自然语言理解

——计算机能思维吗

王开铸 著

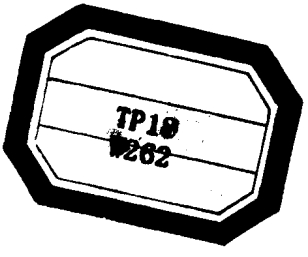
哈尔滨工业大学出版社

2

自然语言理解

——计算机能思维吗

王开铸 著



工业大学出版社

内 容 简 介

全书分六章：第一章叙述了自然语言理解的观点、概念和方法，分析了研究目标和内容；第二章论述了语言与思维的特点、语言与思维的结构及其规律的存在性；第三章介绍了前人在自然语言方面的语法和语义成果；第四章叙述了计算机识别和理解书面语言的过程，在字、词、短语和句子等语法单位上，提出了作者的一些观点，第五章较详尽地介绍了基于理解的中文问答实验系统 CQAES-I 型设计与实现的实例，以说明计算机在有限域内可以理解文章；第六章论述了机器翻译理论与实践。

本书既可作为中文信息处理或自然语言理解研究者的科技参考，又可作为人工智能和自然语言理解的教材。广大读者对“计算机能思维吗？”进行探索，此书可以引其入门。

自然语言理解

——计算机能思维吗

Ziran Yuyan Lijie

—— Jisuanji Neng Siweima

*

哈尔滨工业大学出版社出版

新华书店首都发行所发行

哈尔滨市外文印刷厂印刷

*

开本787×1092 1/32 印张6.625 字数 147 千字

1996年4月第1版 1996年4月第1次印刷

印数1～1 000

ISBN 7-5603-1135-0/TP·82 定价12.00元

序

人工智能所研究的是使计算机能够做那些表现出人的智能特点的事情，使计算机变得聪明些，也定将使人变得更加聪明。我国的“863”高技术研究发展计划中，智能计算机系统被列为17个研究专题之一。这个计划已执行近十年了，这十年中，已取得了可喜的成果。众所周知，听、说、看、写是人的智能行为。如今，智能计算机的发展，在感知信息的获取方面，已取得了长足的进步：文字识别（印刷体和手写体）已有了明显的成果，有的已接近产品了；语音识别（孤立语音和连续语音）在有限词一级，特别是特定人的语音识别也接近实用程度了；让计算机具有语调和情感色彩的语音合成技术，已不是空想了。思维信息处理，或说知识信息处理，即声、图、文不仅被识别，而且被理解，也是国内外研究的热点，已取得了明显成就。

本书作者王开铸教授在自然语言理解方面已研究十多年了，他领导的课题组一直受到国家“863”智能机主题专家组在自然语言理解方面的资助。本书是作者对自然语言研究的立场、观点、目标、方法和研究结果的阐述和总结，因而值得从事本领域研究的学者和学生一读。而书名《自然语言理解——计算机能思维吗？》，本身就是几代学者长期争论的难题。它涉及到关于智能计算、思维、语言和理解等一系列概念的新探索。正像作者书中所说：“计算机理解自然语言的路程方

起于脚下，……”。我想，书中的观点、方法和内容，不求于统一，若能引发出新的“火花”，这才是书的真正价值。

高文

1995年11月于哈尔滨市

前　言

计算机系统能应用在科学计算、数据处理、工业控制、计算辅助设计与制造上已无人怀疑；计算机能够处理数值型数据、非数值型数据也无人怀疑了。而计算机能否处理人类的自然语言，即书面语言和口头语言，不仅能处理中文信息处理前期的录入、编辑、转贮、排版和印刷，而要像人似的，从自然语言中学习到丰富的知识，或退一步说，辅助人去理解自然语言、获得知识和传授知识。在这一点上，持反对态度的大有人在，持怀疑态度的阵营很大，而持赞成态度的，不仅在人数上甚微，而且在态度上也有点暧昧，似乎理不直气不壮。

