刘开瑛 郭炳炎 编著

自然语言处理

科学出版社

自然语言处理

刘开瑛　郭炳炎　编著

国防工业出版社

1991

内 容 简 介

自然语言处理属于高技术学科，是智能计算机系统研究的重要领域. 本书比较全面、系统地阐述了自然语言处理的基础理论和基本方法，同时结合汉语的特点叙述了汉语处理的一些基本技术，并介绍了国内外在自然语言处理方面的一批优秀的最新成果. 全书共分六章. 第一章介绍自然语言处理的概念、发展及其研究对象；第二章介绍自然语言处理的基础理论和算法，包括形式语言、转换生成语法、功能合一语法、语义网络、扩充转移网络和概念从属理论；第三章介绍词汇分析，包括汉语分词、词类划分以及自动分词技术；第四章介绍句法和语义分析的分析策略和技术；第五、六章介绍话语的分析和生成.

本书适用于在计算机科学、语言学、心理学等方面从事自然语言处理研究的工作者、大学本科高年级学生和研究生作为教材或参考书.

自 然 语 言 处 理

刘开瑛 郭炳炎 编著

责任编辑 刘晓融

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100707

山西省激光照排中心排版

山西新华印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

1991 年 8 月第 一 版 开本：787×1092 1/16

1991 年 8 月第一次印刷 印张：11 3/4

印数：0001—2000 字数：264 000

ISBN 7-03-002488-5/TP·185

定价：11.00 元

序

随着我国社会主义现代化建设的发展，信息处理的自动化愈来愈显得紧迫，而我们日常工作环境中的信息有80%以上是用语言文字作媒体记载和传播的，因此自然语言(汉语)处理已成为一个引人瞩目的重要学科.

自然语言处理是一门边缘性、交叉性很强的学科. 它是语言学、人工智能、计算机科学以及心理学等互相渗透的一门综合性学科. 在这个领域里的科技工作者既需要具有专业知识，也需要具有相关学科领域的各种知识. 他们既是"专才"，也是"通才"，并要在融会贯通的前提下，以博识多才取胜. 我们需要培养很多这样的人才.

由于我国在自然语言处理领域里的科研工作起步很晚，因此，注意引进发达国家在这个领域里的先进技术是非常必要的. 又由于汉语与其他国家的语言相比具有很大差别，因而我们就更要花大力气吸取并扩大我国语言学界多年来积累的丰硕成果，扬长避短，提出适应汉语特色的自然语言处理技术.

刘开瑛和郭炳炎两位同志编著的《自然语言处理》一书"介绍了国内外在自然语言处理方面的一些优秀成果，特别是汉语处理方面的一批最新成果". 这部书写得深入浅出，理论联系实际，是一本很好的入门书. 我乐意向广大读者推荐这部很有用的书.

陈力为

1991年2月4日

前　言

自然语言处理属于高技术学科，它是知识信息处理中的核心课题. 计算机处理自然语言的历史较早，远在40年代计算机刚刚出现时，就有人将计算机应用到语言学的研究中来，但由于受计算机功能的限制，只能进行一些编纂词条索引和词语统计方面的工作. 50年代初，机器翻译几乎成了利用计算机处理自然语言的中心课题，历时15年左右，由于开发技术遇到困难，致使机器翻译系统未能获得成功. 此后，自然语言处理的研究工作，大多转向问答系统的探讨，即人们用某些尝试建立计算机系统，让机器理解语言. 但因知识处理的支撑环境尚很薄弱，致使研究工作时起时伏，进展不大，直到近十几年来，才又蓬勃发展起来. 今天，自然语言处理已成为一个极有吸引力的研究领域，它的研究具有重大的理论意义和实用价值. 首先，当前各国正在研制的智能计算机都把能够理解自然语言作为必备的特征之一；其次，自然语言处理在工程上的应用，将导致许多新产业的产生和发展. 通过自然语言（例如汉语、英语等）实现人-机对话；机器翻译，自动标引，自动文摘研究和办公自动化中的词语处理系统等，都将为智能计算机的诞生作出重要贡献. 目前自然语言处理的研究已经打破了传统的语言学、心理学、数学以及计算机科学的界限，通过这些有关学科之间概念的互相渗透、互相影响，已经形成了具有新概念、新理论、新技术的交叉学科 —— 计算语言学. 作者编写本书的目的就在于为我国自然语言处理的研究工作者提供一本有关自然语言处理的著作. 本书取材力求新颖、丰富，并通过大量的实例或典型系统的介绍，使读者不仅了解基本概念，而且掌握实现的方法和技巧.

