破

性贡献和意义。

据黄曾阳先生介绍,上述概念和语句表述模式只是 HNC 理论宏伟目标的一部分,HNC 的宏伟目标是建立以下六个层次上的“自然语言计算机感知模式”:1. 自然语言概念体系表述模式;2. 自然语言语义块和语句的表述模式;3. 句群关联性表述模式;4. 篇章要点表述模式;5. 短时及长时记忆的生成转换模式;6. 知识自学习模式。

综上所述,HNC 理论创立了基于语义的自然语言表述和处理的科学模式,开创了语言研究的新局面,开辟了自然语言理解的新途径。传统的语言表示和处理模式以语法为基础。语法有狭义与广义之分,狭义语法是指以形态变化和虚词搭配为依托的语言法则,这些法则里本来包含语义信息,但语法学从自身研究的便利出发曾长期有意脱离语义而自成体系。这个状况直到乔姆斯基的转换生成语法和菲尔墨的格语法出现以后才发生了变化,随后的功能语法继承了乔姆斯基和菲尔墨的传统,这些语法应称为广义语法,它包含了语义甚至语用。但是,广义语法学虽然融入了语义知识,并未对语义表述给出完善的理论框架。HNC 理论从根本上改变了这一状况,“根本”的具体表现就是建立了表述自然语言概念和语句的两套数学表示式。

在应用上,HNC 理论把以句类格式为基点的语句分析叫做句类分析。句类分析是对大脑语言感知过程的初步模拟,在模糊消解方面,理论上,句类分析应能接近甚至超过常人的水准,这一点已在汉语无声调拼音 - 文字转换方面得到了验证。这使计算机向真正的理解迈出了坚实的第一步。在这第一步的基础上,HNC 理论设计了自然语言处理系统的基本框架,这个框架由 9 个模块组成:1. 单音词感知模块;2. 语义块感知模块;3. 句类分析模块;4. 合理性分析模块;5. 短时记忆知识模块;6. 语境生成模块;7. 隐藏知识揭示模块;8. 要点主题分析模块;9. 短时记忆向长时记忆扩展的模块。目前,部分模块已在计算机上得到实现。

2 HNC 理论开辟了汉语研究的新路子,解决了汉语理解所面临的诸多难题

自 1898 年《马氏文通》问世后的整整一百年来,汉语语法学的研究确实取得了不少成绩,但问题也越来越突出。越来越多的人认识到,问题的根本原因在于,一百年来的汉语研究基本上都是在套用印欧语言的语法学,而汉语同印欧语有巨大差异,语法学不适用于汉语研究。“语法”这个词汉语原来是没有的,是从西方引进的,但这不等于说汉语传统语言学没有语法的概念,只不过表明语法对汉语传统语言学所面临的问题不十分重要罢了。前文提到,语法中本来是包含语义信息的,但语法学从自身研究的便利出发长期脱离语义而自成体系,正是基于这一点,我国著名的音韵训诂学家黄侃先生曾将《马氏文通》戏称为“狗屁不通”,绝不只是戏言。汉语语法学一开始就遇到的问题,诸如词的兼类问题、主宾语问题等,至今没有解决。把不符合汉语特点的语法研究的思路和成果应用于汉语理解,自然遇到了一系列难以解决的问题,使汉语理解难以前进。这些问题主要有:(1)汉语“词无定类”,兼类十分普遍,词类与句法成分之间没有明确的对应关系,难以凭借词类进行有效的句法分析;

(2)汉语是无形态的语言,句法分析没有可利用的词形变化,确定句子的中心动词成为一大难题;(3)汉语的句法结构相当灵活,难以把握;(4)汉语语法上的主语、宾语等句法成分与语义上的施事、受事等论旨角色的关系十分复杂,难以根据句法分析的结果进行语义理解。

百年来的汉语研究证明,汉语是"意合型"的语言,不能套用印欧语的语法学来研究,应该建立基于汉语特点的语言研究理论。二十年来的汉语理解实践表明,从分词开始的每一步都无法彻底实现,根本问题在于每一步都离不开理解,应该开创不依赖形式分析的新路子;更为重要的是,汉语理解需要宏观的理论指导。

汉语研究和汉语理解的困境在呼唤符合汉语特点的新理论、新技术,HNC 理论成功地响应了这一呼唤。解决汉语研究和汉语理解的难题,是 HNC 理论创立之初的首要目标,现在,这个目标已经实现,目标的实现是以前文所述的自然语言表述和处理模式为基础的。HNC 理论建立的语言模型直接从深层语义出发,不再停留于表层形式,摆脱了传统语法学的束缚,尤其适用于汉语研究,它将开创汉语研究的新局面。HNC 理论设计了汉语理解的宏观理论框架,不再走分词、词性标注、句法分析的老路,而是从语义块感知和句类分析入手,直接迈上语义理解的台阶,使老路上的难题或者不复存在,或者得到了解决。

下面仅以 HNC 理论对词性(词类)问题的创见为例来"管窥"它所开辟的汉语研究新思路。

建立在印欧语言形态变化基础上的词性无法落实于汉语。汉语本来是没有"词性"一说的,《马氏文通》以来,语法学界对汉语的词性讨论来讨论去,总难免有"词无定类"的感慨,有"依句辨品,离句无品"的结论。问题主要在于汉语词的兼类太严重,难以处理,《现代汉语词典》中一直没有标词性,自有其苦衷。但是,这不能说明汉语的词没有词性。那么,词性问题的根本在哪里,该怎样解决呢?HNC 理论的五元组从语言深层阐释并解决了词性问题。五元组是前述概念表述体系的组成部分。

任何一个概念都需要从不同侧面予以表达,这种现象叫做概念的多元性表现。具体概念的多元性表现十分复杂,难以给出规范化的表达,抽象概念则有所不同,它的多元性表现在自然语言中有明显的迹象,这就是词性现象。印欧语言的词根或具有词根特色的词,可以加上不同的后缀分别构成动词、名词、形容词和副词,这种词性的转换就是抽象概念多元性的生动表现,也就是说,词根相同词性不同的词是对同一概念不同侧面的表达。汉语对抽象概念的多元性表现则没有相应的形式标示,而往往是同一个词兼有名词、动词、形容词、副词中的几个属性。汉语的词性模糊现象和西语以词缀变化表现不同词性的现象都是抽象概念多元性的生动表现,词缀变化的有无只是一种形式,本质在于抽象概念本身具有这种多元性表现的固有特征。

那么,抽象概念多元性表现的"多"是一个模糊的"多",还是一个确定的"多"?或者说,能否给以规范化的表达?或者再换一个说法,这个多元性表现的"多"是否存在某些基元(primitive)呢?HNC 的答案是肯定的。抽象概念需要从动态、静态、属性、值和效应五个侧面加以表达,这就是抽象概念的五元组特性,简记为:{v,g,u,z,r}特性,它们是抽象概念多元

