

第二届 HNC 与语言研讨会论文集

CESE:={CP1, ICP1, BCP2, ICP2, BCP2, ...}

(HNC1)

SC={GBK1+EK+GBKm(m=2-4)}

(HNC2)

SGU={DOM, SIT, BACE, BACA}

(HNC3)

ABS={BCN/BCD, XYN,XYD,PT,RS, BACEnt, BACAm}

(HNC4)

苗传江 杜燕玲 主编



海洋出版社

第二届 HNC 与语言学研讨会论文集

苗传江 杜燕玲 主编

海洋出版社

2004 年·北京

内 容 简 介

本书是由第二届 HNC 与语言学研讨会的论文汇编成的,所收录的论文展现了 HNC 理论的最新进展以及运用该理论进行语言研究和处理的最新成果,其内容主要包括 HNC 理论的反思与发展、特约报告、语言现象的 HNC 阐释、知识库和语料库建设、机器翻译引擎原理研究、句类分析技术及其应用等六个方面。HNC(Hierarchical Network of Concepts,概念层次网络)是由中国人原创的自然语言理解处理方面的理论体系,是中文信息处理的三大流派之一。本书的读者对象包括自然语言处理、语言学、人工智能等领域的研究者和开发者。

图书在版编目(CIP)数据

第二届 HNC 与语言学研讨会论文集 / 苗传江, 杜燕玲主编. — 北京: 海洋出版社, 2004. 9
ISBN 7-5027-6194-2

I. 第… II. ①苗… ②杜 III. ①机器翻译-学术会议-文集 ②自然语言理解-学术会议-文集 IV. ①H085-53 ②TP18-53

中国版本图书馆 CIP 数据核字(2004)第 094080 号

责任编辑: 阎 安

责任印制: 严国晋

海洋出版社 出版发行

<http://www.oceanpress.com.cn>

(100081 北京市海淀区大慧寺路 8 号)

北京蓝空印刷厂印刷 新华书店发行所经销

2004 年 9 月第 1 版 2004 年 9 月北京第 1 次印刷

开本: 787 mm × 1 092 mm 1/16 印张: 33.625

字数: 820 千字 印数: 1~300 册

定价: 120.00 元

海洋版图书印、装错误可随时退换

第二届 HNC 与语言学研讨会

主办单位：中国科学院声学研究所语言知识处理(HNC)实验室

北京大正语言知识处理研究院

组 委 会：张 全 陈小盟 晋耀红 苗传江 杜燕玲

秘 书 长：杜燕玲

副秘书长：刘向宇 宋 晓

日 期：2003 年 9 月 20—22 日

地 点：北京市海淀区客座公寓

前 言

由中国科学院声学研究所黄曾阳先生创立的 HNC(概念层次网络)理论,是一个关于自然语言理解处理的理论体系,自诞生以来,一直在理论探索和应用开发中不断发展和完善,至今已有十年的历程。

第二届 HNC 与语言学研讨会于 2003 年 9 月 20—22 日在北京召开,中国科学院高科技局桂文庄局长、中国科学院声学研究所所长田静到会祝贺并作了重要讲话。教育部语言信息管理司李宇明司长莅临会议指导并作了题为《语言学与语料库的深度开发》的重要报告。中国计算语言学先驱者之一鲁川先生,从语言学与 HNC 融合的角度,作了题为《电脑理解人类语言需要理论创新》的重要报告。北京大正语言知识处理研究院陈小盟董事长介绍了研究院利用 HNC 理论和技术研发应用系统和产品的最新进展,并传达了许嘉璐副委员长对研讨会和 HNC 探索的殷切期望。本届研讨会共有 80 多位来自全国各地的代表参加,共有 67 篇论文在研讨会上进行了宣读和讨论。

会议第一天,黄曾阳先生作了题为《在反思中前进,在碰撞中成长》的大会报告,详细介绍了 HNC 理论探索近五年来的基本成果,描绘了 HNC 未来 15 年探索的宏伟目标。现在整理成四万余字的论文,第一次全面系统地阐述了 HNC 的理论框架,明确宣布 HNC 理论框架原定的五模式说已改为四模式说,给出了四模式的数学物理表示式,这标志着 HNC 预定建立的全部语言理论模式已告完成。论文将 HNC 的探索目标正式定位于交互引擎的研发,(交互引擎是对人类交际引擎的仿制),并详细阐述了 HNC 理论和技术规划中的三项理论工程、四项技术工程和一项特定基础工程。标志着 HNC 从此将进入一个以理论探索为辅、以应用研究为主的新时期。