本书是作者在近几年来进行自然语言处理研究和为研究生讲授“自然语言处理”课程的基础上写成的，它不仅阐述了有关自然语言处理的基础理论和方法，而且介绍了国内外在自然语言处理方面的一些优秀成果，特别是汉语处理方面的一批最新成果. 本书适用于在计算机科学、语言学、心理学等方面从事自然语言处理有关研究的工作者、大学本科高年级学生和研究生作为教材或参考书. 使用本书的读者需要在计算机软件、人工智能和现代汉语语法等方面有一定的基础.

本书编写过程中得到了清华大学黄昌宁教授的指导和帮助，他审阅了全稿，并提出了许多宝贵意见，在此表示诚挚的感谢. 特别使作者感动的是，陈力为教授在百忙中阅读了书稿，并为本书作序. 老一辈学者的支持和鼓励，使我们获益匪浅，对此，我们深表钦敬并致以衷心的谢意. 山西大学计算机科学系王小鹏、杨尔弘、李涓子等同志为本书的编写提供了许多帮助，在此一并致谢.

由于作者水平有限，书中难免还存在一些缺点和错误，殷切希望读者批评指正.

刘开瑛　郭炳炎

1990年6月于山西大学

目 录

第一章　自然语言处理概述

1.1　什么是自然语言处理

自然语言是指人们日常使用的语言，如汉语、英语、法语、日语等. 它是人类学习环境和互相通讯的工具. 自然语言处理 (Natural Language Processing，简称 NLP)是语言信息处理的一个分支. 所谓语言信息处理，是指用计算机对自然语言的形、音、义等信息进行处理，即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作和加工. 自然语言处理研究使用计算机理解和生成自然语言的基础理论和基本技术，是当前人工智能研究的核心课题之一. 因为处理自然语言的关键是要让计算机“理解”自然语言，所以，自然语言处理通常又叫自然语言理解. 计算机理解自然语言可分为两个方面：(1) 口语的理解. 用口语对计算机讲话，通过语音识别、理解与合成，使计算机能够“听懂”，并作出响应；(2) 书面语的理解. 把文字输入计算机，通过分析和生成，使计算机能够“看懂”，并作出回答. 本书主要是讨论书面语言的理解. 由于一个完整的处理自然语言的计算机系统既要有“理解”的功能，又要有“表达”的功能，因此本书使用自然语言处理这个含义较广的术语.

1.1.1　什么是语言理解?

自然语言理解是人工智能极其活跃的研究领域，也是新一代计算机必须研究的课题. 但什么是“理解”呢? 这是一个富于哲理的问题，目前还难以给出一个准确的定义，尚无令人满意的答案. 然而，长期以来，还是有不少学者，如语言学家、心理学家、哲学家、逻辑学家等致力于有关理解的探讨，他们都站在各自的立场上对其进行解释. 如心理学家认为，理解是“紧张的思维活动的结果”；哲学家认为，理解是“认识或揭露事物中本质的东西”；而逻辑学家则认为理解是“把新的知识、经验纳入已有的认识结构而产生的”. 所有这些解释尽管说法不一，但都是为了弄清楚语言理解的机理和过程.

60 年代以来，计算机科学家，特别是人工智能学者对计算机理解自然语言颇感兴趣. 他们采用人工智能的理论和技术，将设定的自然语言机理用计算机程序表达出来，构造能够理解自然语言的系统. 他们从系统功能的角度出发，把输出对输入文本的反映作为衡量计算机理解语言的判别标准. 美国认知心理学家 G.M.Olson 曾提出四条语言理解的标志：

1) 能成功地回答输入语料中的有关问题.