性表现的基元。任何概念都具有五元组特性,即都需要从五个侧面加以表达,不过,对某个抽象概念各个侧面的表达,自然语言中未必有相应的词语,而且不同语种间存在着差别。反过来,自然语言中的一个表达抽象概念的词语必定是从五元组中的某个或某几个侧面来表达某个抽象概念。例如,"思考、思维、想法"就是分别从五元组的 vg,g,r 侧面对同一概念内涵的表达。五元组是词性的本质内容,是词性的基元。所以,不必为汉语词的大量兼类现象感到困惑。

应该指出,HNC 理论开创的基于深层语义的语言理论,不仅适用于汉语研究,也适用于包括印欧语在内的其他语言的研究;它开拓的不以句法分析为依托的新路子,不仅适用于汉语理解,也适用于世界整个自然语言理解。

3 结束语

我们认为,HNC 理论是相当成熟的全新的理论,它是中国人创立的、基于汉语特点的自然语言理解理论。它的创立为我国开创自己的语言信息产业创造了契机。有人说,中国的信息产业当前面临的是八国联军入侵的局势,有关外国大公司早已看到中文信息处理的巨大市场,他们在向中国进军,凭着雄厚的经济实力,大力"收买"中国的人才、技术和成果,如此长久下去,中国人还哪有自己的信息产业。不久前,IBM 公司推出了汉语语音输入系统,他们有一个不错的语音模型,但是,他们还没有一个好的语言模型。HNC 建立的语言表述和处理模型目前在国内外都是无人可比的,它应该成为中国人的财富,我国应该以它为基础来开创有中国特色的信息产业。我们期待着 HNC 理论大展鸿图。

主要参考文献

[1] 黄曾阳. HNC 理论概要. 中文信息学报, 1997,(4)

[2] 黄曾阳. HNC 理解处理论文选录. 中国科学院声学研究所声场声信息国家重点实验室自然语言理解课题组,1996. 3

[3] 黄曾阳.〈理解问答〉、〈关于 HNC 词知识库的建设〉. 内部资料

[4] 张全. 基于 HNC 理论的语义块感知处理. 中国科学院声学所博士学位论文

[5] 林杏光. 正确引导汉语理解与汉语研究——事关人工智能开发的一个重要前提. 科技导报, 1997,(4)

[6] 张普. 论语义场. 见:陈力为,袁琦主编. 中文信息处理应用平台工程. 北京:电子工业出版社, 1995

[7] 陈群秀,张普. 信息处理用现代汉语语义分类体系:属性分类. 同上

[8] 陈小荷. 汉语语义自动分析的任务与策略. 同上

[9] 鲁川. 现代汉语的语义网络. 同上

[10] 苗传江. 自然语言理解的新进展——简评黄曾阳先生创立的 HNC 理论. 科技导报, 1998 (3)

[11] 姚天顺等. 自然语言理解——一种让机器懂得人类语言的研究. 北京:清华大学出版社, 1995.12

关于汉语词库结构及汉语文本之汉字表示的建议

杜 燕 玲

（中国科学院声学研究所，北京 100080）

引言

语音和语义是一切语言单位的两极，词是音义两极相结合的统一体。“文字只是这些音义结合体的书写符号”。这些提法对于采用拼音文字的语言，其正确性是无庸置疑的。但汉语的情况有所不同，在汉字形成的初期，它当然也“只是音义结合体的书写符号”，这大体相应于象形字和指事字。但随着会意字和形声字的发展，文字不再是语音的附属品，而取得了独立的表意功能，使汉语成了音、形、义三极语言。两极意味着对义的表达只有音一种手段，这种语言基本上独立于文字而发展。三极则意味着对义的表达有音形两种手段，文字在一定时期随语言同步发展，并对后者产生重大影响。对音的运用属于人类的本能，对形的运用则涉及更高级的智能。汉语对音形两极的运用必然表现出更多的智能性，这是它的长处。但另一方面，又限制了语音本能的充分发挥，这是它的弱点。汉语的这种双重性在词汇的构成方面表现得最为明显。语言的发展从词汇起步，词汇起源于对事物的命名。汉语与西语在命名方式上的差异不仅饶有趣味且极富启发性。古汉语的基本命名以单音节为限，几乎不越雷池一步，显得非常原始和笨拙。西语对一个命名的音节数量则不加限制，显得十分灵活和潇洒。但命名的需要随着社会的发展而层出不穷，当新的需要出现时，汉语采取以原有汉字重新组合的方式予以表达，充分显示出其灵活和潇洒的本质。西语则恰恰相反，由于原有词汇的音节数量已不适于再行组合，不得不采取另造新词的原始方式，从而显示出其灵活潇洒中的死板和笨拙。这样，汉字就成了以不变应万变的万能构词基元，两千余年来，基本只减不增，依靠一千多个语义充分基元化的汉字，对一切新概念的表达应付裕如。我们把汉语的这一独特语言现象叫做“字义基元化，词义组合化”。

汉字是形、音、义三者的结合体并不是新观点，训诂学对此早有深刻认识。汉字的这一根本特点对中文信息处理带来了一系列的新问题。首先是汉字的计算机输入和存储，这个问题大体上已经解决了。其次是汉语语音识别结果的音词转换，这个问题还远未解决。最后是汉语的理解处理，表面上似乎与汉字无关，实际上极为密切，音词转换[4]、分段层选处

理[9]、新词及伪词辨识[10]都是源于汉字的特殊需要;音节感知库和字义知识库的必须独立于词知识库而另建,也是源于汉字。但汉字对汉语最本质的影响是将汉语"约束"成单音节语言,而不是像西语那样的音节串语言。词汇以单音词和双音词为主,汉语的这一"单双性"既为汉语连续语音识别和理解处理带来了巨大的机遇,又带来了特殊的困难。这个问题在[4]中已有详细阐述。

本文仅涉及汉字的计算机存储,文中所建议的汉字编码方案已在我们的汉语理解处理系统中使用多年。设计这一编码的目的是为了更好满足理解处理的需要。我们还希望将这一方案推广到其它语言,这当然只是一个设想。无论是已完成的设计或设想都带有建议性和探索性。

1 关于汉语文本之汉字表示的建议

近 20 年来,国内外技术人员为中文信息处理作出了不懈的努力。首先是制定了汉字机内码国家标准 GB2312-80,通称国标码。标准化的汉字内码与汉字字符集有着简明的对应规则。而字符集的区位排列次序与汉字的发音并没有必然的联系。基于中文信息处理过程经常出现语音与文字相互转换的需要,国标码显然不能适应这一情况。其中,长期困扰文语转换的多音字问题就是典型的例子。

这里建议的就是一种寓音形信息于一体的汉字编码方案。

汉语用拼音字母表示的基本音总计 406 个,寓音形信息的编码方案基于汉字的下列分布特征。

	汉字数量 N	基本音个数
(1)	$N \leqslant 32$	353
(2)	$2 \times 32 \geqslant N > 32$	41
(3)	$3 \times 32 \geqslant N > 2 \times 32$	9
(4)	$4 \times 32 \geqslant N > 3 \times 32$	3
(5)	$N > 4 \times 32$	0