研讨会期间,与会的专家和学者肯定了 HNC 近几年来取得的基本成果,并对 HNC 的现状 & 发展提出了很多非常宝贵的意见和建议,概括起来主要有下面四点:(1)HNC 是原创性的,应该坚持自己的特色;(2)HNC 应该努力走向世界,使国外同行了解中国人的探索成果;(3)HNC 的符号体系应该尽量考虑与传统语言学接轨的需求,以方便理解和交流;(4)HNC 的理论研究和应用开发要并重,要不断地推出更多的阶段性成果。

会后,作者根据会上的讨论结果对论文进行了适当的修改,现在结集出版,以便于更广泛、更深入地交流。论文集的主要内容包括五个方面:(1)特邀报告;(2)语言现象的 HNC 阐释;(3)知识库和语料库建设;(4)机器翻译引擎原理研究;(5)句类分析技术及其应用。

自 2001 年初以来,HNC 在两个方面相辅相成地并行发展。一方面,以中国科学院声学研究所为主体,扩展和深入理论探索,并加强面向应用的基础研究和系统建设。另一方面,以北京大正语言知识处理研究院为主体,努力开发实用的系统和产品,创造经济效益和社会效益。这次研讨会的论文大致体现了 HNC 在以上两个方面的最新进展。

相信这本论文集的出版必将会加快 HNC 三项理论工程、四项技术工程和一项基础工程的正式启动,促进 HNC 与语言学界、信息处理学界等专家学者的交流与合作,共同推动 HNC 在语言学和信息科学中的应用及产品开发。

编 者

2004 年 2 月

目次

□ 特约报告

在反思中前进,在碰撞中成长	黄曾阳(3)
语料库中语言知识的标记问题	李宇明(40)
词的具体意义对句子意思理解的影响	陆俭明(48)
当前自然语言处理研究的一些看法	姚天顺(54)
电脑理解人类语言需要理论创新	鲁川(58)
HNC 理论的创新性与实用性	吴泓缈(64)

□ 语言现象的 HNC 阐释

“怀疑”句的功能类型、认知性质与 HNC 思考	萧国政 郭婷婷(73)
HNC 说“的”	池毓煊(79)
“得”字结构的语义类型及特点	唐兴全(86)
基于 HNC 理论的句法结构歧义消解	张克亮(94)
“给”的多义模糊及其消解	白晓革(103)
浅析“得”字句的 HNC 理解	郝惠宁 雒自清 杜燕玲(108)
语言的句子性质与 HNC 的句类分析	柯航(114)
HNC 理论对汉语被动句的理解与处理	游舒(120)
从 HNC 看主谓句和主谓短语	张全(125)
浅析“块”在自然语言理解中的作用	张运良(132)
句蜕中的语义块分离	张玲(138)
面向 NLP 的定语分类研究	熊亮(145)
名量和动量短语的构成与识别	宋晓(150)
现代汉语“以 A 为 B”格式的理解	冯霞(155)
现代汉语配对形式的功能类型	刘一飞(160)
HNC 理论下对具有模糊的词和新词小议	孙雄勇(169)

□ HNC 知识库和语料库建设

HNC 的知识库系统建设	苗传江(175)
HNC 的字知识库	苗传江 唐兴全 刘智颖(182)
对偶性概念的再思考	李 颖 池毓焕(187)
语义块分离现象与语句格式和 EK 构成的关系	杜燕玲 雒自清(200)
动补关联词对句类的影响	刘智颖(208)
简明状态句的句类知识	骆小菊(215)
“坚持”句的句类知识	张 玲(225)
浅识省略句	张 全(231)
“有”字和“是”字存在句三个世界的比较及 HNC 分析	彭淑莉(237)
方式、工具、途径辅块的辨识	雷 静 张 易(244)
否定语式及其指向	朱英姿(249)
现代汉语文言语法残留现象考察	张 易 雷 静(258)
HNC 知识库的句群、语境研究	雒自清 杜燕玲 郝惠宁(270)
HNC 语料库标注体系浅识	张 全(282)
HNC 语料库的作用	张 全(289)