2) 在接受一批语料之后，有就此给出摘要的能力.

3) 能用不同的词语复述所输入的语料.

4) 有从一种语言转译成另一种语言的能力.

无论机器具有什么特性，无论程序设计采用什么样的算法，只要具有上述功能之一，就应该说机器实现了自然语言理解. 它就可以在机器翻译、自然语言接口、篇章理

解、篇章生成等场合获得广泛的应用.

不难看出，上述理解标志实际上是从整个系统的总功能的观点出发而提出的. 这样，自然语言的理解过程，实质上是把一种表达转换为另一种表达的过程，这种转换也可视为映射. 建立自然语言理解系统就是寻求映射的算法，使机器能够得到同人在理解上相当的输出. 判断机器是否理解语言的最直观的方法，当然是依据机器对你所提出问题的回答，来判定机器是否理解了你的问话. 比如，当你把一个自然语言理解系统做为一个航班数据库系统的人-机接口时，你说:"我需要尽快地飞往纽约". 如果该系统着手查找最近一班去纽约的班机，就表明它已"理解"了你的话. 为了进一步探讨语言理解的机理和过程，我国王开铸教授（哈尔滨工业大学）研制了一个基于理解的 CQAES-1 型中文段落问答系统. 通过此系统的问答，验证了人的理解分为四个层次进行. 第一个层次是理解句中每个概念在句中的作用和地位. 第二个层次是理解句子省缺成分和指代关系. 第三个层次是理解句中和句间的关系. 第四个层次是理解某些句子的字里行间的意义，即某些概念的外延知识. 因此，要让计算机理解自然语言，必须用层次结构的观点分析语言现象，使各个层次间在理解上存在单向依赖关系. 即对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是在大的语言单位的制约条件下获得的. 对篇章语言单位的理解模型可分解为五个层次，即分词层、短语层、语句层、段落层和篇章层. 这虽然是一个理想模型，但随着计算语言学的发展，它必将逐步得到实现.

1.1.2 一个汉语理解实例

自然语言处理，国外在 60 年代就开始取得成果，建立了一批自然语言处理系统. 国内在 70 年代末期开始汉语理解研究. 借鉴国外自然语言理解的理论和模型，结合汉语的某些特点，提出了几个初步的汉语理解模型，并建立了汉语理解实验系统和汉语接口实验系统. 例如，最早建成的两个书面对话系统是《机器理解汉语实验 I: CLUS》（中科院心理所李家治、陈永明，1980 年）和《RJD-80 汉语人-机对话系统》（社科院语言所范继淹、徐志敏，1980 年）. 前者属于心理学模型，后者属于语言学模型. 下面以 RJD-80 汉语人-机对话系统为例，简要说明其结构与功能.

RJD-80 系统是人和计算机用汉语普通话进行书面交谈的一个问答系统. 计算机具有句法分析、语义解释、知识检索、话语分析、逻辑推理、根据语义成分生成语句和排除非法输入等功能. 图 1.1 给出了 RJD-80 系统的组织框图.

如图所示，语句输入后，按词查阅词典和句法规则，然后进行句法分析. 如果词汇、句型符合，即根据语义、推理等规则进行语义解释. 同时，查询知识库，对知识内容进行检索，求得答案，组织回答输出. 遇有不符合现存词汇、句型或语义规则的语句，即作为非法输入不予接受.

RJD-80 以中国文学作品的普通常识为主题，贮存词汇 250 余条，其中名词 214 种，包括作家名 34，作品名 139 等. 贮存句型 30 多个，包括不同形式的非问句，特殊问句、选择问句、陈述句、"把"字句、"被"字句、"主谓"句、复句、包孕句等. 句法分析以转换生成语法和扩充转移网络语法为基础. 结合汉语实际制定句法规则.

