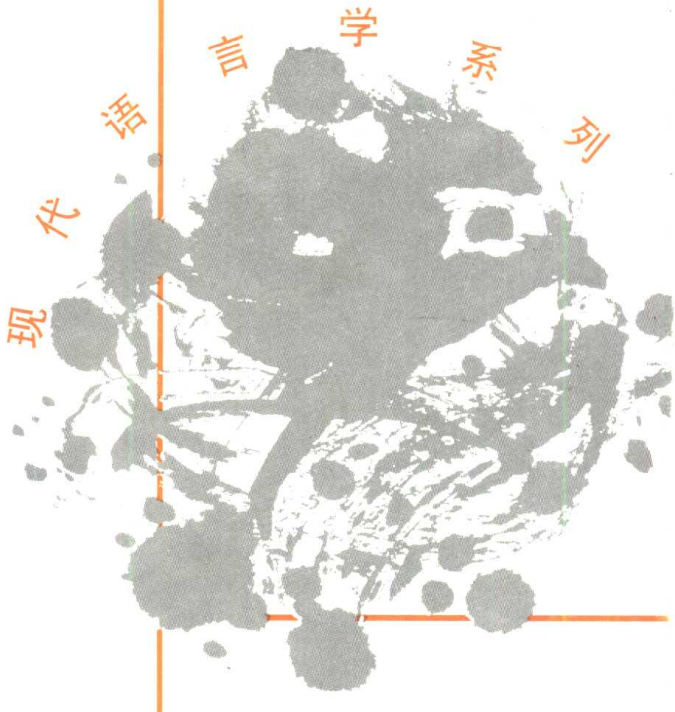
自然语言理解的方法与策略

傅承德　著



现代语言学系列　4　许威汉　主编　河南人民出版社

现代语言学系列之四　许威汉　主编

自然语言理解的方法与策略

傅承德　著

河南人民出版社

图书在版编目(CIP)数据

自然语言理解的方法与策略/傅承德著. －郑州:河南人民出版社,2000.1
(现代语言学系列/许威汉主编)
ISBN 7－215－04001－1

Ⅰ.自… Ⅱ.傅… Ⅲ.自然语言理解－研究 Ⅳ.H0

中国版本图书馆 CIP 数据核字(1999)第 68383 号

河南人民出版社出版发行(郑州市农业路 73 号)
偃师市海洋印刷有限公司印刷　新华书店经销
开本 787×1092　1/32　印张 6.75　字数 134 千字
2000 年 1 月第 1 版　2000 年 1 月第 1 次印刷　印数 1－1 500 册

定价:10.00 元

《现代语言学系列》序

语言学是基础科学，又是领先科学。语言学的新发展，是时代的要求。

从历史上看，19 世纪开始，就有三种因素对语言学的发展起作用：一是科学上的重要发现、发明以及各种学说、社会思潮对语言学产生影响；二是社会的需求推动语言学演进；三是语言学自身要求解决一系列内在问题。而三者之间又相互联系、相互制约。20 世纪电子计算机问世以来，应用范围大大超过单纯的数值运算；它与语言学结合，使语言学发生巨大变化，并相互作用，即计算机给语言学发展的影响，语言学又对计算机的发展起作用。未来社会是信息社会，人机对话更将不断扩大语言交际职能，语言学的功用自必更为突出，语言学的发展自必更为未来社会所关注。

传统语言学的基础雄厚与否，对解决当前问题和迎接未来任务也举足轻重。随着近年文化与文化史研究热潮的掀起，出现了语言与文化相关性的理论，主张科学的语言学应向民族文化和民族语文传统认同并加以转化，创立具有中国特色的语言学。语言学正面临着时代的召唤。

任何学术不是从零开始的，继承与创新是辩证的统一。语言学没有例外，必将循途继轨，不断有所发明发现。况且作为一种学说，向来都有补正发展途地，语言学说借鉴外来新知，发扬优良传统，达到后出特精，正是学科发展所使然。

今天社会主义事业兴旺发达，研究语言热情日益高涨，为适应新形势，“现代语言学系列”的撰写是十分必要的。本系列起点高，眼界宽，内容丰富而系统，具有学术性、科学性、实用性、指导性和资料性特色，适用面广，语言文字工作者、语言文字爱好者，以及其他各有关学科研究者，都可以从中汲取营养。通读全书，当可深会而知之。

许威汉

于上海师范大学中文系

导　言

在计算机上建立起一个能够理解人类自然语言(如汉语、英语、俄语等)的系统,这是人类社会自进入本世纪50年代以来便孜孜以求的一个伟大目标。实现这一目标的主要意义在于:

一,可以使人——机信息交流不再受程序语言的束缚,有助于人类更加方便、自由地利用计算机。

二,可以提供洞察人类语言心理机制的方法和策略。

上述第一项意义,主要是在实践或应用方面的。它是推动这一领域研究工作开展的直接动力,也是迄今为止众多人——机对话系统建成的直接目标。

上述第二项意义,主要是在理论方面的。它为这一领域的研究工作奠定基础,提供方法,指引方向。

上述两个方面是相互补充,相互促进的:面向直接应用的自然语言理解系统的建立,必须以一定程度的有关人类语言理解的知识为基础,同时,在用计算机模拟人类语言理解的过程中,又进一步揭示出人类语言理解某些尚未明了的机制。在对这些尚未明了的机制的研究和探索中,可以产生出许多

新的模拟方法和策略，从而又推动实践或应用的进一步的发展。

关于上述两方面相互补充、相互促进的观点，有些人是持怀疑态度的。他们认为，计算机理解人类自然语言的方式不必跟人类理解自然语言的方式相同或相似。他们的论据是：许多具备某种程度"智能"的计算机系统在执行它们的任务时并不采用类似人类所采用的方法。例如飞机上的自动着落系统，是利用雷达定向的方法来实现着落的。零件分类机器人是利用传感器来感受物体的重量的。这些机制跟人的技巧之间很少有相同之处。然而，评判一个自动着落系统是否成功，跟评判一个自然语言理解系统是否成功毕竟不同。前者只牵涉到客观上一个特殊的目标(即自动着落)是否能实现；只要它能实现，不必管它采用的是什么方法。后者则不同。如果人们断定一个自然语言理解系统是成功的，那么，人们有理由认为，这意味着在该系统的启动中所能观察到的行为应体现出人们自己在理解语言时所具备的特征。也就是说，计算机内部已经形成了人们自己在理解语言时所能形成的相同或相近的功能/逻辑结构，能进行相同的或相近的推论，能注意相同或相近的含义，等等。恰如我们在断言某一个人理解了我们的话的时候，就是推断那个人已经具备了跟我们一样的心理功能或结构一样，尽管我们没有直接的证据来证明这一点。

因此，只有当一个计算机系统能采用一种至少在逻辑上跟人类相当的方法来处理自然语言的时候，我们才能说它真正实现了自然语言的理解。当然，这里是指"在功能或逻辑机

制上”的相当;两者的实际的自然机制自然是完全不同的。也就是说,我们在计算机上用某个数据结构来代表某个意义单元的时候,并不意味着人们头脑中也使用同样的数据结构,也有类似“指针”的东西来指示某个意义单元,等等。我们只是在下面这个意义上认定它们的相同:即以相同的方式利用各自所具备的信息储存技术来联结信息单元,并将其用于相同的目的。

面向应用的自然语言理解系统的实际发展,也显示出计算机理解自然语言跟人类理解自然语言之间的密切关系。经过几十年的努力,我们今天已经能购置到一个实用的数据库系统的自然语言接口了,也可以启动一个用自然语言交谈的专家系统(比如一个“医药咨询专家系统”)了。我们还可以利用计算机作文本自动摘引,或把一种语言自动地译为另一种语言,等等。这些实用系统的建成,是跟它们各自都能在某种范围或程度上成功地模拟人的语言处理能力有关的。但是,跟人的语言处理能力相比,所有这些实用系统的语言处理能力都是极其有限的,这同样是由于它们所能模拟的范围和程度实在太有限了。许多困扰着实用系统发展的因素都跟如何进一步揭示人的语言理解机制并有效地加以模拟这一点有关。例如,语言理解中句法分析是否必要的问题,句法分析跟语义分析的关系问题,语言知识、语境知识和背景知识的组织问题,词语意义的选择和语义歧义的排除问题,词语的所指和照应关系问题,等等。围绕着这些问题,学术界已经并且正在进行深入的研究和探索,提出了种种解决问题的模型和方案。

在这些模型和方案中，有的极富机巧，有的独辟蹊径，从各种不同的角度揭示出了语言理解的某些机制。然而，一个能与人的语言处理能力相媲美的自然语言理解系统的建立，有赖于对人类理解机制的全面揭示和模拟。这是一个长期而艰巨的任务。在这一过程中，往往是老的问题解决了，新的问题又产生了；或者是老的解决问题的方法尚未完善，新的方法又取而代之了。总之，我们绝不能指望有朝一日可以一蹴而就地提出一个十全十美的解决方案来。不过，我们也大可不必灰心丧气。只要想想一个小孩从不会说话到学会说话要花费多少时日，而掌握语言理解机制的大脑的形成，又要花费多少年代，我们就会对自然语言理解这门学科的前途充满信心。