□ HNC 机器翻译引擎原理的研究

原型句蜕的汉英机器翻译初探	李 颖(299)
多元逻辑组合的汉英对比初探	池毓焕(308)
试论原型句蜕到多元逻辑组合的变换	池毓焕 李 颖(314)
汉英机器翻译中的辅块及多元逻辑组合问题	雒自清 张雪荣 郝惠宁(323)
HNC 承受句的汉英句类转换	张克亮(331)
作用效应句在汉英翻译中的句类转换	熊 亮(342)
汉英翻译中因果句的句类转换	熊 亮(350)
汉英翻译简明状态句的句类转换研究	孙雄勇(357)
汉英翻译中一般转移句格式转换	孙雄勇(362)
过渡处理系统的模块设计	韦向峰(368)
过渡处理系统规则的形式化	韦向峰(374)
汉英机器翻译系统中动词时态的处理	孙广范(380)
寻找理性法官	张 全(386)

□ HNC 句类分析技术于应用

HNC 语言理解技术及其应用	晋耀红(397)
汉语分析中动词/名词兼类的处理	孙广范(405)
汉语动词连见的类型与分布	宋 晓(411)
从“坚持”看动词连见	韦向峰(417)
汉语存现句的分析和处理	白晓革(423)
汉语分析中并联结构的判定	孙广范(430)
基于属性和规则的未登录词识别	孙广范(437)
浅识 HNC 的解模糊思路	谭 露(442)
一个基于语义的智能信息过滤算法	晋耀红 邬 郑(448)
HNC 领域分类系统的设计思想与工程实现	邬 郑 晋耀红(456)
基于句类的因特网语言知识处理	张 全(465)

□ 其 他

俄语文本信息获取的通用接口:目标、模型和应用	傅兴尚(479)
一种基于 TM 技术的机助翻译系统的语义处理策略 易绵竹 南振兴 武 斌(487)	
汉语詈语的结构方式及其文化蕴涵	罗耀华 向婷婷(494)
面向中文信息处理的现代汉语“n+n”结构歧义问题研究	杨 泉 冯志伟(507)
一种汉语语料逐词索引工具的设计方法	萧国政 胡 慊(516)
个人语料库的构建原则及 Excel 检索方法的实现	黄慧英(523)

特约报告 ☐

在反思中前进,在碰撞中成长

黄曾阳

中国科学院 声学研究所,北京 100080

摘要 本文是《HNC(概念层次网络)理论》的续篇,概述了 HNC 理论探索近五年来的基本成果,对 HNC 的某些传统论述进行了反思与改正,同时给出了 HNC 未来 15 年探索的素描。本文正式将 HNC 的探索目标定位于交互引擎的研发,将语言概念空间描述的原五模式说改为概念基元空间、句类空间、语境单元空间和语境空间的四模式说,给出了这四个模式的数学表示式;将交互引擎的基本模块凝练成句类分析、语境单元萃取和语境生成的三部曲或三个基本环节。为达到交互引擎的彼岸,本文阐释了这一科学征途的三项理论工程、四项技术工程和 HNC 研发平台的概貌。为说明这一彼岸的理论可行性,本文诠释了 HNC 的两个基本假设和“概念无限而概念基元有限、语句无限而句类有限、语境无限而语境单元有限”的基本论点。

1 引言

HNC 理论基本构架的形成已经整整 10 年了。

HNC 这 10 年的历程是一个幸运的历程。

小托尔斯泰有一部不朽名著叫《苦难的历程》,该名著三部曲的名称是《两姐妹》、《一九一八年》和《阴暗的早晨》。HNC 的幸运历程也有一个三部曲,可类比定名为《两仙翁》(按:指语法与逻辑或语言学与哲学)、《一九九七年》和《明媚的晨曦》。

小托尔斯泰完成那“苦难历程”的写作时间是在小说描述事件的 20 年之后,如果以这个时间为参照,那么 HNC 幸运历程的描述放到 2023 年以后比较适当,是今天座中年轻才俊进入“知天命”或“从心所欲”的时候,你们到那个时候看着办吧。