造成这种局面的因素有两个：其一是研究自然语言理解的最终目标难，正如书中所说，研究思维和语言的机制、结构和规律，要达到这个目标，非一日之功；其二是模拟智能活动，在当前来讲，周期长，致使一些人望而生畏。

幸喜，在自然语言识别上，包括书面语识别和口头语识别上，在国内外已取得较大成果，如计算机对印刷体字的识别早已达95%以上，对手写工整的楷体字识别已达80%以上，计算机对标准普通话的识别(在孤立字或连续语音词级上)率已达80%左右。同样，在自然语言生成上，包括口语生成和篇章生成，也取得较大成果。计算机发出的声音，在音量、音色和音调上已可调，单字音已接近人的要求，在句子级的发音上，已有了语调的控制，已能使人完全听清楚了。估计，自然语言识别和生成技术，到本世纪末，会有更长足的发展，接近实用的产品定会问世。

上述二项技术必然促进自然语言理解的发展，会吸引更多的人来研究自然语言理解。同样，自然语言理解研究的技术，又能推动上述二项技术指标的改进。如在语音识别上，加上自然语言理解的后处理，可使首音（或首词）识别率由70%上升到90%以上；在手写体汉字识别上，加上自然语言理解的后处理，可使首字命中率由80%上升到95%以上。这种良性循环，将促进自然语言理解研究的蓬勃发展。

本书提出的观点、技术和方法，并不是想著书立说，实因这方面的书籍太少了。作为抛砖引玉，以引出真正的著作，也可以说，出版本书的真正心愿，是为了促使自然语言理解研究的蓬勃发展时代早日到来。

本书是作者多年来从事计算机理解自然语言的研究工作和自然语言理解教学的成果和经验的总结。与其说是一本学术著作，不如说是一本经验汇集。全书注重在方法的探索上，而不在于知识的完整性上。

本书的前五章由王开铸执笔，第六章由吴岩执笔。全书由王开铸统稿完成。在成书之际，想到了跟随我研究自然语言理解的学生们，他们是冯寅、王英伟、常雅冬、何卫东、王建波、王晓龙、李俊杰和吴岩等人，他（她）们先后跟着我研究了十多年自然语言理解，走了许多弯路，也正是这些弯路，使我们都成长起来了，在此顺便向他们致意。同时，还要向国家高技术智能计算机系统专家组致谢，在四轮八年中，支持我完成了五项“863”智能机组的项目。正是这些项目的完成，使我们的队伍壮大起来了。

由于此书形成仓促，定有不足之处，万望读者批评指正。

王开铸

1995年10月于哈尔滨市

目　录

第一章　自然语言理解概述

本章提出研究自然语言理解的几个基本概念、观点和认知模型。其目的只是想使自己在探索自然语言理解时，不至于终日彷徨在“自然语言理解能行吗？”的迷途中，而扎实地去做些计算机理解自然语言的研究工作。确实，计算机理解自然语言的路程方起于脚下，尚需大家去探索这些问题的本质。

1.1　自然语言理解的基本概念

自然语言是相对于人工语言而言的。自从计算机问世以来，人工语言的研究蓬勃发展，此起彼伏，已有数百种语言诞生，其中大部分人工语言随着计算机机型的淘汰而消失了，一部分语言更适用于用户需求而得以发展。目前，尚在流行的计算机的人工语言只有数十种。自然语言随着人类的诞生而产生（见 2.1 节），随着社会的发展而发展。经过漫长的社会演变，已形成了如今的八大语系：汉藏语系、印欧语系、亚非语系、阿尔泰语系、乌拉尔语系、尼日尔-刚果语系、马来-玻里尼西语系和德拉维达语系。现已查明全世界有 5 651 种不同的语言或方言。使用人口最多的自然语言如汉语、英语、俄语、日语、法语和德语等。自然语言是人类用于交际和思维的主要工具。

