

自然语言理解的人本中心论

易绵竹 汤庆国

解放军外国语学院计算语言学研究室

中国河南洛阳471003, 036信箱10号

摘要: 在本文, 我们提出以人为中心的语言观来探讨自然语言理解的智能化机制, 充分强调人性因素在智能型人机界面系统中的重要作用。随后, 我们将详细论述自然语言文本理解的哲学、逻辑学、认知学、心理学基础, 揭示理解的层次性, 并刻划各层次的基本特征, 最终试图设计一个具有相当解释力和心理现实性的文本理解模型。

关键词: 自然语言理解 智能化 人本中心论 人性因素 人机界面系统

An Anthropocentric Approach for Natural Language Understanding

Yi Mianzhu Tang Qingguo

Computational Linguistics Section, PLA Foreign Languages University

Box:036-10, Luoyang 471003, Henan P.R.China

Abstract: In this paper, we provide an anthropocentric point of view on language to research the intelligentualization mechanism of natural language understanding and fully emphasize the importance of human factors in intelligent human-machining interface system. Then, we'll detailedly discuss the philosophic, logical, cogitological and psychological principles for understanding text in natural language, bring to light the levels of understanding and describe their basic characteristics and, at last, attempt to design a text understanding model which processes enough explanatory power and psychological reality.

Keywords: natural language understanding, intelligentualization, anthropocentric approach, human factors, human-machining interface.

1 引论

处于世纪之交的人类社会正面临以信息技术为核心的新技术革命(即信息革命)的严峻挑战, 世界各国纷纷兴建的信息高速公路(information superhighway)被认为是通向未来的必由之路。在未来高度信息化社会里, 一个国家的综合实力及其现代化程度将按其信息技术水平的高低来衡量和判别。

起初以自然语言处理为本职专业、而今已成为当代最有影响的未来学家N. Negroponte宣称：“我们已经进入了后信息时代(post information age)”，它在本质上完全不同于生产原子(atom)的工业时代，而是制造比特(bit)的时代。按照Negroponte的理解，比特就如同人体内的DNA，是信息的最小单位，是一种存在(being)状态，同时也是数字化计算中的基本粒子。作者认为，“计算不再只和计算机有关，它决定我们的生存”，这就是“数字化生存”(Being Digital) [1]。毋庸置疑，“数字化生存”的确是一种新思维、新观念，它能够反映我们置身其中的信息社会、网络文化之基本形态。其实，信息数字化的思想早在我国古代的《易经》中就有体现和演示。潘启明先生用现代科学观点去研究易学，提出易象数是宇宙的全息数学模式，应当走易象学之路，而不能拜倒在西方实证科学门下。作者通析《周易参同契》，阐明宇宙是由气(energy)、数(information)和象(mode)三部分组成的道理，进而解释气是能量、数是能量存在的指令信息，而象则是人所确认的能量的存在形式。数在人体科学中就是遗传信息(DNA)，而在信息科学中则称之为比特(bit)，一串比特通常代表的是数字信息(numerical information)。以易经64卦而论，若以卦为单位，它所包含的信息量为64比特；若以爻为单位，它所包含的信息量为384比特[2]。我们可把比特想成“1”或“0”，这两个数码正好组成二进制数字系统。二进制正是十七世纪德国著名数学家G. Leibniz发明的计算机的运算法则，他认为世界上最早的二进制就是中国的八卦[3]。须知，语言信息的计算机处理，必须将输入的文字符号转换成由0、1组合的数据形式才能被计算机所接受、存储、加工及输出。

本文研究的对象是人如何提高计算机对自然语言理解和语言信息处理的智能化水平，从而实现有效的人机双向交流，创造更加完美的人性化界面。我们主张应当把人类行为学的知识应用于高智能计算机及其设备的设计之中，以使计算机和人都发挥最佳功能，这也是所谓“人类工程学”(human-factors engineering)的研究内容。我们知道，人的智商(IQ)毕竟高于计算机，要使计算机能够识别人类知识的表达方式，真正达到美国科学家J. Licklider提出的人与电脑共生(man-computer symbiosis)的目的，人性因素至关重要。