但是,我今天还是要说一点,HNC 幸运历程的最大幸运是什么?那就是遇到了众多的伯乐。由于近年我成了一个经常阅读圣经的无神论者,所以他们成了我心中的科学而非神学的弥赛亚。现在,让我们以热烈的掌声向这些科学弥赛亚表示我们的无限敬意与感谢。有两位科学弥赛亚——陈力为院士和林杏光教授已经离开我们了,让我们以肃穆的沉思表示我们永恒的缅怀。

中国历史最辉煌的年代是唐朝,唐朝最伟大的文豪兼哲人是韩退之(愈)先生,退之先生最富于哲理的名篇之一是《进学解》,《进学解》里最著名的警句之一是:“业精于勤,荒于嬉;行成于思,毁于随。”从某种意义上说,本文是学习《进学解》的一点体会。

我的体会之一是,退之先生说的“思”就是哲学意义的反思,“随”就是现代意义的拒绝创

新与改革。“随”的对偶概念是“撞”，无“思”无“撞”就不能前进与成长。所以本文定名为《在反思中前进，在碰撞中成长》。

世俗意义下的反思仅对内而不对外，碰撞仅对外而不对内。本文的反思和碰撞则兼对内外，且都以对内为主。

不过，本文并不直接以反思与碰撞为中心展开论述，而是把两者融合到关于自然语言之电脑理解的以下三个基本问题的阐释里。这三个基本问题是：

(1)为什么已经具有许多奇妙智能的电脑到今天还不能理解自然语言？这里的根本障碍到底是什么？怎样克服这一障碍？

(2)假定电脑能够理解自然语言，那么，让它成为一个“一目千行，过目不忘”的超级才子（语言超人）存在什么障碍？怎样克服这一障碍？

(3)假定电脑能够理解自然语言，那必然有一个从低级到高级的成长过程，那么，这一过程主要阶段或基本步调是什么？我们怎样推动这些阶段或步调的实现？

围绕着第一个问题的阐释，本文将着重介绍 HNC 理论规划中的三项理论工程，简称三部《手册》——《HNC 概念基元符号体系手册》、《句类知识手册》和《语境单元知识手册》。

围绕着第二个问题的阐释，本文将简要介绍 HNC 技术规划中的四项技术工程——句类分析技术 SCA(Sentences Category Analysis)、语境单元萃取技术 SGUE(Sentences Group Unit Extraction)、语境生成技术 ABS(ABSTRACT)和翻译引擎技术。

围绕着第三个问题的阐释，本文将简略说明 HNC 的一项特定计算工程——计算机自然语言理解度的测试与学习平台，这个平台的首要目标是进行电脑理解度的量化与类化研究，将简称 HNC 研发平台。

如果把这三项阐释所范定的研究目标的实现比做一次重大战役，那么，第一项阐释关系到该战役的战略方针的确定，第二项阐释关系到主战场或主攻方向的选择，第三项阐释关系到关键信息的搜集与分析、先进武器的供应与保证。毋庸论证，这三者对于赢得一场战役的胜利具有决定性意义。

HNC 通过语言概念空间研究语言现象。语言概念空间是存在于人类大脑之中的一个符号体系，这个符号体系既是人类进行语言思维的载体，又是人类进行语言交际的引擎，这个符号体系将简称交际引擎。计算机要获得理解自然语言的能力就必须也拥有一台在功能上类似的引擎，否则计算机就永远不可能成为名副其实的电脑。这个引擎将叫做交互引擎。交互引擎是交际引擎的仿制，仿制就必然存在简化与不足。交际引擎由千亿数量级的神经元构成，交互引擎将由当前电脑的二进制数字符号构成。HNC 的基本定位就是催生交互引擎的问世，这是天方夜谭式的神话么？HNC 不持这种悲观态度。HNC 已经看到，如果能够实施上述三项理论工程、四项技术工程及其特定计算工程，则完全有可能在 15 年左右的时间里，使交互引擎在中华大地诞生。

2 关于交互引擎的三项理论工程——三部《手册》

本章是本文的重点。将分六节进行论述。2.1 节对 HNC 探索的重要先行者作简要回顾；2.2 节介绍 HNC 理论的基本定位及其两个基本假设；2.3 节论述 HNC 探索预定的三部《手册》，不仅是三项理论探索，而且是三项巨大的理论工程；2.4 节是本章的第一重点，介绍