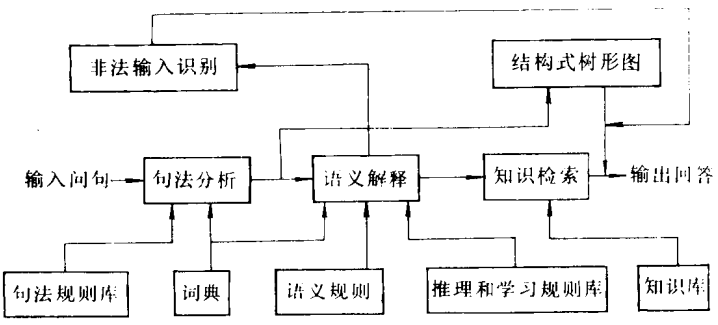

图 1.1 RJD-80 系统的组织框图

一个人看到一个句子或一段话，总是充分运用他的知识和智能去理解这个句子或这段话. 这就是说语言理解不仅靠语法知识，最重要的是它对主题事物的了解. 主题事物即背景知识，是人类对客观世界的感性认识和理性认识，是理解语义内容的基础. 机器没有知识，所以设计任何主题的人-机对话系统必须建立知识库，把有关的背景知识全部贮存在内. 如果提出的问话超出了贮存的知识内容，即使句法分析完全正确，也无济于事. 这就说明句法分析不等于语义理解. 这里举出 RJD-80 中的几组问答（编号是为了叙述方便）:

1) 人：日出的作者是不是巴金　　机：不是，是曹禺
2) 人：你看过日出吗　　机：我没看过
3) 人：北京人的作者是曹禺吗　　机：是的
4) 人：老舍是北京人吗　　机：语义不通

2) 中的“日出”一词，在本系统中指书名，不是指早晨的“日出”现象. 4) 中的“北京人”一词在本系统内指书名，所以尽管词汇、句型都符合本系统的贮存，计算机的回答仍是“语义不通”，因为人不能是作品. 只要知识库中不列入老舍的籍贯，机器就无法回答. 所以包括背景知识的知识库，是建立理解系统的基础工程.

RJD-80 系统中还有话语分析(Discouse Analysis)功能，即对整段话的理解. 在整段谈话中，由于一定的语言环境，经常使用代词或省略某些词语. 下面选用另一片段，说明系统具有确定代词所指，理解省略词语的句子的能力. 同前段对话并不连贯，只是编号相连.

5) 人：他写过很多诗歌　　机：他是谁
6) 人：艾青　　机：啊，他是诗人
7) 人：你知道谁是诗人　　机：郭沫若
8) 人：他是剧作家怎么又是诗人　　机：他写过诗歌，又写过剧本，所以既是诗人，又是剧作家.
9) 人：老舍呢　　机：老舍是小说家，又是剧作家
10) 人：小王喜欢诗歌，你认为他是诗人吗　　机：喜欢诗歌，不是诗人

从以上讨论中，我们看到了句法、语义、知识库、推理等在研制语言理解系统中的作用.

1.1.3 什么使自然语言理解难

自然语言处理是一个很难的课题. 实践证明，自然语言难于理解的三个主要因素是目标表达的复杂性，映射的歧义性，以及句子成分之间的相关性. 这可以说是国外 30 年来研究工作的总结. 这里，我们仅举例对自然语言在各个层次上存在的各种各样歧义给理解造成的困难进行说明. 先举一个有歧义的英语句子为例:

The boy saw a girl with a telescope.

这里有两种可能的解释. 第一种是"男孩用望远镜看见了一个女孩"，第二种是"男孩看见了一个拿着望远镜的女孩". 人们在阅读或会话时，可以根据上下文和语言环境进行判断，但让计算机孤立地分析这一句话，很难得出正确判断. 而对汉语来说，理解就更为困难，这主要是由其特点所决定.

歧义是语言中大量出现的现象. 按照汉语中词组和句子的构造原则基本一致的语法特点，汉语中的歧义现象主要反映在词和词组两个层次上. 在词这一层，体现为词的多义（多义词），在词组这一层，则体现为词组的多义（多义词组）. 这些歧义现象给计算机处理自然语言设置了巨大障碍，成为我们不能回避也无法回避的一个困难问题.