2 人本中心论

数百年来，生理学家、心理学家、哲学家、语言学家，更不用提社会学家和人类学家，都围绕人的问题展开了深入细致的研究。在一切人文科学中，虽然哲学被称为科学之科学，但是现在国内外学术界已达成这样一种共识：语言学不仅是各种科学的基础部分，而且又是一门领先科学(pilot science) [4]。自本世纪20年代以来，人们逐渐清醒地认识到，语言学不仅是哲学和人文科学发展的突破口，同时又是社会科学、自然科学与思维科学的接合部，因而成了一门带头的科学[5]。纵观现代学术思潮总的发展趋势，我们会发现哲学中的语言转向(由此形成“语言哲学”分支)和语言学中的认知转向(由此产生了“认知语言学”)。此外，学科之间的交叉研究已成为一种时尚，应运而衍生出许多边缘性、交叉性学科。仅以语言学为例，就有一系列所谓“带连字符的”(hyphenated)语言学科，诸如：人类语言学、社会语言学、心理语言学、神经语言学、计算语言学、模糊语言学、生态语言学等。必须指出，一切科学，尤其是人文科学归根结底是人学，而人学研究的一个重要方面似是语言，

因为“语言是(人类)存在的家园”(德国哲学家M. Heidegger语)。首先,语言最基本的功能是交际(人与人、人与机器),语言交际的计算机模拟是鉴定语言学理论有效性的试验场,这也是计算语言学的根本任务。其次,语言具有逻辑推论(discursive)或认知(cognitive)功能,因此,语言学被列为认知科学群的核心学科,它们共同关注的是人类思维的动态组合、智能活动的本质特征,以及知识的组织、表示、加工和使用问题。最后,语言还具有载贮(cumulative)功能,它是人们在认识世界、改造世界过程中所获取的各种信息之主要载体。我们要让计算机理解人类语言并有效地处理语言信息,就必须使后者形式化、代码化、数字化、机器化,这样才能达到人机交流的最优化(optimization)目的。

在当代语言学界已普遍接受以人为中心的语言观,认为研究语言不可能脱离人,因为“无论作为说话者,还是作为叙述世界的主要行动者,人都是语言的中心角色(central figure)。人是语言的创造者和操纵者,而语言是作为人的交际工具在千百年历史长河中得以确立和存在”[6]。关于语言和人的问题,本世纪德国著名心理语言学家 K. Buler有过精辟论述:“自古以来,劳动工具和语言就被认为是人之所以为人的最鲜明表现。拉丁语词homo faber(即:人是劳动工具创造者)表示人使用精选并加工的物体作为工具,而希腊语词zoon politikon(即:人是社会动物)表示人运用语言与自己的同类进行交际”[7]。人们正是通过语言表达自己的内心世界,描绘素朴的物质世界,构拟想像的可能世界。

计算机刚刚问世,就有人把这种方便快捷的工具用于语言学的研究,诸如词频统计和词典编纂。随着科学技术日新月异的发展,自然语言的计算机处理能力不断提高,有识之士将机器翻译这项属于高科技领域的语言工程学科作为主攻方向,致力于自然语言文本理解、转换与生成的研究,取得了长足的进步,有些机译系统已经商品化、产业化,如美国的SYSTRAN系统,中国的科译1号英汉机译系统,IMT/EC-863智能型英汉机译系统,高立英汉、日汉机译系统等。我们认为,一个成功的实用化机译系统或人一机对话系统研制者的关注焦点应集中在:提高系统的智能化程度,创造良好的人性化界面。

如何提高计算机的智能?北京大学教授俞士汶先生对此颇有高见,他明确指出,人的介入和人一机共生是使机器智能不断提高的关键因素。作者开展《自然语言的计算理论》专题研究,敏锐地发现智能科学的三个分支ABC(即人工智能: Artificial intelligence, 脑科学: Brain science, 认知科学: Cognitive science)同人与机器的自然语言理解、自然语言习得有密切关系[8]。作为有志于语言学及其交叉学科的研究者,我们完全赞同上述学术观点,并在自己的博士论文[9]和专著[10]中有意识地涉猎认知语言学、心理语言学、神经语言学、语言哲学(主要是阐释学: Hermeneutics)以及人工智能等知识领域,且取得了一些初步得到国内外学术界公开承认的研究成果。在此恕不赘述,下面转入另一话题的讨论。

3 理解层次论

人类具有理解语言、运用语言进行正常交际的能力,而计算机只有理解了人类语言,才能有效地处理语言信息。形象地说,计算机只有变得更加聪明(“知书达理”)、更通人性(“善解人意”),才能更好地为人类服务。我们知道,计算机作为人类制造出来的信息加工工具,它代替人脑对各种信息进行处理,可以说是人类大脑的延伸。正是在这个意义上,计

