

计算机理解语言研究的新突破

——《HNC(概念层次网络)理论》述评

New Breakthrough in the Research of the Language Understanding of the Computer

林杏光

(中国人民大学对外语言文化学院, 教授      北京 100872)



黄曾阳先生的专著《HNC(概念层次网络)理论》于1998年11月在一流大学的一流出版社(清华大学出版社)出版了。作为HNC理论产品化研究的一个成员,喜读这部70余万字的鸿篇巨制,我

深感可贺,可赞!

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汇合是历史发展的必然趋势。为了顺应这一历史发展潮流,笔者作为一个积极推动语言学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业委员会首届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委员和中国中文信息学会理事,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,我认识到用长年来流行在我国语言学界的“语素—词—词组—句子成分—单词句—复句”这一套汉语语法学去解决汉语的计算机理解是走不通的。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,汉语语法学出现了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,这显然是不对路的。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,词的下界面跟语素划不清界线,词的上界面跟词组划不清界线。要标词性嘛,名、动、形的界限也划不清楚,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。黄曾阳先生一针见血地指出,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。这并不是我国的语言学家和语言信息处理专家无能,而是汉语语法研究的路子不对。我国著名的老一辈语言学家张志公先生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黄曾阳先生创立的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在流行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理论提出了可在信息上工程上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文理解的强大而完备的语义描述体系,包括语句处理、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,并正在实现产品化。

自然语言的计算机理解的发展主要围绕三个方面:自然语言的表述和处理模式;自然语言知识的表示、获取和学习;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8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诊断。(2) 建立网络式概念基元符号体系, 即概念表述的数学表达式。这个符号体系或表达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达式。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, HNC 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“自然语言的计算机理解”中的“理解”这一含义贪大求全, 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)的全部功能。人脑由 150 亿至 180 亿个脑细胞组成, 恰似人体司令部。在人类科技发展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 8 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, 计算机就可以运用这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述, 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要素, 没有这两个基本要素便不成其为概念。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一起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, 由它可构成 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。鉴于 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, 所以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关系之中。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。因此过程和转移、关系和状态也是一种效应的表现形式。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单克(Schank)的概念从属理论(Conceptual Dependency Theory)是一致的。可惜单克只主要考虑了“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的作用效应链完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理论得到工程的实现, 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。有了句类知识的指导, 句类分析就可赢得语句理解的能力。

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, 首先要抓住语句的核心。汉语的语句核心没有形态标志, 拿到一个汉语的句子, 计算机如何能抓住句子的核心呢? 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理论终于突破了这个难关。抓住了语句核心之

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(语句的数学表达式)和不同性质的语义块(语句的物理表达式), 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现现象, 抓住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。

自然语言的计算机理解, 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之上的。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), 而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, 对我国的高新技术领域的国际竞争具有重大的意义。

HNC具有巨大的应用潜力和广阔的应用前景。多年来, 全球科技界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。今人可喜的是, 为了HNC理论的产品化, 在

(下转第35页)

业,并抓紧制订《中小企业促进法》或《中小企业振兴法》等。第二,完善中小企业的管理机构。国家经贸委已设立中小企业司,负责制订中小企业的扶持政策,指导中小企业的改革和发展,并促进和健全中小企业的服务体系。各地也应相应建立中小企业管理机构,负责本地区有关中小企业的行政管理工作。中央和地方的中小企业管理机构不应直接管理企业,干预企业的生产经营活动,而应同各中介组织一道,通过政策引导、行业规划、信息服务、咨询策划、技术指导、人才培养和国际合作等方式,扶持中小企业健康发展。第三,减轻中小企业的负担。国家应当通过规范税收制度、清理额外收费的方式,切实减轻中小企业的各种负担。第四,加强对中小企业的宏观引导。有关部门定期发布行业信息或具有指导作用的《中小企业行业发展指导目录》,引导中小企业进入或退出某些行业和领域,以避免盲目竞争和重复建设。

其次,切实解决中小企业的金融服务问题。一方面抓紧落实已出台的改善中小企业金融服务的各项政策。国有商业银行新设立的中小企业信贷部应增加信贷投入,简化操作手续,提高办事效率,探索适应中小企业资金运行的信贷管理办法,履行好为中小企业服务的职能;城市商业银行和城市信用社把服务重点放到城市中小企业;农业银行和农村信用社应加大对乡镇企业的支持力度。另一方面,在借鉴别国经验的基础上,大力探索适合我国国情的支持中小企业发展的新路子。借鉴日本设立中小企业公库的经验,建立中小企业投资公司,采取国家财政拨付部分资本金、再以资本金的一定倍数为限发行债券的方式筹资,用于支持中小企业进行技术改造和风险性投资;借鉴日本根据《中小企业信用保险公库法》设立中小企业信用保险公库的作法,改革我国的保险体系,成立中小企业信用保险机构,对中小企业的债务给予担保,解决其贷款难的问题;借鉴美国小企业管理局的作法,设立“小企

业出口流动资本项目”,使许多商业银行得以利用该项目为小企业提供短期出口信贷。

再其次,帮助企业提高素质。国家应加强对中小企业进步工作的支持:制订有关规划和计划,并由财政拨付相应的经费,鼓励中小企业增强技术创新能力;设立技术推广中心,通过各种渠道向中小企业推广成熟的适应技术,帮助中小企业提高技术水平。当前,应充分利用结构调整的有利时机,在发展大企业、大集团的同时,加快扶小步伐。推广日本企业的“下承包”模式,发挥大企业、大集团的优势,将中小企业纳入大企业、大集团的发展体系,二者之间通过利益纽带形成配套协作关系。这将使大企业能够通过其内含的交易、组织关系,带动中小企业进入高层次的生产空间,从而使中小企业的生产活动同社会的产业结构变动方向保持一致;与此同时,由于在下承包制度中大企业与中小企业之间形成独特的技术传递链条,大企业在市场交易的基础上,通过组织行为,主动地将质量管理经验、新产品的生产方法、新技术情报、新产品开发技术等向下承包的中小企业传递。

第四,规范和指导中小企业改革。通过改制,选择适当的财产组织形式,实现政企分开,从而达到转换经营机制的目的。在进行企业改制时,要严格进行资产清理评估、产权界定和股权设置,妥善处理债权债务和产权转让收入,并妥善安排富余职工,健全公有资产的管理和监督体系,确保中小企业产权制度改革能够规范、有序地进行。改制有多种形式,不能强求一种模式、搞“一刀切”,而应以“三个有利于”为标准,从客观实际情况出发,选择适宜的作法。对普遍使用的股份合作制要突出其“合作”的性质,应当在资产构成中保留一部分公共积累,在构造股权结构时应当对大股和小股的差额有所限制,在股东的权利上要实行一人一票而非一股一票。

(责任编辑 王宏章)

(上接第 64 页)

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。

HNC(Hierarchical Network of Concepts)是概念层次网络理论的英文简称。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。黄曾阳先生的论著《HNC(概念层次网络)理论》,从一个全新的角度,对这一挑战作出了回应。这部专著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。总之,凡研究或关心与自然语言有关的科学问题的读者都将从该书中获益匪浅。

(责任编辑 蔡德诚)