如果对上列 1 4 类分布分别给以 1 4 个编码,总共需要 474 个编码,可用 9 位表示,命名为音码,每一音码内的不同汉字用 5 位表示,命名为序码,总计 14 位。仍然是用两个字节表示一个汉字。但音码表示了拼音,序码表示了声调和字形,汉字的音形信息完整地寓于一体。

当然,在实际进行编码时,不应限于 474,而应该用满 512,如第 3 节的音码表所示。

音码的意义在于构造音码矩阵,由于汉语的非单音词以双音词为主,这个音码矩阵可为音词转换带来极大的便利。

所谓音码矩阵就是一个 512×512 的方阵,方阵的每一结点(元素)用一位或两位表示,仅占用 31.25K 或 62.50K 字节。沿着这个矩阵的某一行搜索,可找出以该音为第一音的全部词汇,沿着这个矩阵的某一列搜索,可找出以该音为第二音的全部词汇。如果有必要的

话,不难把这一穷极搜索功能扩展到指定音节在多字词中的任一位置的情况。

音码矩阵以其高效的双向搜索功能为基础,其作用不仅在于便利语音与文字的转换,而且是实现词义库与词库同构的关键;是实现对语音识别进行二次引导处理[4]的保证;是实现分离结构词库的基础。最后一点下一节会详细说明。总之,音码矩阵是一种灵活高效的数据结构,它充分体现了汉语的特点,以它为基础建立的理解处理工作平台,就能同时满足汉语语音及文字处理的各种需要。“85”期间,我们为中科院“汉语人机对话系统”研制的理解处理软件,以及相应的汉语词库、字义库、词义库和音节感知库,都是以音序码为依托的。

最后应该说明,在显示或打印汉字时,当然应将音序码转换成国标码,以利用已有的成果。但汉语文本的存储,则改用音序码为宜,因为,它为文本的理解处理消除了与多音字有关的一切障碍。

2 汉语词库的数据结构

中文信息处理系统早已由字处理过渡到词处理阶段。以从键盘向计算机输入汉字为例,输入一个词一般比孤立地输入构成词的单字的重码率要低得多。语音识别系统中,计算机所能得到的也是一系列声母、韵母或音节信息。与键盘输入的差别仅在于,它不是一个个确定的音,而是一个包含多个候选音的模糊阵列。这样的音—字转换系统,同样需要得到词库的支持。通常最自然的做法是:词库是一个独立的结构。只要建立索引表,即可直接由音找到词形。在搜词过程中,需要对组成该词的汉字机内码及对应的显示字库频繁地访问与调用。如果该处理系统还要调用词性、语义和其他有关信息,计算机的查询负担则会相应增加,搜索速度很难达到工程上实时的要求。特别是在面对一个模糊的语音阵列时,往往需要对大多数不满足组词条件的相邻音尽快予以排除。在这种情况下,基于国标码的一般词库的查询方式,将不得不为大量的无效或冗余信息付出宝贵的时间代价。

音码和序码概念的引入,必然对汉语词库的数据结构带来新的思路。这就是上面所说的“分离结构的词库”。

所谓分离结构,就是词库由音码矩阵、结点说明库和扩展库三部分构成。

结点说明库采用规范化结构,对每个结点统一用两字节进行说明。正是这一规范化措施使词义库得以与词库同构[7],从而实现了两库寻址的合一。

引入扩展库,是结点说明库得以规范化的关键。

结点说明库用 2 位说明结点是否需要扩展,即结点的类型说明。它是说明库的固定部分。

不需要扩展的结点只有一个双音词,这是结点的大多数情况。这时,用 10 位标明双音词的两个序码,用 2 位标明它的级别,另外 2 位表明它是否儿化及能否插入。级别分 4 级:0 —— 一级常用词,1——二级常用词,2——专业词汇,3——非常用词。

需要扩展的结点分三种情况,一是有同音词,二是多音词,三是双音词与多音词并存。这时,说明内容为三类扩展库的地址。

在结点说明库中只需要给出双字词的两个序码。作为一种数据结构,不仅与汉字输入的‘双拼’方式,即对每一汉字的声母和韵母各击一次键的输入方式,最相匹配,而且可以将双字词词库的存储空间将近节省一半。

3 具体的音码表

在下列音码表中,“拼音”栏内的字符表示汉语拼音的一个音节。如‘拼音’栏为空,则表示该音码的音节与前一个音码的音节相同,依次类推。

具体的音码表如下:

音码	拼音	包含的调号	音码	拼音	包含的调号	音码	拼音	包含的调号
32	a	(1 2 3 4)	33	ya	(1 2)	34		(3 4 5)
35	ba	(1 2 3 4 5)	36	pa	(1 2 4 5)	37	da	(1 2 3 4 5)
38	ta	(1 3 4)	39	cha	(1 2 3 4)	40	sha	(1 3 4)
41	zha	(1 2 3 4)	42	ca	(1 3)	43	sa	(1 3 4)
44	za	(1 2 3)	45	ga	(1 2 3 4)	46	ka	(1 3)
47	ha	(1 2 3 4)	48	ma	(1 2 3 4 5)	49	na	(1 2 3 4 5)
50	la	(1 2 3 4 5)	51	fa	(1 2 3 4)	52	jia	(1 2)
53		(3 4 5)	54	qia	(1 3 4)	55	xia	(1 2 4)
56	ai	(1 2 3 4)	57	bai	(1 2 3 4)	58	pai	(1 2 3 4)
59	dai	(1 3 4)	60	tai	(1 2 3 4)	61	chai	(1 2 4)
62	shai	(1 3 4)	63	zhai	(1 2 3 4)	64	cai	(1 2 3 4)
65	sai	(1 4)	66	zai	(1 3 4)	67	gai	(1 3 4)
68	hai	(1 2 3 4)	69	kai	(1 3 4)	70	mai	(2 3 4)
71	nai	(3 4)	72	lai	(2 3 4)	73	ao	(1 2 3 4)
74	yao	(1 2)	75		(3 4)	76	bao	(1 2 3 4)
77	biao	(1 3 4)	78	pao	(1 2 3 4)	79	piao	(1 2 3 4)
70	dao	(1 2 3 4)	81	diao	(1 3 4)	82	tao	(1 2 3 4)
83	tiao	(1 2 3 4)	84	chao	(1 2 3 4)	85	shao	(1 2 3 4)
86	zhao	(1 2 3 4)	87	cao	(1 2 3)	88	sao	(1 3 4)
89	zao	(1 2 3 4)	90	gao	(1 3 4)	91	hao	(1 2 3 4)
92	kao	(1 3 4)	93	mao	(1 2 3 4)	94	miao	(1 2 3 4)
95	nao	(1 2 3 4)	96	niao	(3 4)	97	lao	(1 2 3 4)
98	liao	(1 2 3 4)	99	rao	(2 3 4)	100	jiao	(1 2)
101		(3 4)	102	qiao	(1 2)	103		(3 4)
104	xiao	(1 2 3 4)	105	an	(1 3 4)	106	yan	(1 2)
107		(2 3)	108		(4)	109	ban	(1 3 4)
110	pan	(1 2 4)	111	dan	(1 3 4)	112	tan	(1 2 3 4)
113	chan	(1 2 3 4)	114	shan	(1)	115		(3 4)
116	zhan	(1 3 4)	117	can	(1 2 3 4)	118	san	(1 3 4)
119	zan	(1 2 3 4 5)	120	gan	(1 3 4)	121	han	(1 2 3 4)