算机又被称为电脑。智能计算机应当具备模拟人脑的直观感知、认识、学习和推理功能，可以自动积累知识，根据执行情况修改计划、接受训练，甚至通过分析大量数据发现规律。近年来，为了使计算机模拟人类大脑的功能，一些发达国家注意开展人工神经网络的研究，向计算机智能化迈出了重要的一步。

计算机的本质功能是符号处理，语言符号系统具有相当严密的层次，语言信息可分为语法信息、语义信息和语用信息。我们以位语法 (Positional grammar, 简称PG) 理论为先导，采用认知科学的模式化方法，将语言系统内部各层次的信息分别表示在句法位谱配模式、语义位框架模式和语用位序列模式之中 [11]。清华大学教授黄昌宁先生应邀在香港第17界东方语言计算机处理国际会议 (ICCPOL' 97) 上所做的学术演讲题目是：“Can computers process the language information level by level?” 我们对这个不言而喻的反问应作肯定的回答：计算机处理语言信息必然按层次进行。应当指出，语言信息分层模式化乃是语言知识表示的有效方法，它对计算机自然语言理解模型和机译系统的设计与研制具有重要的指导意义。

随着对语言本质属性认识的不断加深，我们将语言三个层面信息之间的关系表述为：语法信息是通过词法、句法分析而得出的表层显性信息，它是语言分析的基础；语义信息是通过逻辑、概念分析而揭示的深层隐性信息，它是语言理解的核心；语用信息是通过语境、常识分析而提取的外部情景信息，它是语言交际的目的。简言之，自然语言理解过程是通过语法实体描写解释语义内容，从而确定其语用交际价值。前二者属记载于文本中的客观信息，后者是文本解读者 (即理解主体) 通析全文后而获得的主观信息。

众所周知，人工翻译有三条准则：“信、达、雅”。为使机器翻译达到理想境界，我们应朝这三个方向努力。语义层面、语法层面和语用层面的集成化或一体化分析，是使人译和机译符合“信、达、雅”标准的根本保证。所谓“信”就是要求译文忠实地传递原文的语义内容；“达”就是要求译文生成符合语法表达习惯；而“雅”则要求译者在修辞、语用方面下功夫，尽可能反映原文作者的交际意图和行文风格。唯有如此，才能提高译文的质量，实现译准率高、可读性强的目标。当然，这个目标是分层次、按步骤来实现的，不可能一蹴而就。

近年来，人工智能学者区分出自然语言文本理解的四个基本层次：1) 语言层次上的理解又称语言学理解 (Linguistic understanding)，只需输入可被解读的文本，人一机对话系统就能找到该文本中出现的词汇单位，弄清它们之间的联系，揭示各语块的句法—语义结构，并对涉及文本内容的直接问题作出回答；2) 逻辑层次上的理解又称逻辑理解 (Logical understanding)，需要对文本进行逻辑扩展，使文本在解读过程中的信息更加丰富，从而反映外部世界构造的特点及其活动过程的规律，揭示这些特点和规律在人的意识和语言中是怎样割断的。时间、空间、动作、因果等逻辑关系的推导，能使人—机对话系统对超出文本内容的问题作出解释性回答；3) 认知层次上的理解亦称认知理解 (Cognitive understanding)，需要调用存储于知识库中的、与原文所描写的外部世界和情景相关的所有信息，从而赋于人—机系统以更高的智能，这就需要设计出在知识库中检索信息的有效程序，许多人工智能专家正致力于这项研究；4) 价值观念层次上的理解亦称价值理解 (Valuable understanding)，需要让机器了解有关社会或某社团普遍遵循的行为规范、价值判断等方面的信息，从而使机器更通人性。除此，有可能还需要更为复杂层次上的理解，如隐喻式或寓言式理解 (metaphoric

or allegorical understanding)。但在近期内，人工智能研究者在这方面未必能取得突破性进展[12]。

4 文本的理解策略及模型构拟

如前所述，我们主张以人为中心的语言观，对自然语言文本逐层分析，以求深刻理解。文本是表现为言语链的语符串，它要被人或机器理解，其各组成部分(语块: chunks)必须植根于读者所熟悉、机器能识别的代码、框架、完型(格式塔: gestalt)等，从而构筑可解度强的连贯性模型(model of coherence)。智能化的计算机模拟系统，实际上是以透明的可操作程序，为揭示隐含于文本中的各种信息提供“可视的”(visual)窗口。基于心理学和认知学的原理，我们把文本理解可定义为一种循序渐进的信息动态组合过程，其最终目的是将作为整体的文本纳入同时感知的意义图式之中，使人或机器能动把握并自动提取文本的语言信息及概念信息。

