语言学研究新视界文库

面向自然语言处理的现代汉语词义基元结构研究

胡　惮◎著

中国出版集团
世界图书出版公司

语言学研究新视界文库
教育部人文社科研究青年基金项目（09YJC740060）成果

面向自然语言处理的现代汉语词义基元结构研究

胡　惮◎著

中国出版集团
世界图书出版公司
广州·上海·西安·北京

图书在版编目（CIP）数据

面向自然语言处理的现代汉语词义基元结构研究 /
胡惮著 .—广州 : 世界图书出版广东有限公司 , 2014.9
ISBN 978-7-5100-8588-8

Ⅰ . ①面… Ⅱ . ①胡… Ⅲ . ①现代汉语—词义—
自然语言处理—研究 Ⅳ . ① TP391

中国版本图书馆 CIP 数据核字（2014）第 211706 号

面向自然语言处理的现代汉语词义基元结构研究

责任编辑　宋　焱
出版发行　世界图书出版广东有限公司
地　　址　广州市新港西路大江冲 25 号
http://www.gdst.com.cn
印　　刷　北京天正元印务有限公司
规　　格　710mm × 1000mm　1/16
印　　张　12
字　　数　200 千
版　　次　2014 年 9 月第 1 版　2014 年 9 月第 1 次印刷
ISBN　978-7-5100-8588-8/H · 0872
定　　价　36.00 元

目　录

绪　论

第一节　自然语言处理：人类知识处理的金钥匙

在人类语言学研究的漫漫历史长河中，自然语言处理的研究还只是一支十分年轻、十分活跃的支流。然而，其重要性和应用前景，已经跃居到了语言学各部门、各领域的前沿，潜移默化地改变着整个人类社会生产、生活的各个角落。

一、自然语言处理对人类社会的影响

语言能力是人类作为高等智慧生物区别于其他生物物种的独一无二的本质属性。语言是人类思维的工具，人类的多种智能也离不开语言。人类知识与文化的传播与传承主要是以语言为载体的。

人类对自己语言的观察和研究史，可以追溯到公元前 6—前 5 世纪。在纷繁的学科体系中，语言研究从来都不是孤立的。在语言学的各个历史发展阶段，学者们一直在不断尝试将语言学和其他学科结合起来进行研究。自 20 世纪中后叶计算机发明以来，将语言学这个最古老的学科与计算机科学这个最新兴的学科结合起来，研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法，又成为这两个学科都高度关注的研究新方向。这个领域的研究，被称为自然语言处理。

其实发明计算机的最初目的是为了实现复杂、海量的，人工难以完成的数

值计算。随着研究的深入和应用领域的拓展，这个目的已经远远不能满足实际的需要。科学家们不断进一步探索如何用计算机去模拟人脑的功能，乃至用机器部分实现或实现人类的智能。因此，它又被冠以一个响当当的名字——电脑。

当计算机的应用日益普及，不断渗透到人类社会的各个角落，协助甚至替代人类完成各种工作的时候，这必然会涉及人与机器之间的交互。因此，用自然语言与计算机进行通信，是人们长期以来孜孜以求的目标。要实现这一目标，我们必须解决两个方面的问题，即自然语言理解与自然语言生成。也就是说，要使计算机既能理解自然语言文本的意义，也能以自然语言文本来表达一定的思想意图。这两个方面，就是自然语言处理的主要内容。这也是人工智能研究的核心部分之一。

因此，几乎自计算机诞生之日起，人们就开始构想和尝试将其用于对人类自然语言的处理。这种尝试，是从机器翻译领域开始的。

其实机器翻译的思想与理论研究的历史由来已久。早在17世纪，法国哲学家、数学家笛卡尔和德国数学家莱布尼兹等就提出了使用机器字典克服语言障碍的设想。（巩茗珠，2009）当时这些设想只是停留在理论层面，没有研发出实际的机器。