音码	拼音	包含的调号	音码	拼音	包含的调号	音码	拼音	包含的调号
122	kan	(1 3 4)	123	man	(1 2 3 4)	124	nan	(1 2 3 4)
125	lan	(2 3 4)	126	ran	(2 3)	127	fan	(1 2 3 4)
128	bian	(1 3 4 5)	129	pian	(1 2 3 4)	130	dian	(1 3 4)
131	tian	(1 2 3 4)	132	jian	(1)	133		(3)
134		(4)	135	qian	(1 2)	136		(3 4)
137	xian	(1 2)	138		(3 4)	139	mian	(2 3 4)
140	nian	(1 2 3 4)	141	lian	(2 3 4)	142	ang	(1 2 4)
143	yang	(1 2 3 4)	144	bang	(1 3 4)	145	pang	(1 2 3 4)
146	dang	(1 3 4)	147	tang	(1 2 3 4)	148	chang	(1 2 3 4)
149	shang	(1 3 4 5)	150	zhang	(1 3 4)	151	cang	(1 2)
152	sang	(1 3 4)	153	zang	(1 3 4)	154	gang	(1 3 4)
155	hang		156	kang	(1 2 4)	157	mang	(2 3)
158	nang	(1 2 3)	159	niang	(2 4)	160	lang	(1 2 3 4)
161	liang	(2 3 4)	162	rang	(1 2 3 4)	163	fang	(1 2 3 4)
164	jiang	(1 3 4)	165	qiang	(1 2 3 4)	166	xiang	(1 2 3 4)
167	e	(1 2 3 4 5)	168	er	(2 3 4)	169	de	(2 5)
170	te	(4)	171	che	(1 3 4)	172	she	(1 2 3 4)
173	zhe	(1 2 3 4 5)	174	ce	(4)	175	se	(4)
176	ze	(2 4)	177	ge	(1 2)	178		(3 4)
179	ke	(1 2)	180		(3 4)	181	he	(1 2)
182		(4)	183	me	(5)	184	ne	(2 4 5)
185	le	(1 4 5)	186	re	(3 4)	187	ye	(1 2 3 4)
188	bie	(1 2 3 4)	189	pie	(1 3)	190	die	(1 2)
191	tie	(1 3 4)	192	jie	(1 2)	193		(3 4)
194	qie	(1 2 3 4)	195	xie	(1 2)	196		(3 4)
197	mie	(1 4)	198	nie	(1 4)	199	lie	(1 3 4 5)
200	bei	(1 3 4 5)	201	pei	(1 2 4)	202	zei	(2)
203	gei	(3)	204	hei	(1)	205	mei	(2 3 4)
206	nei	(3 4)	207	lei	(1 2 3 4 5)	208	fei	(1 2 3 4)
209	en	(1 4)	210	ben	(1 3 4)	211	pen	(1 2 4)
212	chen	(1 2 3 4 5)	213	cheng	(1 2 3 4)	214	shen	(1 2 3 4)
215	sheng	(1 2 3 4)	216	zhen	(1)	217		(3 4)
218	zheng	(1 2 3 4)	219	cen	(1 2)	220	ceng	(1 2 4)
221	sen	(1)	222	seng	(1)	223	zen	(3 4)
224	zeng	(1 4)	225	gen	(1 2 3 4)	226	geng	(1 3 4)
227	hen	(2 3 4)	228	heng	(1 2 4)	229	ken	(3 4)
230	keng	(1)	231	men	(1 2 4 5)	232	nen	(4)
233	ren	(2 3 4)	234	fen	(1 2 3 4)	235	beng	(1 2 3 4)
236	peng	(1 2 3 4)	237	deng	(1 3 4)	238	teng	(2)

音码	拼音	包含的调号	音码	拼音	包含的调号	音码	拼音	包含的调号
239	meng	(1 2 3 4)	240	neng	(2)	241	leng	(1 2 3 4)
242	reng	(1 2)	243	feng	(1 2 3 4)	244	o	(1 2 4)
245	yo	(1 5)	246	bo	(1 2)	247		(2 3 4 5)
248	po	(1 2 3 4)	249	mo	(1 2)	250		(3 4)
251	fo	(2)	252	ou	(1 3 4)	253	you	(1 2)
254		(3 4)	255	pou	(1 2 3)	256	dou	(1 3 4)
257	tou	(1 2 3 4)	258	chou	(1 2 3 4)	259	shou	(1 2 3 4)
260	zhou	(1 2 3 4)	261	cou	(4)	262	sou	(1 3 4)
263	zou	(1 3 4)	264	gou	(1 3 4)	265	hou	(2 3 4)
266	kou	(1 3 4)	267	mou	(1 2 3)	268	lou	(1 2 3 4)
269	rou	(2 4)	270	fou	(3)	271	dong	(1 3 4)
272	tong	(1 2 3 4)	273	chong	(1 2 3 4)	274	zhong	(1 3 4)
275	cong	(1 2)	276	song	(1 3 4)	277	zong	(1 3 4)
278	gong	(1 3 4)	279	hong	(1 2 3 4)	280	kong	(1 3 4)
281	nong	(2 4)	282	long	(1 2 3 4)	283	rong	(2 3)
284	yong	(1 2 3 4)	285	jiong	(1 3)	286	qiong	(2)
287	xiong	(1 2)	288	yi	(1)	289		(2)
290		(3)	291		(4)	292		(4)
293	bi	(1 2 3)	294		(4)	295		(4)
296	pi	(1 2)	297		(3 4)	298	di	(1 2)
299		(3 4)	300	ti	(1 2 3 4)	301	chi	(1 2)
302		(3 4)	303	shi	(1)	304		(2)
305		(3)	306		(4)	307		(4 5)
308	zhi	(1)	309		(2)	310		(3)
311		(4)	312		(4)	313	ci	(1 2 3 4)
314	si	(1)	315		(3 4 5)	316	zi	(1 2)
317		(3 4 5)	318	ji	(1)	319		(1)
320		(2)	321		(3)	322		(4)
323	qi	(1)	324		(2)	325		(3 4)
326	xi	(1)	327		(1)	328		(2)
329		(3 4)	330	mi	(1 2 3 4)	331	ni	(1 2 3 4)
332	li	(1 2)	333		(3)	334		(4)
335		(4 5)	336	ri	(4)	337	yin	(1 2)
338		(3 4)	339	ying	(1 2)	340		(3 4)
341	bin	(1 4)	342	bing	(1 3 4)	343	pin	(1 2 3 4)
344	ping	(1 2)	345	jin	(1)	346		(3 4)
347	jing	(1)	348		(3 4)	349	qin	(1 2 3 4)
350	qing	(1 2 3 4)	351	xin	(1 2 4)	352	xing	(1 2 3 4)
353	min	(2 3)	354	ming	(2 3 4)	355	nin	(2)