第一部手册——《HNC 概念基元符号体系手册》的要点,重点阐释了 HNC 概念基元符号体系设计五项基本原则的延伸原则;2.5 节介绍第二部手册——《句类知识手册》的要点,对句类概念的 HNC 传统论述中的失误作了系统的反思;2.6 节是本章的第二重点,提出了交际语境和交互语境的概念,阐释了交互引擎研发的基本思路与框架,对 HNC 理论体系的整体描述进行了反思,将 HNC 理论体系的原五模式说改成四层级与三提升说,列举了语境单元萃取和语境生成研究的基本课题,概述了 HNC 理论体系四组表示式——(HNC1)、(HNC2)、(HNC3)和(HNC4)的物理意义。

2.1 对 HNC 先驱者的简短回顾——交互引擎的背景说明 1

已经具有许多奇妙智能的电脑到今天还不能理解自然语言,为什么?这个问题的答案很复杂,但是又似乎很简单,有一个简单答案是这样的:要理解人类语言就必须拥有世界知识,而世界知识浩瀚无垠,需要软件工程师在语言学家的配合下一点一滴地教给电脑,这个工程太浩大了。面对这一浩瀚无垠的困难,人们思考着并实际采取过许多对策,这些对策可以概括成三种基本类型:最大简化策、愚公移山策和最大似然策。

最大简化策的基本思路是:设定一个受限环境,让这个环境的世界知识十分有限,且语言表达十分规范,此策的代表人物之一是维诺格拉德教授。

愚公移山策的基本思路是:将世界知识划分出一定的类型和层次,用谓词逻辑形式或结构化数据形式写出来交给电脑,前者的代表人物是美国 CYC 计划的主持者 Lenat 教授,后者则代表人物众多。

最大似然策当前最走红,领唱者甚多,自称语料库学派。其核心思想是把言语当做一个随机过程,依托大规模真实语料制造一个计算机的进化环境,让计算机在这个环境里以大体类似于宇宙进化的方式自行进化。

前两种对策基本以所谓理性主义为依托,第三种对策基本以所谓经验主义为依托。

三种对策的基本思路朴素而简明,都具有朴素性的固有力量。

但是,对于“已经具有许多奇妙智能的电脑到今天还不能理解自然语言”这样重大的科学问题,仅仅依靠朴素性的固有力量是不够的,需要进行多侧面多层次的反思。我们至少需要反思以下三个基本问题:第一,为什么开始只有极少世界知识的 1~5 岁正常儿童都是学习自然语言的天才?第二,为什么儿童在完成这一学习的过程中,既不需要学习语法课程,也不需要学习逻辑课程?第三,那高深而且越来越高深的语法学和逻辑学与儿童在潜移默化中获得的基本语法和逻辑知识是否存在某些奇妙的区别?

有两位美国人——乔姆斯基先生和山克先生在 HNC 之前进行过这一反思,还有一位欧洲人维特根斯坦先生在电脑尚未出现之前就进行过这一反思,并为此写了《哲学研究》这一不朽巨著。

乔姆斯基先生对上列三个基本问题的前半部分作过刺猬式反思,山克先生则对三者都进行过狐狸式反思。乔先生的答案是:儿童的语言天赋是由于人类大脑里存在一个经过百万年进化而形成的东西,他把这个东西叫做普适语法(UG)。随后他用毕生精力研究这个东西,在 20 多年间写下了参考文献中列出的八部专著^[1~8]。山克先生的答案是:语言就是用概念基元写出的脚本,语法不过是一件人为的装饰品。这位先生以典型的狐狸研究风格搞了一种概念基元和若干个小脚本之后就转向其他相关领域了。

上列三位先行者的创新都体现了西方文化的优势特征:熟谙分析与归纳的逻辑。但是理论创新还需要综合与演绎的逻辑,这四位最高理性法官携手合作才能做出重大的创新,近代和现代物理学的重大创新都是四位理性法官携手合作的光辉典范,而我国的两部古老名著——《老子》和《孙子兵法》则是古代的光辉典范。

一个极为常见的认识误区是,这四位最高理性法官需要一位老大。这个老大观念在东方和西方文化里都具有挥之不去的魔性。实际上,这四位最高理性法官需要轮流坐庄,不能由某一位理性法官永远充当老大,就像桥牌游戏的东南西北四家不能由某一家老是坐庄一样。在 HNC 的形成过程中,演绎起过特殊重要的作用,但绝不是永远的老大,可是 HNC 经常流露出演绎老大的意识,这是首先需要反思的。因为 HNC 已经进入不再是演绎坐庄的新时期了,八项工程的实施都需要四位最高理性法官的通力合作。