20世纪30年代初，法国科学家阿尔楚尼（Georges Artsrouni）提出了用机器进行翻译的想法。他申请了一项“翻译机”发明专利，实际上是一个使用纸带的自动双语词典。1933年，苏联发明家特罗扬斯基（Peter Troyanskii）也设计了把一种语言翻译成另一种语言的机器，并在同年9月5日登记了他的发明。这项发明包括双语词典和一种根据世界语处理语法的方法。该系统被分为三个阶段：①由一位讲源语言的本族语编辑将要翻译的文字按照预先设定的逻辑形式和语法规则进行改编。②让机器将这些改编过的文字“翻译”成目标语言。③由讲目标语言的本族语编辑将机器翻译的结果加工润色使之符合目标语的表达习惯。但是，由于20世纪30年代技术水平还很低，这些翻译机器都没有真正制成。[1]

直到1946年世界上第一台电子计算机ENIAC诞生，机器翻译再度被提上议事日程。1947年，信息论的先驱、美国科学家W.Weaver和英国工程师A.D.Booth

[1] 根据百度百科词条“机器翻译”（http://baike.baidu.com/view/21352.htm）及维基百科词条“History of machine translation”（http://en.wikipedia.org/wiki/History_of_machine_translation）编译。

在讨论电子计算机的应用范围时提出了利用计算机进行语言自动翻译的想法。后来，学界普遍把这一年认定为机器翻译的诞生之年。然而，Weaver 的设想在当时并没有得到普遍的认可，大多数人对此持怀疑态度，认为不同语言的词界限过于模糊，情感及跨语言方面的内涵过于广泛。这些质疑，并没有阻挡学界先驱探索前进的步伐。1949 年，Weaver 发表了机器翻译的备忘录，并提出了机器翻译的可计算性。1954 年，美国 Georgetown 大学与 IBM 公司合作实现了世界上第一个真正的机器翻译系统，迎来了机器翻译研究的高潮。（巩茗珠，2009）

半个多世纪过去了，在全球无数科学家和语言学家的共同努力下，机器翻译发展到今天已经得到了长足发展，在理论、技术与应用方面都取得了突破性的进展。虽然目前机器翻译系统仍然不可避免地存在着诸多问题，但这并不影响其欣欣向荣的发展势头。现在，机器翻译技术已经被广泛应用于国际交往和人们日常生活的各个层面，在全球一体化的背景下，为世界各国的政治、经济、文化、科技的交流做出了重要贡献。

除了机器翻译外，自然语言处理的其他应用领域还包括文字识别、语音识别、语音合成、人机对话、信息检索、文本分类、自动文摘、信息过滤、自动问答等等。从目前的理论和技术现状看，虽然通用的、高质量的自然语言处理系统仍然是人们较长期的努力目标，但是针对一定应用、具有相当自然语言处理能力的实用系统已经出现，有些已商品化，甚至开始产业化。典型的例子有多语种数据库和专家系统的自然语言接口、各种机器翻译系统、全信息检索系统、自动文摘系统等。

人类社会的发展史，也是一部知识海量增长的历史。尤其是近 1 个世纪内，人类知识呈几何级数递增。21 世纪以来，随着信息科学的发展和信息技术的迅速普及，人类社会全面进入大数据时代，知识信息已经达到了天文数量级。自然语言是人类知识的主要载体，数千年人类文明积累和传承下来的知识大部分是用自然语言来记载和表述的。在科学技术和文化发展日新月异的今天，每天如井喷般产生的大量新的知识，也主要是以自然语言来表达的。对这些知识和信息的获取、挖掘、加工、存储、传播和应用，仅仅依靠人力的处理已经远远无法胜任，必须借助于越来越高效的计算机技术。因此，自然语言的计算机处理正逐渐成为知识工程与知识管理的研究核心。可以毫不夸张地说，自然语言处理的效率已经

成为了制约整个人类知识产业发展的瓶颈。

人们的日常生活，也越来越离不开自然语言处理技术。现在网络已经渗透到了我们生活里的方方面面，现代社会的正常运转已经完全离不开网络，生活在大数据时代的现代人，几乎都要与互联网打交道。中国互联网络信息中心 2014 年 1 月发布的《第三十三次中国互联网络发展状况统计报告》表明[1]，截至 2013 年 12 月，我国网民规模达到 6.18 亿，网页数量达到 1 500 亿，互联网普及率为 45.8%。这些网络资源中的各种知识和信息，70% 以上是以自然语言为载体的。网民在使用互联网的过程中，都会或多或少地用到自然语言处理的研究成果，从这些浩如烟海的数据洪流中定位、挖掘、获取所需要的各种知识和信息。因此，自然语言处理技术也是影响互联网信息的有效传播和利用，乃至整个互联网事业良性发展的重要因素。