音码	拼音	包含的调号	音码	拼音	包含的调号	音码	拼音	包含的调号
356	ning	(2 3 4)	357	lin	(2 3 4)	358	ling	(1 2 3 4)
359	ding	(1 3 4)	360	ting	(1 2 3 4)	361	wu	(1 2)
362		(3)	363		(4)	364	diu	(1)
365	jiu	(1 3 4)	366	qiu	(1 2 3)	367	xiu	(1 3 4)
368	miu	(4)	369	niu	(1 2 3 4)	370	liu	(1 2 3 4)
371	bu	(1 2 3 4)	372	pu	(1 2 3 4)	373	du	(1 2 3 4)
374	tu	(1 2 3 4)	375	chu	(1 2 3 4)	376	shu	(1 2)
377		(3 4)	378	zhu	(1 2)	379		(3 4)
380	cu	(1 2 4)	381	su	(1 2 4)	382	zu	(1 2 3)
383	gu	(1 2)	384		(3 4)	385	ku	(1 3 4)
386	hu	(1 2)	387		(3 4)	388	mu	(2 3 4)
389	nu	(2 3 4)	390	lu	(1 2 3)	391		(4 5)
392	ru	(2 3 4)	393	fu	(1)	394		(2)
395		(2)	396		(3)	397		(4)
398	wa	(1 2 3 4 5)	399	shua	(1 3 4)	400	zhua	(1 3)
401	gua	(1 3 4)	402	kua	(1 3 4)	403	hua	(1 2 4)
404	wai	(1 3 4)	405	chuai	(1 3 4)	406	shuai	(1 3 4)
407	zhuai	(1 3 4)	408	guai	(1 3 4)	409	kuai	(3 4)
410	huai	(2 4 5)	411	wei	(1 2)	412		(3)
413		(4)	414	dui	(1 4)	415	tui	(1 2 3 4)
416	chui	(1 2)	417	shui	(2 3 4)	418	zhui	(1 4)
419	cui	(1 3 4)	420	sui	(1 2 3 4)	421	zui	(1 3 4)
422	gui	(1 3 4)	423	kui	(1 2 3 4)	424	hui	(1 2)
425		(3 4)	426	rui	(2 3 4)	427	wan	(1 2 3 4)
428	duan	(1 3 4)	429	tuan	(1 2 3 4)	430	chuan	(1 2 3 4)
431	shuan	(1 4)	432	zhuan	(1 3 4)	433	cuan	(1 2 4)
434	suan	(1 4)	435	zuan	(1 3 4)	436	guan	(1 3 4)
437	kuan	(1 3)	438	huan	(1 2 3 4)	439	luan	(2 3 4)
440	ruan	(3)	441	wang	(1 2 3 4)	442	chuang	(1 2 3 4)
443	shuang	(1 3)	444	zhuang	(1 3 4)	445	guang	(1 3 4)
446	kuang	(1 2 3 4)	447	huang	(1 2 3 4)	448	weng	(1 3 4)
449	wen	(1 2 3 4)	450	dun	(1 3 4)	451	tun	(1 2 3 4)
452	chun	(1 2 3)	453	shun	(3 4)	454	zhun	(1 3)
455	cun	(1 2 3 4)	456	sun	(1 3)	457	zun	(1 3)
458	gun	(3 4)	459	kun	(1 3 4)	460	hun	(1 2 4)
461	lun	(1 2 4)	462	wo	(1 3 4)	463	duo	(1 2 3 4)
464	tuo	(1 2 3 4)	465	chuo	(1 4)	466	shuo	(1 4)
467	zhuo	(1 2)	468	cuo	(1 2 3 4)	469	suo	(1 3)
470	zuo	(1 2 3 4)	471	guo	(1 2 3 4)	472	kuo	(4)

音码	拼音	包含的调号	音码	拼音	包含的调号	音码	拼音	包含的调号
473	huo	(1 2 3 4)	474	nuo	(2 4)	475	luo	(1 2 3 4 5)
476	ruo	(4)	477	yu	(1)	478		(2)
479		(3)	480		(4)	481		(4)
482	ju	(1)	483		(2 3)	484		(4)
485	qu	(1 2)	486		(3 4 5)	487	xu	(1 2)
488		(3 4 5)	489	nv	(3 4)	490	lv	(2 3 4)
491	yue	(1 4)	492	jue	(1 2)	493		(3 4)
494	que	(1 2 4)	495	xue	(1 2 3 4)	496	nue	(4)
497	lue	(3 4)	498	yuan	(1 2)	499		(3 4)
500	juan	(1 3 4)	501	quan	(1 2 3 4)	502	xuan	(1 2 3 4)
503	nuan	(3)	504	yun	(1 2 3 4)	505	run	(4)
506	jun	(1 4)	507	qun	(1 2)	508	xun	(1 2)
509		(4)	511	dia	(3)			
				lia	(3)			
				dei	(3)			
				tei	(4)			
				zhei	(4)			
				shei	(2)			

4 结束语

语音识别系统输出一个模糊阵列,需要程序自动判断哪些音可以组成词,并由此作出引导处理。由于模糊音的候选集可能很大,音的各种组合数将是一个天文数字。本发明的音码矩阵和分离结构的词库可以高效地迅速排除掉不可能的组合,并进一步得到包括语义在内的词汇信息。音码矩阵还可容易地实现双向搜索同效率。

本发明的音序码以及从语音到词库的快速搜索方法,其特征在于:以语音为入口线索,制定寓音形于一体的汉字内部编码——音序码;以音码为基础,构造音码矩阵。表示拼音的音码和表示声调和字形的序码,确定汉字音与形的一一对应关系。在分离结构词库的支持下,通过对音码矩阵的查询,快速确定矩阵中某一元素对应两音的组词特征,从而保证音-词转换的高效性。

1995 年

关于词汇知识表示框架的设计与实现*

刘 志 文

（中国科学院声学研究所，北京 100080）

HNC（概念层次网络）理论是中国科学院声学所黄曾阳研究员创立的。词汇知识表示框架是该理论中最富特色的内容之一。人工填写的可把握性和程序运作的可实现性是这一框架结构的生命力所在。依此所建立的各类知识库，均以有利于多义选一处理、实现模糊消解为主要目的，成为 HNC 理解处理技术的知识来源。