在文本理解的动态过程中，理解主体(人)一般运用两种基本的认知手段：纯语言知识和非语言知识。前者指语言语义学知识(如语义原语: semantic primitives, 语义算法: semantic algorithms等)，后者指超语言(extralinguistic)语义学知识(如逻辑学、心理学、社会学、伦理学、美学以及人类文化学诸方面的知识)。二者有机地联系起来，就能够反映自然语言理解的实际过程。

我们为文本理解制定以下五种策略(strategies): 1) 直观透视(visual observation): 对输入文本逐句进行可见层次上的语法分析，构建句子线形的句法位谱配模式；2) 直觉判断(intuitive judgement): 对句子各组成要素进行认知层次上的语义角色分析，构建其语义位框架模式；3) 交际前瞻(communicative perspective): 确定不同语境中句子的语用交际价值，构建其语用位序列模式；4) 激活情景框架(activation of situational frame)，以获得情节记忆，它属于短期记忆(short-term memory)；5) 形成语义网络(formation of semantic network)，以获得语义记忆，它属于长期记忆(long-term memory)。顺便指出，人工智能学者提出了中期记忆(mediate-term memory)概念，使人的记忆的三级结构与计算机的存储结构对应起来，即：长期记忆对应外存，短期记忆对应寄存器，中期记忆对应内存。

综上所述，文本理解模型可图示如下(见附图)。按照这个模型，我们曾以俄罗斯著名作家列夫·托尔斯泰的长篇小说《战争与和平》为实例，大致勾勒了该文本的理解框图[9]。

5 结束语

自然语言理解只是人类言语活动的一个方面，本文强调了人性因素在文本理解过程中的核心作用，阐明了人本中心论和理解层次论的基本观点，制定了文本理解的策略，并构拟了一个具有相当解释力和心理现实性的理解模型。我们期望本文的研究能为言语转换与生成机制的探索、为智能化人机界面系统的设计提供一些理论思考，其技术实现需要同软件工程师通力合作才能获得成功。因为研制一个实用化的人机对话系统或机译系统，必须具备一套相对完整的知识表示方法，以及技术上可操作的程序保障手段。以上所述，乃是我们针对NLP中

一些热点问题的初步建议与设想，敬请计算语言学家、人工智能和机器翻译专家提出宝贵意见。

参考文献

- [1] N. 尼葛洛庞帝著，胡泳、范海燕译：《数字化生存》，海南出版社，1997.
- [2] 潘启明著：《〈周易参同契〉通析》，上海翻译出版公司，1990.
- [3] 周述歧编著：《数学思想和数学哲学》，中国人民大学出版社，1993.
- [4] 伍铁平编著：《语言学是一门领先的科学》，北京语言学院出版社，1994.
- [5] 姚天顺等著：《自然语言理解：一种让机器懂得人类语言的研究》，清华大学出版社，1995.
- [6] G. A. 佐洛托娃著：《俄语句法的交际面面观》（俄文版），科学出版社，1982.
- [7] K. 比尤列尔著：《语言理论：语言的表征功能》（俄译本），进步出版社，1993.
- [8] 俞士汶、朱学锋主编：《计算语言学》（教学参考资料），北京大学计算语言学研究所，1993.
- [9] 易绵竹著：《现代俄语三位动词的系统研究》（俄文版，普希金俄语学院语文学博士论文），莫斯科，1994.
- [10] 易绵竹著：《模式分析与知识表达论：位语法理论与实践》，黑龙江教育出版社，1993.
- [11] 易绵竹等：“基于 PG 文法的俄语动词句型描述”，载 陈力为、袁琦 主编：《计算语言学进展与应用》，清华大学出版社，1995.
- [12] K. E. 列维金，D. A. 波斯佩洛夫编著：《人工智能的未来》（俄文版），科学出版社，1991.

附图：