正是基于其重要的战略地位，世界各国都非常重视自然语言处理的相关研究，投入了大量的人力、物力和财力。自然语言处理研究的历史虽不很长，但目前已有的成果足以显示它的重要性和应用前景。在美、英、日、法等发达国家，自然语言处理不仅作为人工智能的核心课题来研究，而且也作为新一代计算机的核心课题来研究。对各类计算机应用系统而言，自然语言处理技术无不占据重要地位。专家系统、数据库、知识库、计算机辅助设计系统、计算机辅助教学系统、计算机辅助决策系统、办公室自动化管理系统、智能机器人等，无一不需要用自然语言做人—机界面。从长远的目标来看，将具有篇章理解能力的自然语言理解系统用于机器自动翻译、情报检索、自动标引、自动文摘等方面，必然具有十分广阔的应用领域和令人鼓舞的应用前景，给人类社会的发展带来革命性的飞跃。

二、人脑的语言思维与电脑的运算

美国计算机科学家 Bill Manaris（1999）曾经这样定义自然语言处理：自然语言处理是研究人际交际与人机交际中的语言问题的一门学科。它要研制表示语言能力（linguistic competence）和语言行为（linguistic performance）的模型，建

[1] 参见中国互联网络信息中心官网文件 http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/P020140305346585959798.pdf。

立计算框架来实现这样的语言模型，提出相应的方法来不断完善这些模型，并以此为依据设计各种实用系统，探讨这些实用系统的评测技术。

按照 Manaris 的定义，自然语言处理不但研究人与机器之间通过自然语言进行信息交互的语言问题，也研究人与人之间的语言交际问题。

然而，因为人脑和电脑的巨大差异，这导致它们对自然语言处理的策略并不完全相同。

早在 20 世纪 50 年代中期，现代计算机之父、著名匈牙利裔美籍数学家约·冯·诺依曼就对人脑的思维过程与电脑的运算进行了比较研究。他指出二者都是一种自动机。人脑是天然的自动机，电脑是人造的自动机。但是，人脑和电脑，无论在控制或逻辑结构上，都有巨大区别。他认为，虽然人脑的“逻辑深度”和“算术深度”都比计算机小得多，但有许多现代计算机所不能比拟的优越性。比如，同样容积的神经元比人造元件能完成更多的运算，能同时处理更多的信息，记忆容量也大得多，每个神经元件的准确度较低而其综合的可靠性较高等等。他还特别指出，人脑的语言绝不是数学语言。（诺依曼，1965）

让机器自动处理人的自然语言，一直是人类的科学梦想。早在计算机出现以前，英国数学家图灵（A.M.Turing）就预见到未来的计算机将会对自然语言研究提出新的问题。他在 1950 年发表的《机器能思维吗？》一文中指出：“我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。”图灵提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语。他天才地预见到计算机和自然语言将会结下不解之缘。（冯志伟，2008）

半个多世纪以来，无数的语言学家、数学家、哲学家、逻辑学家、心理学家、计算机学家、人工智能学家等一直在朝着这个伟大的梦想艰难地跋涉前进，试图揭示人脑处理语言的机制并用电脑模拟这一过程。

然而，心理学和认知科学还远远没有明确揭示出人脑思维的奥秘。人脑对客观世界意义化与符号化的认知原理与操作过程、对知识之间关联路径与网络的

构建算法、对模糊意义的处理策略等等问题，我们尚知之甚少。因此，在这样的前提下，我们要模拟人脑对自然语言的处理，显然困难重重。就目前的计算机技术与人工智能的发展水平而言，电脑与人脑在数据处理能力方面表现出来的差异是显而易见的：