本书的宗旨，是力图对迄今为止已经发展起来的有关自然语言理解的各种方法和策略作一概要介绍、回顾和总结。

本书共分7章。第1章到第3章是从介绍一个简单的、面向微型世界(一个作图世界)的系统着手，引进有关自然语言理解的一些基本原则、方法和概念。诸如“语法的形式化定义”、“语法或知识的内部表达”、“模式匹配”、“转移网络”、“分析程序”等等。第4章到第6章进而探讨有关自然语言理解的一般方法，即面向真实世界的句法分析方法、语义分析方法和篇章分析方法。最后一章即第7章则着重讨论自然语言理解的策略问题，以及跟理解的策略有关的心理学问题。在所有这些讨论中，我们的立足点是在一般的自然语言理解问题上面，但是仍把注意力着重放在有关汉语的自动理解方面。由于迄今为止学术界所提出的各种自然语言理解的方法和策

略，绝大部分是针对西洋语言的自动理解的，虽然许多方面也能适合于汉语的自动理解，但毕竟是不能完全照搬的。必须按照汉语的特点加以改造。在这一方面，本书也力争能有所贡献。

国家文科基地
上海师范大学中文学科点项目

目　录

1　一个简单的自然语言理解系统

1·1　自动作图世界

设想我们已经在微型计算机上建成了一个自动作图系统，该系统能接受由简单的作图语言（一种人工的机器内部语言）编制的指令，并能按照指令的要求储存信息，或在显示屏上作图。

现在我们给该系统装备一个用户友好接口，该接口能够接受用户用汉语发出的指令，并把这些指令翻译为相应的作图语言的指令。这样，我们就建成了一个类似于维纳格雷德的“积木世界”[①] 的作图世界。显然，在这个简单的作图世界里所需使用的自然语言词汇、句子等是十分有限的，因此，要建成这样一个用户友好接口并不是很困难的。然而，由这个简单的作图世界出发，我们可以对涉及自然语言理解的方方

① 参见 Winograd(1972)。

面面的问题展开充分的阐释，并进而引向一些更为复杂的方法和策略方面问题的探讨。

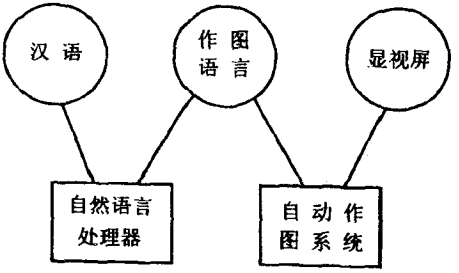


图 1·1

图 1·1 显示了我们这个作图世界的结构框架。由用户终端输入的是数量有限的一些汉语句子。这些句子受到一个严格而有限的语法控制，构成汉语句子集合中的一个子集。输入的句子经过自然语言处理器的处理，被翻译成意义上等值的作图语言的指令，然后再输入到自动作图系统中去。自动作图系统则按照作图语言表示的指令作出反应，把点、线、几何图形等输出到显示屏的相应位置上去。

从自然语言理解角度看，这里的关键就在于能否把汉语的句子翻译成计算机内部相应的作图语言的指令。如果一个输入的汉语句子能被译成作图语言的指令，该句子就是有意义的，或能被理解的；不然，就是没有意义的，或不能被理解的。前者处于作图世界之内，后者则在这个世界的边界之外。

1·2 内部文本的分类及其形式

能直接促进自动作图系统动作的，是用作图语言表达的指令，我们称之为“内部文本”。内部文本可以分为两类。一类是点的赋值语句，如在一个从标为 1000×1000 的显示屏上确定一个点。这类语句的形式为：

point(〈用户定义的点的名称〉〈数字〉〈数字〉)，“point”是关键词，表示要在显示屏上确定一个点。后面括弧内的第一项是用户可以任意定义的点的名称，它可以用一个小写字母来表示，也可以由一个以小写字母开头的字符串来表示(字符串的长度一般都有一定的规定)。后两项是两个 0 到 1000 之间的数字，代表用户定义的点的座标值。下面这些都是合格的点的赋值语句：

point (p, 123, 456)
point (q, 678, 901)
point (r, 987, 654)

内部文本的另一类是连接显示屏上任意两个点的动作语句。这类语句的形式为：

line－seg(〈用户定义的点的名称〉,〈用户定义的点的名称〉)

或：

line－seg(〈数字〉,〈数字〉,〈数字〉,〈数字〉