对《现代汉语八百词》统计表明，在 800 多个词中 22.5%的词有兼类现象，不同的兼类情况约 50 种. 又如《中学生词典》的 14 万词条中兼类词占 5.86%. 初看起来，兼类词只占汉语词汇的一小部分，但我们应当注意到常用词的兼类现象严重，而兼类词的使用频度并不低. 也就是说，汉语中往往越是常用词，其不同用法越多. 譬如动词使用频度较高，动名词兼类占到全部兼类词的 49.8%. 而且在一个句子中，随兼类词的增多，词类的组合数将剧烈增长，如图 1.2 所示的 21 个词的句子，有 5 个兼类词，即可导致 $3\times2\times2\times2\times2=48$ 种可能的组合.

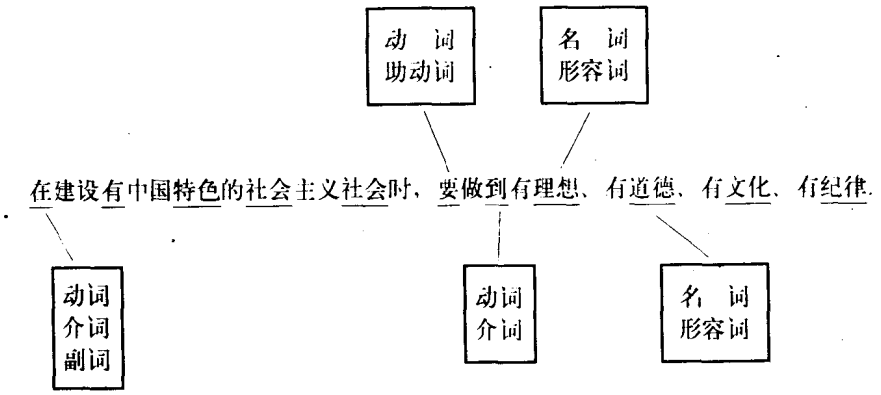

图 1.2 有 5 个兼类词的句子

当然，并不是说多义词一定就是兼类词. 例如，"他走了三个小时了"，可理解为"他在路上步行了三个小时"，也可理解为"他离开了三个小时"，其中"走"是多义词，但两种情况下均属动词，并非兼类词.

汉语中词组的歧义也是普遍存在的，有些词组，虽然词相同、词序也相同，但在不同的语言环境里可表示几种不同的意义. 现举例如下:

1)"援助的是中国"，这显然是一种主谓结构，但它却有两种意义，或理解为"援助

国是中国”，或理解为“受援国是中国.”

2)“穿好衣服”，可理解为“穿好／衣服”，或“穿／好衣服”，两者均为述宾结构.

3)“研究计划”，虽只有一种切分形式“研究／计划”，但可有两种解释，“研究”为动词时，是述宾结构，“研究”为名词时，是定中结构.

4)“爱护人民的军队”，有两种意义，“爱护／人民的军队”或“爱护人民的／军队”，前者为述宾结构，后者为定中结构.

要想脱离语言环境和上下文把上述种种歧义都辨别出来，解释清楚，是很困难的.

汉语中有两种特殊的句型，即连动句和兼语句. 这些句型对印欧语来说，由于有词形变化，使用计算机分析还比较容易，但对汉语来说，就显得很困难了.

连动句是谓语由两个或两个以上连用的动词或动词短语构成的句子. 例如句子

他跑着回来告诉我们这个消息.

共有“跑着”“回来”“告诉”三个动词，究竟哪一个是中心动词，由于无明显的形态标志，机器很难区分. 与其相应的英语句子

He came running back to tell us the news.

由于不定式记号 to 加动词原形构成了动词不定式，动词原形加词尾-ing 构成现在分词，所以机器容易处理.

兼语句的谓语是由一个动宾结构和一个主谓结构套在一起构成的，即谓语中前一个动宾结构的宾语兼作后一个主谓结构的主语，例如:

小孩子笑他是一个大胖子.