2.2 HNC 的两个基本假设与电脑未能理解自然语言的症结——交互引擎的背景说明 2

自然语言理解是一个从自然语言空间到语言概念空间的映射过程,两个空间各有自己的一套符号体系。语言交际过程、即交际引擎的运作过程实质上是这两种符号体系相互映射的过程。说者//^① 写者将语言概念空间的符号映射成语言空间的符号,这是语言生成过程,听者//读者将语言空间的符号映射成语言概念空间的符号,这是语言理解过程。

语言空间符号体系千差万别,据说当今世界上还存在 6 000 种之多,但语言概念空间符号体系只有一个,全人类的语言概念空间具有同一性。这是一个假设,是 HNC 理论的第一假设。这个假设一定会引起质疑,因为不同文化背景和不同民族的语言概念空间必然存在差异。这里的质疑者不了解理论探索的基本特征,那就是必须有所舍弃——舍弃枝节,抓住要害,没有舍弃就不会形成任何理论。这就是黑格尔先生所说“哲学的开端就是一个假设”的真谛。拙著^[9]里写了一段不符合科学论文规矩的话,但由于关系到 HNC 第一假设是否合理这一根本问题,这里仍然全文引录下来供本文读者思考。

关于这个模型问题,可以说存在两种态度,一是得过且过,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,必须谨记有所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,拿着那些惟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。

① 符号//表示“和”或者“或”的情况,全文同。

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,根据自然的普遍合法性,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。

HNC 第一假设关系到对语言本体的认识,也关系到对“自然语言理解”的理解。依据这一假设,语言的本体就不是语法,而是语言概念空间。语言概念空间就是拙著^[9]里所说的概念联想脉络或语义网络。乔姆斯基先生把他本人精辟阐释的那个人类百万年进化的大脑产物叫做普适语法并不妥当,因为语法这个概念还是停留在语言空间,没有升华到语言概念空间。语言空间和语言概念空间是相互依托但又毕竟具有本质区别的两类空间,语言空间依托于语音和文字而存在,语言概念空间依托于概念联想脉络而存在。语音、文字、概念联想脉络都是符号体系,语音符号体系具有可听性(现代技术还能使之可见、可存储、可复现),文字符号体系具有可见、可存储、可复现性,概念联想脉络符号体系既不可听,亦不可见。脑科学研究的现代技术虽然正在朝着使之可观测、可分析的方向努力,但远没有达到像基因复制或蛋白质合成机制研究的技术水平,目前还得主要依靠演绎式探索“鸣锣开道”。

对 HNC 第一假设需要作两点重要补充,两者是对下述两个基本问题的回答。

第一,语言概念空间是概念空间的全部么?回答是:否!语言概念空间只是概念空间的一个子空间。概念空间的其他子空间还有形象概念空间、情感概念空间、艺术概念空间、科学概念空间等。

第二,语言概念空间具有历时性么?回答是:可以忽略!当然,某些概念子空间具有非常鲜明的历时性特征。如科学概念空间在以牛顿和爱因斯坦为标志的先后三个历史时期就具有天壤之别,人类社会物质文明的发展主要依靠这一“天壤之别”的出现。但语言概念空间的主体是稳定的,其历时性远不像语言空间那样突出,仅有某些局部呈现出极为缓慢的历时性。

在 HNC 第一假设的基础上,还需要引入第二假设——关于语言概念空间结构特征的假设。这一结构特征必须假设,因为,如同上面所指出的脑科学的当前研究水平还不能提供足够的启发性知识。

HNC 的第二假设是:语言概念空间是一个四层级——基层、第一介层、第二介层和上层——的结构体。基层对应着语言概念基元符号体系,也称概念基元空间;第一介层对应着句类符号体系,也称句类空间;第二介层对应着语境单元符号体系,也称语境单元空间;上层对应着语境符号体系,也称语境空间。这四层级符号体系及其对应的世界知识将分别简称概念基元世界、句类世界、语境单元世界和语境世界。空间对应着相应科学对象的数学描述,世界对应着相应科学对象的物理描述。这里有一个新术语——语境单元,这里的语境也有它特定的意义,都将在 2.6 节说明。