一、自然语言理解的目标

研究语言是个古老的命题。近年来，研究自然语言理解确为多种学科所重视。哲学、语言学、心理学、逻辑学、工程学和计算机科学等学科，都开展自然语言理解的研究。概括起来有两个总目标：长期目标和短期目标。长期目标中，第一个是研究人的思维、思维机制、思维过程和思维规律；第二个是研究语言、语言规律和语用规律。

各个学科按照自己研究的侧面，又制定出各自的短期目标和子目标。如工程学和计算机科学在这两个总目标指引下，突出了自己的一个短期目标，即研究自然语言人机工程，包括自然语言的人机接口设备和人机间的理解机制。

二、立足点

计算机科学工作者或人工智能工作者研究自然语言理解的立足点，与其它学科研究自然语言理解的立足点是不同的。如古今的语言学家，特别是传统的语言学家们，他们的立足点是指人理解自然语言，是研究人从懂得的语言中说明懂，从理解了的材料中说明如何理解。他们所提出的理论，如词法、语法等法则，是解释所占有语料的合法性。我们研究自然语言的立足点是指计算机理解自然语言，是研究计算机从不懂中如何学会懂，即从占有语料出发，研究计算机从不懂到懂，从不理解到甚少理解，到理解的过程。

尽管语言学家们与我们研究的立足点不同，而各自研究的方法、获得的成果与结论，可以相互借鉴和促进。事实上，人工智能就是力图用计算机来模拟人的智能活动，其中包括人对自然语言理解的过程。

站在我们的立足点上，对我们的短期目标——自然语言人机工程进行分解，建立的目标树如图 1.1 所示。

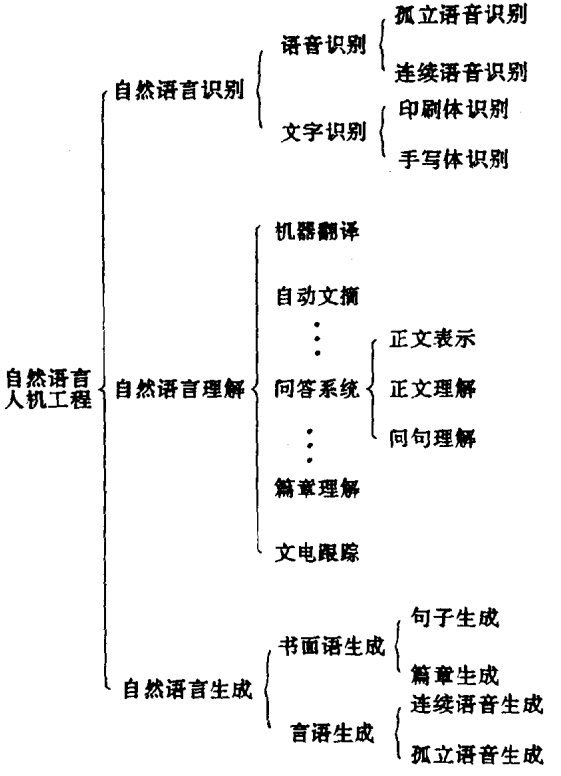


图1.1　自然语言人机工程目标树

三、自然语言理解的观点

自然语言理解的观点，第一是系统工程观点。用系统论的观点来观察分析对象，用系统工程的观点来求其实用。所谓系统就是由相互作用和相互联系的若干组成部分结合而成的具有特定功能的有机整体。我们研究的系统是人工系统，它

也要遵循整体性和有机性。当我们把对象分解为若干个组成部分时，要密切注意他们间的相互联系和相互制约的关系，要密切注意成员与系统的界面和系统与环境的界面。当实现对象时，要注意系统工程的观点，各部分要有机地进行组织。因为系统中的每一个成分在系统中的形式和作用，并不等于它独立于系统之外的形式和作用，因此，系统工程的观点要求我们用“整体大于部分之和”的目标来组织系统，这才能做到系统是有机的整体。