这个句子的主语是“小孩子”，谓语是套在一起的动宾结构“笑他”和主谓结构“他是一个大胖子”，其中“他”既是动宾结构的宾语，又是主谓结构的主语，是一身兼二职的兼语. 谓语中只有一个动词“笑”与主语存在主谓关系，第二个动词与主语不存在主谓关系. 而且，“他是一个大胖子”是引起“小孩子笑”的原因. 由于汉语中动词无任何标志，要想让机器作出这样的分析是有一定困难的. 与该句子相应的英语句子为

The children laugh at him for being a big fatso.

由于有 him 指出宾格，而-ing 指出分词，再加上 for 的作用，相比之下，机器分析起来就显得容易一些.

还有，汉语词汇极其丰富，成语众多，量词也特别丰富，这些虽然使得汉字文字表达优美、准确、生动、形象，但也给计算机汉语理解和汉语生成增添了困难.

总之，自然语言理解难，汉语理解就更难.

1.2 自然语言处理研究的历史

自然语言计算机处理研究的历史可追溯到 50 年代初期，其发展大致可分为三个时期.

1.2.1 早期系统（60—70 年代）

当通用计算机问世时，人们想到的第一件事就是用计算机把一种语言翻译成另一种语言. 在开始的 15 年（1950—1965 年），机器翻译几乎成了所有自然语言处理系统的

中心课题. 最初，人们以为翻译包括两个基本过程，即查词典和语法分析. 一篇源语文章可以首先通过查词典，析出文句中每个词的目标语等译词，然后再进行第二步，即语法分析——调整词序、词尾和形式等等. 人们相信，好的译文可以通过分别处理查词典和重新排列词序两步操作来得到，但是使用这种方法没有能够达到预期效果. 我们知道，当一个人听语言时，所以能理解所听到的语言，他不只是依赖他的语法知识，而且还要运用与所讨论的世界有关的知识. 因为某个词和句子在上下文中有一种特定的意义. 许多句子就是依据这种知识而被正确理解的. 早期机器翻译系统未获成功是因为没有去尝试理解它所翻译的内容究竟是什么，所以机器输出的新语言不能精确复示源语言的同样意义.

自然语言处理的一些崭新的人工智能方法受到了 60 年代许多学科发展的影响，其中包括：高级程序设计语言和表处理语言，以及 N.Chomsky 在语言学理论上的突破. 所以在 60 年代中，人工智能学者开发了一批新的计算机程序. 这些早期的自然语言程序标志着人工智能在研究自然语言理解方面的开端. 下面举几个例子.

BASEBALL 是 1963 年由 B.Green 建立的一个情报检索程序. 它的数据库中存放的是关于美国一年内全部球队比赛的各种事实 (month, day, place, scores…). 从用户那里输入的问题只允许有一个简单句子，没有逻辑连接词(and, or ,not)和比较级(higher, longer…)，而且大多数词必须被一部大字典所认识. 分析系统采用 14 种词类和从右到左的扫描，把输入的句子转换为功能短语，提出关键字，再把该功能短语改写成一份规范表达式. 例如:

How many games did the Yankees play in July?

(七月间 Yankees 队进行了几次比赛?)

这个问题经过 BASEBALL 系统处理后,变为如下的规范表达式:

TEAM	=YANKEES	(队名=YANKEES)
MONTH	=JULY	(月份=七月)
GAMES(number of)=?		(比赛次数=?)

通过在数据库中搜索能够同该问题相匹配的数据条目，把它们存放在一张"已求"表中，经过处理和输出来形成问题的回答.

SIR(Semantic Information Retrieval)是由 B.Raphael 于 1968 年完成的，这是他在 M.I.T（美国麻省理工学院）的博士学位论文的一部分. 这是一个原始型的"理解"机器，它能累积事实，然后对这些事实进行演绎以便回答问题. SIR 接受英语的一个有限制的子集，对英语提出 24 种匹配模式. 例如，SIR 中提出了如下模式:

* is *

* is part of *

There are * on *

Is * * ?

How many * does * have?

What is the * of * ?