（一）并行计算能力

我们知道，中央处理器（CPU）是电脑的核心部件，功能相当于人类的大脑。早期的计算机，一般只有一个运算引擎。随着技术的发展，这种结构越来越难以满足日益复杂的计算任务，多核处理器应运而生了。多核处理器就是指在一枚处理器中集成两个或多个完整的计算引擎（内核），让他们协同工作，进行并行计算，以实现同时处理多个不同的任务，从而提高计算效率。由于技术条件的限制，电脑的内核并不能无限制地增加。有技术专家断言："一味增加并行的处理单元是行不通的。并行计算机的发展历史表明，并行粒度超过 100 以后，程序就很难写，能做到 128 个以上的应用程序很少。CPU 到了 100 个核以上后，现在并行计算机系统遇到的问题，在 CPU 中一样会存在。如果解决不了主流应用并行化的问题，主流 CPU 发展到 100 个核就到头了。"[1] 而人脑则不同。针对不同的数据处理任务，比如颜色、形状、气味、温度、运动等等，人脑有相对独立的处理中心。也就是说，人脑中存在着成千上万个独立的计算引擎，可以协同处理并行计算任务。这就相当于一个超大型的、由无数的电脑组成的分布式计算集群。虽然电脑的单个 CPU 内核的运算速度远远大于人脑，但是电脑的内核数量及其并行计算能力与人脑有着天壤之别。所以，对单一的计算任务，比如数值计算而言，人脑远远不如电脑。但是对复杂的并行处理任务而言，比如对自然语言的语音、语形符号、语法、语义的综合协同处理，电脑则根本无法跟人脑相提并论。

（二）数据存取效率

虽然从理论上讲，人脑存储数据的能力是无限的，而电脑的容量受其存储器物理空间的制约，但是事实上电脑的存取效率远远大于人脑。对我们人类而言，要获取并记住某项知识，往往需要对大脑进行反复刺激。所谓"书读百遍，其义自见"，实际上就是一个反复存储的过程。能够"一目十行，过目不忘"的奇才，

[1] 百度百科：多核处理器，http://baike.baidu.com/view/2797908.htm。

也仅仅存在于传说之中。而且存储在人脑中的知识或信息，时间长了还会慢慢遗忘。而对电脑而言，只需要一次输入，即可永久保存，除非存储器损坏。在提取数据的时候也是如此。我们大部分人都有过这样的经验：某件非常熟悉的事情或者某个熟人的名字想不起来了；某件重要的事情发生的时间和地点想不起来了；某个重要的物件放在哪里想不起来了，如此等等。这些信息，其实并没有遗忘，也就是说并没有从大脑的记忆数据库中消失，也许慢慢想就想起来了，也许某天突然就记起来了。这些现象，都是因为从记忆库中提取数据的效率或数据存取机制出现了问题而导致的。对电脑而言，就不存在这样的现象。无论其记忆库中的数据量有多庞大，只要给定了检索条件，即可快速准确地提取所需数据。

（三）数据索引效率

无论人脑和电脑，都可以以数据库的形式存储信息。对数据进行定位检索、比较和排序是数据库最重要的基本运算形式。这样的工作对电脑而言是轻而易举的，而人脑在这方面的能力则显得捉襟见肘。比如，从大量文本中提取单词并计算词频、将数万词的词表按一定规则排列、从一个庞大的语料库中提取所有的动名组合、比较两个类似文本的差异等等，电脑都可以在一瞬间准确完成。而对人脑而言，这几乎是不可能的任务。

（四）逻辑容错能力

对于一个计算或自动化处理系统而言，容错能力是保证系统稳定性的重要指标。电脑运算是基于脉冲信号所携带的信息来完成的，如果失掉了一个脉冲，那么其结果必然是信息的意义完全被歪曲了，变得毫无意义。但是，在人脑的神经系统中，即使失掉了一个脉冲，甚至失掉了好几个脉冲，其结果也仅仅是与此有关的频率（即信息的意义）只是有一点畸变而已。在很多情况下，这种畸变并不会对主要信息的传递造成决定性的影响。而且人脑在运算的时候，碰到信号丢失导致信息不全的情况，还会根据已有的认知经验，从记忆数据库中存储的类似的认知图式提取对应的信息来填补，从而保证运算顺利完成。就这个方面而言，人脑的神经系统比电脑的数字系统具有更加优越的修改错误信息的逻辑容错能力。虽然在电脑的软硬件设计中也加入了很多容错的技术，但是其性能远远无法和人脑相比。比如对自然语言中的普遍存在的省略和零范畴现象，人很容易理解，对