这个四层级概念空间和概念世界共同构成语言思维(不是全部思维)的载体,这个语言思维载体的名称就是概念层次网络,简称 HNC。为什么使用这么一个人们难以习惯的奇特名称呢?因为语言概念空间的基本特征是它的层次性和网络性。为了便于与传统语言学接轨,我们曾把概念层次网络叫做语义网络;为了便于一般读者的理解,我们也曾把它叫做概念联想脉络。

HNC 第二假设隐含着下述三项“无限与有限”的假设,这就是:“语言概念无限而语言概念基元有限,语句无限而语句的概念类型(句类)有限,语境无限而语境单元有限”。

如果在上述两项基本假设的基础上仅探讨概念层次网络的基本理论特征,那就是狐狸

式研究,如果锲而不舍地去具体构造出概念层次网络的四层级符号体系,那就是刺猬式研究。显然,概念层次网络的研究需要这两种研究方式的紧密配合,刺猬需要狐狸的智慧,而狐狸需要刺猬的执著。

语言概念空间的上述四层级符号体系一旦构造出来,而且该符号体系能够直接用二进制数字符号加以表示,那么,交际引擎的仿制(即交互引擎的研制)就不再是一个炼金术的幻想,而是一项意义深远的科学工程了。

结论是一清如泉的:仿制交际引擎的关键举措就是首先把体现语言思维载体的上述四层级符号体系装进电脑,并在电脑中形成一个四层级概念世界。这个载体当然有低级、中级与高级之分。人类大脑里存在着该载体的最完善、最高级形态,而现在的电脑连该载体的最简陋雏形都没有,这就是电脑至今不能理解自然语言的症结所在。HNC从1998年开始试图向电脑装入一个最低级的语言思维载体,2001年基本完成,2002年显示出初步效果,但其喜人之处不在于这些效果本身,而在于他//她所预示的成长潜力。今天晚上,晋耀红博士将向大家展示这一婴儿的憨态,明天晚上,苗传江博士将向大家介绍这一婴儿的知识结构,张全博士则将介绍对该婴儿的托儿所教育设想。

2.3 三部《手册》是一项巨大的理论工程,而不只是一个理论体系

为了叙述的方便,下面将把《HNC概念基元符号体系手册》、《句类知识手册》和《语境单元知识手册》分别简称《手册1》、《手册2》和《手册3》。《手册1》将描述概念基元符号体系及其相应的世界知识;《手册2》将描述句类符号体系及其相应的世界知识;《手册3》将描述语境单元符号体系及相应的世界知识。三部《手册》所描述的符号体系及其相应的世界知识相当于概念基元世界、句类世界和语境单元世界的“数字地图”。

概念基元世界大体对应语言空间的词语,句类世界大体对应语言空间的语句,语境单元世界大体对应语言空间的句群。这里的对应不可能是严格的,因为概念基元、句类和语境单元这三个世界所拥有的自身符号体系并非与语言世界的符号体系一一对应,而且对应性的具体表现与具体语言世界(语种)密切相关。“大体对应”是一个极为粗略的描述,确切的描述是三部《手册》的重要内容之一。

我在文献[10]中曾用“7-2-1”、“57-3192”和“7-57-7”这三组数字对上述基层与介层概念世界的宏观特性作了一个最高层次的概括。这里首先需要对这三组数字进行反思,本文将把这组数字改成下面的描述:“8-2-1”、“2-8-57-3192”和“10-57-7”。下面三节将分别说明对这三组数字变动的反思。

这里需要着重指出的是:三部《手册》不只是对基层与介层概念世界的理论阐释,还要描述对应于概念基元世界的“典型”词语、对应于句类世界的“典型”例句、对应于语境单元世界的“典型”句群;描述捆绑于每一概念符号的世界知识。《手册1》将重点描述捆绑于每一延伸概念的世界知识,《手册2》将重点描述捆绑于每一基本句类和常用混合句类的世界知识,《手册3》将重点描述捆绑于每一领域句类的世界知识。

上列“典型”的含义是指典型语种,不只是汉语,还包括其他重要语言,首先是英语。对基层与介层概念世界的世界知识捆绑不能只是自然语言的叙述,还必须使之符号化和结构化,形成相应的数据库和知识库。

由此可见,三部《手册》不是单纯的理论性专著,而是三项巨大的理论工程。