第二是用层次结构观点来分析归纳语言现象。语言是分层次的符号系统。有声语言系统的底层是一套音位，如现代汉语普通话有二十几个声母、三十几个韵母和四个声调。上层中的第一层是音义结合的最小单位“汉字”；第二层是由一个或多个汉字组成的词，它是语言系统中能独立使用的单位；第三层是短语，再后是句子、段落和篇章层次。

第三是层次间单向依赖观点。语言系统的各个层次间存在单向依赖关系。这是因为在语言系统中，任何一个大的语言单位的理解，必须在小的语言单位理解的基础上进行，而小的语言单位的理解，又受大的语言单位的制约。

1.2 语言认知模型

一、人类智能的进化

人类的进化史可以说明：人类的智能是后天演变得来的，而不是想象中的智慧女神似的神秘而不可测。考古、地质等学科发现：

①100 多亿年前产生宇宙。宇宙起源有几大学说。宇宙大

爆炸学说认为：大约在一百多亿年前的某个瞬间，宇宙从温度和密度都极高的状态中，由一次突然的大爆炸中而诞生。

②46 亿多年前有了地球。地球开始是一个热球体，由于温度不断冷却，表面形成了地壳、气圈和水圈，从而形成了现在的地球。

③35 亿多年前有了生命。海洋是生命的摇篮。最初的生命是没有细胞结构的，它们是蛋白质、核酸等有机分子的凝聚体，从海水中摄取有机物作为营养。而后由非细胞形态进化到单细胞形态的生物。

④4 亿年前开始形成动植物。大约 4 亿年前，海洋生物朝陆地迁移。植物沿着苔藓植物、蕨类植物、裸子植物的方向发展。动物沿着无脊椎动物、鱼类、两栖类、爬行类、哺乳类的方向发展。

⑤6 千万年前有了灵长类猿猴。在哺乳类动物中进化出一支最高级的动物，即灵长类的猿猴。

⑥3 百万年左右出现了古人类。由于地壳的变化，气候的变冷，造成了一些森林的毁灭。一直生活在森林中的古猿，有的被淘汰了；有的迁移到别处森林里去；有的来到地面生活。正是这些在地面上生活的古猿，逐渐学会直立行走，手脚分工，大脑也得到发展，以致会制造工具和使用工具。从此，主宰世界的人类出现了。

古人类从制造石器工具中开始了真正人类的物质资料生产的发展，也就发展了真正人类的高级的抽象思维能力。当然，三百万年前人类的抽象思维能力不能和现代人相比，只是说明，现代人的高级抽象思维能力是原始人的抽象思维日积月累年复一年发展而来的。

原始人在聚群中以手势语为主，辅以少量语音符号的方

式交流思想。当直立人能将自己懂得的特殊语言符号，转化为通用的一般性语言符号时，才能以口语为主来交流思想。大约在一百万年前，人类才开始以口语为主辅以手势语的方式交流思想。这种交流思想的语言就是自然语言。

二、自然语言的功能模型

语言是一种很重要而又很复杂的社会现象，它吸引了很多专家的注意。但究其功能模型而言，不外乎有两个。

功能模型 1：

$$\text{思想(A)} \xleftrightarrow[\text{语言}]{\text{语境}} \text{思想(B)}$$

功能模型 2：

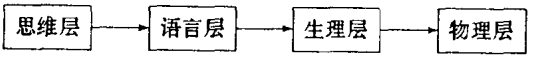
$$\text{思维(A)} \xleftrightarrow[\text{语言}]{\text{语境}} \text{思维(A)}$$

其中，语言是一种由个体特殊符号转化为群体一般符号的系统。A、B 是人类的一分子。语境是指一定的语言环境，即 A、B 共同受制约的语境。

功能模型 1 是表示人们利用自然语言交流思想和表达思维成果的模型。而功能模型 2 是表示人们利用语言进行思维，进行感知知识的再加工的模型。

三、言语链模型

功能模型 1 表示 A、B 二人利用语言在特定的语境下进行交流思想的模型。思想这种信息在听说之间沿着四个层次组成的言语链上传播着，这四个层次是：



思维层：人们在认知模型的基础上，展开积极的思维活