电脑而言就是一个难题。

（五）模式识别能力

在生理学、生物学、神经生理学、心理学、认知科学等学科领域，模式识别（pattern recognition）主要是研究生物体（包括人）是如何感知对象的，即生物体对外界环境的自然信息综合感知的机制与过程。在人类的生命活动中，模式识别是时刻都在使用的一种高级智能活动。例如：通过视觉、听觉、触觉等感官接受图像、文字、声音等各种自然信息去认识外界环境；将感性知识加工成理性知识的能力，即经过分析、推理、判断等思维过程而形成概念、建立方法和做出决策；经过教育、训练、学习不断提高认识与改造客观环境；对外界环境的变化和干扰做出适应性反应等等。（杨光正，2001）

在信息科学和人工智能领域，模式识别就是一个要用机器去完成人类智能中通过视觉、听觉、触觉等感官去识别外界环境的自然信息，对表征事物或现象的各种形式的（数值的、文字的和逻辑关系的）信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程。

如上所述，人脑因为其强大的并行运算能力，在大多数情况下其模式识别能力远远优于电脑。电脑的模式识别是通过提取识别对象的一些关键的区别性特征并配合一定的算法来实现的，在某些需要大规模数据、比较特定的领域，比如指纹比对方面可能超过人脑。

对自然语言的识别和理解，也是一种模式识别。目前，电脑在语音识别、文字识别领域，技术已经较为成熟，甚至已经接近人脑。而在自然语言的其他层面，比如句法、语义的模式识别方面与人脑差距还很大。

（六）模糊运算能力

客观世界的大部分事物在其表象之下，都存在着其本身性质的内在不确定性。因此，对客观世界的认识需要强大的非精确、非线性的信息处理能力，这就是模糊运算能力。人脑天生就具有模糊处理能力。基于以模糊集理论为基础的模糊计算技术，电脑可以在一定程度上模拟人脑的非精确、非线性的信息处理能力，从而实现很多应用。

但是，在自然语言的模糊性处理方面，人脑具有得天独厚的优势。语言表

达和理解都具有不同程度的模糊性。自然语言中的很多概念，比如“大小”、“高矮”、“胖瘦”、“好坏”等等都是模糊不定的。体重达到什么水准算“胖子”？头发掉了多少算“秃顶”？什么长相的人算“漂亮”？等等，这些问题，我们很难给出客观的的标准。对人脑而言，迅速评判这类对象并非难事，而且往往可以在大多数个体中达成共识。而对电脑而言则很难回答这些问题。

由于结构原理和工作方式的差异，导致人脑和电脑存在着的这些功能差异，尤其是在对自然语言处理方面的差异，要求我们在揭示和描写自然语言的内在规律的时候，面向人和面向电脑的描写，应该采用不同的策略，以分别适应二者的功能特点。

第二节　自然语言处理的路线之争：经验主义还是理性主义

正确认识事物的本质和规律是人类处理自身与客观世界关系的前提。认识论就是探讨人类认识的本质、结构，认识与客观实在的关系，认识的前提和基础，认识发生、发展的过程及其规律，认识的真理标准等问题的哲学学说。（张东荪，2011）经验主义（empiricism）与理性主义（rationalism）作为两种经典的认识论流派，对哲学与很多其他学科（包括自然语言处理）的发展产生了深远的影响。

一、哲学史上经验主义与理性主义思潮的论争

在欧洲哲学史上，经验主义与理性主义这两种重要的哲学思想论争了几百年，并且一直持续到今天。经验主义的主要代表人物有英国的弗兰西斯·培根、霍布斯、洛克、巴克莱和休谟等；理性主义的主要代表人物则包括笛卡尔、斯宾诺莎和莱布尼茨等。他们虽有分歧，但并非完全对立，而是在既对立又统一的矛盾中共同发展，构成了一段哲学史的丰富内容。整个16—18世纪的欧洲哲学史就是一部经验主义和理性主义哲学产生、发展和终结的历史，也是一部经验主义和理性主义既相互斗争又相互促进的矛盾发展史。（马云泽，1999）

二者分歧的核心思想主要体现在以下两个方面。