1 词汇知识框架的基础

词汇知识表示必须有层次，有分工，任务明确。

要划分概念知识与语言知识，语言知识与常识性知识，可自学知识与框架知识的界限，分别在概念、语言和常识三个层面建立相应的框架知识库。

要强化概念知识和语言知识的分工。词汇知识库必须以概念知识库为先导，以常识知识库为后援，不宜将三者混同于一体。

要明确词汇语言知识的服务目标。当前的首要服务目标是自然语言五重模糊的消解。依据服务目标选定框架知识的项目，框架项目及其基本内容必须先教给计算机，在这一阶段，不能企望于计算机自学习。

2 需要重建两个符号体系

——要重建概念表示的符号体系

在这个符号体系里，概念之间的基本关联性应在符号自身的表示结构中得到充分体现，并有高中低三个层面的明确分工：高层面表达概念集合的类型，并蕴涵概念集合之间的交式及链式关联性；中层面表达概念的对偶、对比及包含特性；低层面表达概念的个性，中低层面表示可多次重复。HNC 的三个概念层次网络（基元概念，基本概念，语言逻辑概念）是对这一符号体系的奠基性探索。

---

* 本文系作者 1997 年 8 月 16 日在全国第四届计算语言学联合学术会议（JSCL '97）上作的专题发言。

——要重建语句构成表示的符号体系

传统的语句构成表示方式不是语义表述,基本不反映人类语言感知过程的语言信息处理特征。重建的目的就是通过对语言感知过程的模拟建立语句构成的语义表述模式,HNC 的句类理论是对这一表述模式的探索。根据 HNC 理论关于作用效应链的基本假设,自然语言的语句可划分为 7 个基本句类和 36 个二重混合句类。7 个基本句类是:作用句、过程句、转移句、效应句、关系句和状态句,最后是判断句。前 6 个基本句类的两两混合(6 × 5 = 30),再加上它们与判断句的混合,共形成 36 种二重混合句类。每一基本句类各有确定数量的子类,例如作用句有五个子类,它们是:一般作用句、承受句、反应句、免除句和约束句。

这些句类及其子类都有自身所必需的主语义块,主语义块的个数和每一主语义块的语义特征是句类的函数。这些主语义块是语句的基本构成。用它们可写出相应的语句构成表示式,这些表示式属于概念层面的知识。

主语义块有 4 种基本类型,命名为 E、A、B、C。E 是特征语义块,A、B 是对象语义块,C 是内容语义块。E 表示句类,7 种基本句类分别用 X、P、T、Y、R、S、D 表示,这 7 个字母是句类标记,混合句类以这些字母的连用表示,例如 XP 和 PX 分别表示作用过程句和过程作用句,XT 和 TX 分别表示作用转移句和转移作用句。子类信息用句类标记后的数字表示,例如 X1、X2、X3、X4 分别表示承受、反应、免除和约束。主语义块用句类标记字母(对特征语义块)或它们与语义块类型字母 A、B、C 连用构成,例如 X2、X2B、XAC、X2C 分别表示反应句的反应、反应者、反应引发者及其表现、反应者的后续表现等 4 种语义块。

主语义块的不同排列顺序形成语句的不同表述格式,例如:3 主块句可有 6 种表述格式,4 主块句可有 24 种表述格式。但不同类型的自然语言各有自己的约定或习惯,对约定顺序,所有主语义块都不加标记,而对非约定顺序,某些主语义块则必须添加标记,具体地说,就是在对象或内容语义块之间必须加切分标记。这些语义块标记由所谓虚词(相应于 HNC 的语言逻辑概念)来承担。不带语义块标记的语句称为标准格式,带标记的称为非标准格式。通常,每一句类仅有一种标准格式,个别情况有两种甚至多种标准格式。

每一句类的标准格式及其各种变形格式(即非标准格式)是概念层面的知识,但对变形格式的具体选择则是词汇层面的知识。汉语的有名例句‘鸡不吃了”,既有转移句的两种变形格式的模糊,又有多义词‘鸡’的多义模糊,这两种模糊可同步消解,消解的手段只能是上下文提供的语境知识。这里要指出的是:转移句的句类格式知识极有利于语境知识的运用。

上述语句和语义块表示式就是 HNC 理论关于语句块构成表示符号体系的要点。

3 词汇框架知识的基本内容

——句类格式知识

句类格式即语句表示式是句类知识的总纲。句类格式决定了主块的数量及每一主块的基本内涵,其他各项知识都是它的派生物。

句类格式的有关规则属于概念层面的知识,但格式的具体选定则是词汇层面的首要知

识,是词汇框架知识的第一项。

——语义块构成知识

人类的言语感知以语义块感知为切入点,这是 HNC 理论模拟言语感知的基本假设。汉语最小的语义块可以是一个汉字或一个音节,但最大的语义块可以由一个语句退化而来,因此,对语义块的构成,必须也只能在词汇层面予以尽可能详尽的描述。

语义块的构成有良性与非良性之分,上面给出的"反应引发者及其表现"语义块 XAC 就是一种非良性构成,因为这个语义块中的反应引发者 XA 和它的表现 XC 在一个语义块中的表述方式及排列顺序是不能事先确定的。但是,有些句类的语义块构成可以事先确定,并写出相应的构成表示式,这是语义块的构成的良性表现。例如一般作用句的对象语义块通常可表示为 B = XB + YB + YC 的形式,其中 XB 是作用对象,YB 是效应对象,YC 是效应内容,这三项块素各有自身的概念关联性。这些知识对于模糊的消解往往具有一针见血的功效。

语义块在非标准格式时会出现分离现象。例如"张三打断了李四的腿"这个一般作用句,其对象语义块中分别有作用对象"李四"和效应对象"腿",在标准格式里它的各个语义块都比较规范。但如果采用非标准格式,例如"李四被张三打断了腿",这时对象语义块就分离了。三主块的一般作用句出现了反常的"四个"语义块。对于这一类的语义块分离现象必须在词汇层面给出信息,是词汇框架知识的重要组成部分。

——主辅语义块转换知识

除了主语义块之外,一个句子通常还要配置若干辅语义块。辅语义块有 7 个基本类,它们是:方式辅块、工具辅块、途径辅块、比照辅块、条件辅块、因辅块、果辅块(或因果辅块)。这 7 类辅块中,除条件辅块的某些表述(如时间条件)外,不论它在句子里的位置如何,都需要配置辅块标记,这一点可视为辅块与主块的区别特征之一。

但是,主辅语义块之间并不存在不可逾越的界限,同一内容的不同表述方式(指传统语言学中的"陈述、疑问、祈使、感叹"4 种表述方式)会伴随主辅语义块之间的转换,这是众所周知的语言现象。从语言理解处理来说,更重要的是同一表述方式下,不同表述格式也会出现主辅语义块的转换,特别是主块向辅块的转换。对这一转换现象的信息提示,也是词汇框架知识的组成部分。

——句类转换知识

同一内容不仅有不同的表述方式和表述格式,还可以采用不同的句类,这就产生了句类转换的现象,例如,一般作用句和被动承受句的相互转换,反应句向一般承受句和一般作用句的转换。句类转换的规则属于概念层面的知识,但具体转换的实现则属于词汇层面,也应纳入词汇框架知识。

