

“自然语言理解”的新进展

——简评黄曾阳先生创立的 HNC 理论

A Brief Review on Hierarchical Network of Concept

苗传江

(北京语言文化大学语言信息处理研究所, 硕士 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 Concept(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。

中国中文信息学会理事长、中国工程院院士陈力为教授给黄曾阳先生的论文写了评语:“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。《HNC 理论概要》的作者黄曾阳先生创立的面向‘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,因此,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。”

一、HNC 理论包含极其丰富的内容

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些模式对自然语言进行理解和处理了。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述。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。主语义块有 4 种,辅语义块有 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)短时记忆向长时间记忆扩展的模块。

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。现已建成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。

二、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,这里主要介绍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,对“自然语言理解”的研究和成果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。知识库建立的首要目标应定位于对自然语言模糊成份的消解,这是HNC理论对迄今为止的各种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。

三、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,汉语语法学上出现了以西方语言学理论研究汉语的态势,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,而汉语是无形态语言,用形态语言的理论去描写无形态的汉语,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。而由黄曾阳研究员研究、探索和创立的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,从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。

四、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,从理论上说,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。正是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,他们正在向中国进军,凭着雄厚的经济实力,大事“收买”中国的人才、技术和成果,如果长久下去,中国人还哪有自己的信息产业?我们国家正面临丢失中文信息处理

(下转第27页)

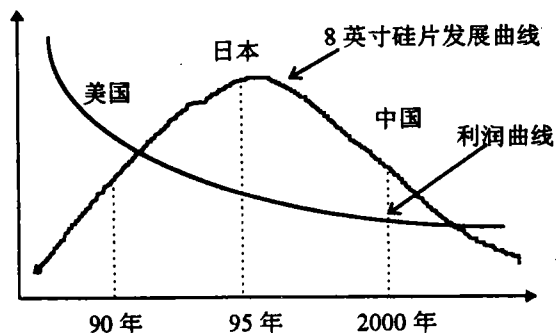


图1 材料发展和利润曲线

日本；而我国主要生产和利用的则是下降期的材料，其中技术含量已不再能产生多少利润。由此可

见美国、日本和中国之经济现状。

表1是硅材料的发展历程和趋势，以国际的发展状况为基准，随着集成电路集成度越来越高，硅材料的直径也越来越大，到现在2~3英寸的硅片已不生产；而我国却仍处于世界80年代的水平，生产的利润率很低。

三、对中国材料科学发展的一些思考

上面分析了中国材料科学的发展状况和经济水平，两者是相互影响、相互作用的。由于材料发展水平低，所以导致经济水平比较低；而由于经济水平较低，又不能大规模投资提高我们的材料水平，所以现实是比较严峻的。我们必须充分认识我国的现状，加倍地努力，提高我国的材料水平和经济水平。

表1 硅材料发展历程及生产情况

年份(年)	1970	1973	1976	1979	1982	1985	1988	1992	1995	1998	2001	2004	2007	2010
集成度(位)	1K	4K	16K	64K	256K	1M	4M	16M	64M	256M	1G	4G	16G	64G
硅片直径(英寸)	2	2.5	3	4	5	6	6~8	8	8~10	12	12~14	14	14~16	16

中国

日本

美国

综观我国的材料科学研究，存在的问题有：一是材料研究和生产的结合不密切，研究不能尽快转换为生产力；二是整体水平落后，制约了优势项目的发展。作为一个大国，我国的材料科学必须全面发展，又要兼顾重点。因此，我国材料科学的发展战略应是：既要增强关联交叉学科，提高系统性和综合性水平；又要重点发展目前能够用于生产的新型材料研究，以提高经济实力和材料整体研究水平，

为未来走向世界前列奠定基础。

参考文献

- [1]“硅低温外延及其应用与硅基材料”高级研讨班，1996，浙江奉化
- [2]吴光宇等主编，现代科学技术革命与当代社会，北京航空航天大学出版社，1995，北京
- [3]谢常青，面向21世纪的世界集成电路工业，科技导报，1997(5) (责任编辑 刘先曙)

(上接第25页)

软件这块我国计算机产业仅存的软件阵地的危险。不久前，美国的IBM公司推出了汉语语音输入系统，这是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企及、无人可相比的，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情况下紧张地工作，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，这预示着HNC理论的发展和应用有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) (责任编辑 蔡德诚)