其中，符号 * 表示名词，这样的名词前可用限定词 a，the，every，each 或数词所修饰. 匹配到一种模式便会在程序中触发相应的行动，作为对用户的响应. 对于 24 种匹

配模式之外的句型，机器是不能识别的.

STUDENT 是另一个基于模式匹配的自然语言处理程序，它是 D.Bobrow 于 1968 年作为他在 M.I.T 的博士研究工作而编写的. STUDENT 能阅读和解决高中代数应用题，列出方程求解并给出答案. 被它所识别的整个英语子集是根据下列基本的模式集推导出来的:

(what are * and *)
(what is *)
(how many * 1 is *)
(how many * do * have)
(how many * does * have)
(find *)
(find * and *)
(* is multiplied by *)
(* is divided by *)
(* is *)
(* (* 1 / verb) * 1 *)
(* (* 1 / verb) * as many * as * (* 1 / verb) *)

其中，符号 * 表示任意长度的一个词串，* 1 表示一个词，(* 1 / verb) 表示必须用词典来识别的一个动词. 系统采用这种简单的模式匹配方式，再加上一些启发式，在分析典型的高中代数文字题时表现出了较强的能力.

ELIZA 程序是 J.Weizenbaum 于 1966 年在 M.I.T 编写的，这是一个模仿心理治疗学家行为的程序. 它或许是这些基于"模式匹配"的自然语言系统中最有名的一个. 为了简化输出过程，有些单词被立即翻译成适于响应的新形式. 例如，若有个患者说: "I cried"，ELIZA 便会以"Why do you cried"，将"I"变成了"you". 这个程序同前述的系统相比并没有采用更多的智能机制，但其对话还是显示了惊人的真实性.

总之，这些早期的自然语言系统中没有一个以任何成熟的方式涉及到语言的句法分析，它们的主要技术是采用关键词匹配，而且只能接受英语的一个受限很强的子集，在受限的专门领域内达到有限的目标，因此理解是相当肤浅的. 这恐怕还不能算真正理解了自然语言.

1.2.2 第二代系统 (70—80 年代)

在这个时期，研制出了一些很有名的系统，标志着自然语言处理进入了一个新的阶段. 这些系统的主要特点是在句法、语义的分析中采用了所需要的知识表达形式和处理模型，尽管它还是局限在某个领域内，但却能更好地理解自然语言. 下面我们列举几个较著名的系统.

LUNAR 是 W.Woods 在 1972 年设计的一个实验性的自然语言信息检索系统，它协助地质专家查找、比较和评价阿波罗-11 火箭上从月球获得的岩石和土壤组成成分的化学分析数据，通过人-机接口，用普通英语来回答有关问题. 它们使用的大型数据库是美国国家航空与航天管理局提供的，是第一个采用扩充转移网络(ATN)分析程序来处

理英语语法问题的系统. LUNAR 还利用了过程语义学的思想，即首先把询问转换为一段"程序"，然后由信息检索模块来加以执行. 该系统有能力处理时态和语气、某些回指的指代和比较，限制性的关系子句和某些形容词的修饰成分. 它也许是把 ATN 分析方法应用于真实世界的一个最佳例子. 由于该系统的专业范围有严格的限制，在语言处理中尽量解决那些常见的语法现象和地质学家们经常用来提问的那些英语句型. 因而这个系统有一定实用性，是一个有用的系统.

SHRDLU 是 T.Winograd 于 1972 年在 M.I.T 的 AI 实验室开发的一个自然语言理解系统，它是一个在"积木世界"中理解英语的计算机系统，它可以回答用户的问题，也可以执行用户的命令，通过操纵机器人手臂，采用相应的行动，移动桌上的积木块的空间位置. 系统还可在显示屏幕上表示相应的景象. 该系统包括一个句法分析程序，具有一部基于 M.Halliday 系统语法的大型英语语法；一个语义分析程序，含有为解释词和结构的意义所需要的知识；一个问题求解器，可为执行命令和寻找问题答案作出安排. 这是一个句法、语义和推理的组合系统. 由于它在语言理解过程中试图把语言学的方法和推理方法结合起来，所以该系统十分引人注目，在自然语言理解研究中向前迈出了一大步，成为人类语言理解的一种比较有生命力的模型.