——E 语义块构成知识

E 语义块的构成知识应予以特殊表述,它具有下列一般表示式:

$$E = QE + EQ + EH + HE$$

式中,QE 表示 E 要素的时态、情态、势态和性态特征;HE 表示 E 要素的基本概念(时、空、

数、量、质、度)特征 ;EQ 与 EH 是 E 要素的高低概念搭配表示,或一般与特殊概念的搭配表示 ;除上列四项构成外,汉语的 E 块还有助词成分 hE 或 qE。

上列 E 块的块素除 QE 之外,都需要在词汇层面予以表述,其中的 EQ 和 EH 更是汉语词汇框架知识的重点,因为汉语在非标准格式时,经常对 E 块采用高低概念搭配的表示方式。值得指出的是,EQ 和 EH 还可以再作‘ QH ’分解。

——块素的概念关联性

一个句子中的概念关联性有句内与句间、块内与块间之分,对块内与块间,关联性的表述都必须以块素为参照对象,关联的类型有启发性与强制性之分,前者可依靠人工输入,但后者则需主要依靠机器的自学习。

4 汉语词汇框架知识的两个特殊问题

上述知识表示框架,仅满足书面语言处理的需要。对汉语的语音处理,还需要增加音节知识库 ;对汉语书面语处理,则需增加单字知识库。这两个知识库是汉语的特殊需要。

汉语是单音节语言,尽管现代汉语趋向双音化,但音节感知仍是汉语的特色。一个音节通常包含若干汉字,每个汉字又包含多个义项,一个音节所对应的语义集合通常非常庞大。因此汉语音节感知的第一步不是直接进入义项感知,而是从义类感知入手。音节知识库的任务是以义类感知进行引导,以便迅速对单字词位置作出假设。起引导作用的义类目前分为以下八项:

1. 作为语义块切分、组合标志的逻辑概念
2. E 语义块的逻辑型 QE 知识
3. 基本命名概念
4. 构造新词的活跃语素
5. 数词
6. 量词
7. 基本概念
8. 动词

单字知识库的任务分工主要是对每个汉字能够独立使用的义项给以标注,标注的形式、内容与非单字知识框架类似,即单字动词的句类知识。但除此之外,单字知识库的一项特殊使命是为新词和伪词的辨识提供相应的知识。这项知识的表示与音节知识库相呼应。

The Perception Processing in Chinese Understanding

Quan ZHANG(张全)　Dinghua GUAN(关定华)

(State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences
P.O. Box 2712, Beijing 100080, China)

ABSTRACT

The technology of spoken language processing utilizes the technology of the speech understanding processing more and more in order to enhance the processing ability. This paper introduces a perception processing method about Chinese. It is based on a novel theory of natural language understanding processing--Hierarchical Network of Concepts theory. The perception processing completes chunks perceiving and Sentence-Category recognizing. The way of realization is hypothesis-test. An experimental system has been developed by this method. Although this system is just an experimental system, we consider that there is not any difficult obstacle to use this theory to process natural language in computers.

1. INTRODUCTION

The speech technology (speech recognizing and speech synthesis) has reached a high level. It needs the technology of understanding processing to enhance the ability of processing. The semantic plays an important role for understanding processing. Professor Zengyang HUANG founded a novel theory for natural language understanding processing based on semantic[1,2]. This theory is named Hierarchical-Network of Concepts (HNC) by its feature.

The chunk is a very important concept of HNC. In the view of HNC, the chunk perception and Sentence-Category (SC) analysis are the two basic processing modules of understanding processing. The result of SC analysis is the input of other processing modules based on HNC, such as, rationality analysis of sentence, context generation, hidden information revealing, topic and gist analysis. The realization of chunks perception determines whether the HNC theory becomes the HNC understanding processing technology.

The chunk perception is difficult. If the input is an ambiguous set (as the result of speech recognition), the chunk perception is more difficult. There are many probable routes of composing chunk. The longer the input, the more the routes. The chunk perception needs SC knowledge to guide it. But

only the Sentence-Category is determined, the SC knowledge can be used. The chunk perception is the foundation of the determining Sentence-Category. This is a circularity of using knowledge. This situation is like that, a hen laid an egg, and a chicken is hatched from an egg, which step is the first step, it forms a "hen-egg" circularity.

The main idea of this paper is to introduce a way to break the circularity.

2. THE FOUNDATIONS OF CHUNKS PERCEPTION

The HNC theory offers the theoretical foundation for the perceptive processing of chunks[1, 2]. In order to perceive chunks, the HNC knowledge bases are necessary. They are the foundations of chunks perception.

2.1 The theoretical foundation of chunks perception

HNC classifies the chunks into two types: the main chunks and the supplement chunks. The main chunks are the main parts of a sentence. The supplementary chunks provide the background knowledge of a sentence.

There are four kinds of main chunks according HNC[2], i.e., Agent (A), Object (B), Content (C) and Feature (E). Supplementary chunks are classified as follows[2]: Condition (Cn), Means (Ms), Instrument (In), Ways (Wy), Reference (Re), Premise (Pr) and Result (Rt).

The theory of HNC classifies the basic sentence-categories into seven types by the primitive concepts[2]. The names of the seven types are Action, Process, Transfer, Reaction, Relation, State and Judgment. Each Sentence-Category has its own knowledge. This knowledge includes the knowledge of the chunks: format knowledge, priority concepts knowledge and constitution knowledge. The knowledge determines what kinds of chunks should appear in what sequence, and what kinds of concepts have priorities to a definite chunk in a Sentence-Category. Chunks are functions of a Sentence-Category.

2.2 The HNC knowledge bases

The HNC theory classifies the knowledge involved in natural language into three layers, the conceptual layer, linguistic layer and common sense layer. The conceptual layer knowledge embodies the correlation of abstract concepts, and gives the association clue among concepts. It does not consider the character of a specific language. The linguistic layer need embody the character for one specific language. The common sense layer gives expression to the special knowledge.

The HNC knowledge bases are organized according the three layers. We designed a geographic common sense knowledge base[4], and use it to answer some geographic questions[5]. Now, in conceptual layer, we have designed conceptual correlation knowledge base and SC format base. These knowledge bases offer three kinds of chunks' knowledge in the conceptual layer. We have designed a series knowledge base about mandarin Chinese, such as vocabulary knowledge base[3], Chinese character

knowledge base, syllable perception knowledge base. (Each Chinese syllable has its own independence meaning. Many chunks' flags in Chinese use single syllable words.) These knowledge bases provide the three kinds of chunks' knowledge about mandarin Chinese.

The knowledge of conceptual layer and linguistic layer plays the more important roles in understanding processing.

3. CHUNK PERCEPTION PROCESSING

How to use a computer to perceive chunks is a key processing step when HNC is applied to understanding processing. We find a way to complete the perception and to determine the SC. The basic operation is "hypothesis-test".

We design an experimental system according the way of chunks perception processing chunks[6]. The diagram of how to perceive chunks and to determine the SC is shown as Fig.3.1.

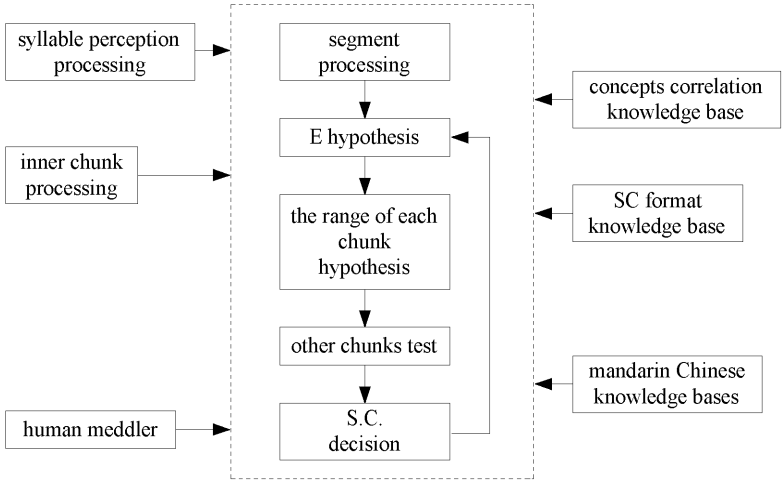


Fig. 3.1 The diagram of Perceptive Processing

3.1 E chunk hypothesis

The input of the experimental system is Chinese phonetic alphabet without tone. When a sentence input the system, the system first matches the vocabulary knowledge base to form a segment (continuous words' string). In a segment, the syllable perception processing finds the position of single syllable words by the syllable knowledge base. The segment processing finishes. The E-hypothesis starts.

The E chunk always embodies the SC information of the sentence. If the E chunk can be determined, the other chunks can be determined. The key of perceptive processing is the E chunk perception.

In order to embody tense and mood, there are usually some language logic concepts[1] around the

E of the sentence. We can assume the E of the sentence according the concept attribute of words and the logic concepts. After E-hypothesis, we can use the SC knowledge offered by the assumed E to test other chunks.

3.2 Other chunks test

The SC knowledge includes standard format knowledge. The format information points out the chunks' positions. The position information is the clue of other chunks test. If the sentence deviates from the standard format, the logic concept always appears before the chunk. According the standard format and logic concepts, we can hypothesize the range of each chunk.

In the each hypothesis range, the system detects if there are the priority concepts to a chunk by the priority concept knowledge of chunks. Then, it tests the constitution of the chunk by the chunks' constitution knowledge of chunks.

If the system finds all the chunks of a Sentence-Category in the suitable position, with the priority concepts and according the constitution of chunks, the sentence-category decision will approve this Sentence-Category. The system starts the inner chunk processing. If there is not any conflict in a chunk, the sentence-category is confirmed. The system completes the chunks perception and SC recognition. Else, the system will start another E-hypothesis.

If there is not a Sentence-Category passing the test well, the system will ask human to meddle.

3.3 Examples of processing

Here is an example to demonstrate the processing. The input Chinese phonetic alphabets are shown as follows:

· wo men xiang shang ji hui bao guo zhe ge qing kuang.

(We have reported this news to the leader.)

· 我们 向上* 汇报* 情况*

上级*

The result of matching the vocabulary base is shown. The "*" indicates that there are other candidate words: (the other words in the braces)

· 情况 {轻狂} 向上 {相商}

· 汇报 {回报} 上级 {简计}

The three syllables "xiang shang ji" form a segment. The system assumes the single syllable word in the "xiang" position.

We can assume the E is "hui bao" (in Chinese character is **汇报**, in English is Report). Then, we know that the meaning in HNC symbols is v923, the SC is Transfer (T), the SC format is TA + T + TB + TC, the priority concept of TA is p (HNC symbol for people), TB is p and TC is gw or g23 (gw and g23 for information).

The meaning and the position of "**我们** (in HNC symbol is p)" is suitable for TA. The TB

should follow the E, however,"向（102）"is the flag of TB, then"上级（p）"is suitable for TB. Behind E,"情况"is suitable for TC. Here, all the other chunks pass the test, the E is"汇报".

Then, we can translate the Chinese phonetic alphabet into Chinese character as follows:

我们（TA）向上级（TB）汇报过（T）这个情况（TC）。

Here are some other examples:

jiang jie shi tong zhi l zhong guo.

蒋介石(A) 统治了(X) 中国(B)

gan jin ba zhe ge xiao xi tong zhi ta men.

赶紧 把这个 消息(TC) 通知(T) 他们(TB)

fang xiang chao zhe nan mian.

方向(SB) 朝着(S) 南面(SC)

zhang tong zhi chao l 1 ge cai.

张同志(A) 炒了(X) 1个菜(B)

4. CONCLUSION

We use some corpora to test this experimental system, when the SC of the input belongs to the basic SC, and the chunks are not too complex, the result is in keeping with the expected well.

We are doing some work about the complex chunks with the experimental system, such as clause molting chunks, chunks with"有"or"是"which are active elements of composing chunks and the chunks containing multiply modifier without"的"which is the modifier flag. The result will be written in other papers. However, we do not consider that there are some difficult obstacles in theory to solve these problems.

5. ACKNOWLEDGMENTS

We would like to thank those people who join this work or provide valuable suggestions: Mr. Zhiwen LIU, who does a lot of work of knowledge acquisition; Ms. Yanling Du, who accomplishes the input interface and programs some functions for E hypothesis; Ms. Yu MAO, who offers the various knowledge bases' interfaces in this work; Mr. Yaohong JIN, who programs many fundamental functions in this work and undertakes the main part of inner chunk processing and Mr. Ling SHEN, who gives many useful suggestions.

REFERENCES

[1] Zengyang HUANG et al. The Fundamental Structure and Feature of the Natural Language Semantic Network. Collection of HNC, China. 1996

[2] Zengyang HUANG et al. The Deep Structure of the Natural Language and Sentence - Category Analysis. Collection

of HNC，China．1996

[3] Quan ZHANG et al．A Scheme of Modern Chinese Words＇Knowledge Base．JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE，1995，17：(Sup．162)．China

[4] Quan ZHANG et al．The Application of Object-Oriented Technique on Geographical Knowledge Base．Collection of ＇95 National Workshop on New Technology on Computer．China．1995

[5] Quan ZHANG et al．Sentence-Semantic-Class (SC) Analysis in Chinese Interface System on Geographical Knowledge．Proceedings of National Symposium of Phonetics，Picture and Communication．China．1995

[6] Quan ZHANG．Perceptive processing of chunks based on HNC．Ph．D．Dissertation of Institute of Acoustics，Chinese Academy of Sciences，China．1996

编者按：本文选自《 Proceedings of 1997 China-Japan Symposium on Advanced Information Technology**》。它在** HNC **理论走向工程化的道路上进行了拓荒性的探索，迈出了最先的一步。**