MARGIE(Meaning,Analysis,Response Generation and Inference on English) 是由 R.Schank 和他的学生们于 1975 年在斯坦福大学 AI 实验室建立起来的一个程序. 其目的是提供一个自然语言理解的直观模型. 早在 1973 年 Schank 就提出了概念从属理论，简称 CD 理论，用来表示自然语言中的短语和句子意义. MARGIE 系统是依据 CD 理论建立的. 一是原语分解假设：在任何语言中，对于任何两个具有相同意义的句子，其意义的表达式只有一种. 另一是把理解作为本能的推理假设：理解过程至少是局部地推理，这种推理由表达句子意义的概念得出. 因此，MARGIE 系统由三部分组成. 第一部分是一个概念分析器，它把英语句子转换为概念从属表达式. 第二部分是一个推理器，它接受已经转换成 CD 表达式的语句，根据在系统中存储的当前语景中的命题来演绎出大量事实. 最后一部分是一个篇章生成模块，它一方面通过一个甄别网络(Discrimination Net)来区别不同的词义，根据英语的上下文选择词，使它适合输出的要求；另一方面利用一个 ATN 把一个内部的概念从属表达式转换为类似英语的输出形式. MARGIE 是一个实验系统，它是 Schank 在计算语言学方面进一步研究的基础.

SAM(Script Applier Mechanism)和 PAM(Plan Applier Mechanism)是 Schank 和他的学生们于 70 年代后期在耶鲁大学开发的故事理解程序. SAM 和 PAM 系统的输入都是故事，而且都利用一个从英语到 CD 的分析器产生该故事的一种内部表达（概念从属方式）. 它们都能够对故事释义以及根据故事进行智能推理. 它们的差别是建立了 CD 表达之后的处理方式. SAM 通过寻找与故事相匹配的一个或多个脚本来理解故事，使故事和脚本相匹配的过程包括三部分：即分析器(PARSER)，记忆模块(MEMTOK)和脚本的使用(APPLY)，这些模块协同工作. PAM 通过判断故事要达到的目标，以及把故事的行动同为达到目标而采取的方法进行匹配，以便理解故事. 理解过程为：决定目标；决定满足该目标的子目标；根据一个由已决定的子目标所调用的一组规划的可能实现来分析输入的概念内容. 这两个系统都以概念从属理论为基础.

GUS(Genial Understanding System)是 Bobrow 于 1977 年在 Xerox Palo Alto 研究

中心开发的一个基于框架的实验性旅行咨询系统，它可以通过自然语言对话帮助用户安排一次空中旅行计划.

1.2.3 第三代系统：展望未来

国外有人把基于知识的篇章处理系统，作为第三代系统的开发目标. 尽管目前还不存在这样的系统，但从已建立的系统似乎可以看出，第三代系统的设计必须基于语言理解、生成、推理和知识库管理的紧密结合、较强的人机交互的支撑以及知识领域之间信息传递方法的应用. 国外关于第三代系统的设想正在探讨中. 这里仅选取以知识工程方法编写的《自然语言处理》(Richard E. Cullingford, Natural Lnguage Processing) 一书中提出的模型，作为本节的例子，以供参考.

第三代系统的模型是采用由 Schank 等提出的针对自然语言处理的"概念信息处理"的方法. 这种关于自然语言处理的观点，主张将理解和生成的重要过程进行在概念深层，而不是在诸如句法和词法的表层. 这样，存储结构、信息的存取、问题求解以及期望形成等问题，都被认为是人类和计算机使用语言的核心. 人工智能的研究当然还包括对自然语言处理的其他方法，而它们或多或少都持有这样的观点.

第三代系统的典型应用如同一个用于医疗、财经、地质和制造等领域的专家咨询系统. 图 1.3 给出了这种典型系统的简单框图.

图中自然语言接口接受用户的查询，并将其转换成意义结构，送至专家推理系统. 推理系统包含其应用知识的符号表达以及基于这些知识而得出结论的推理机制. 自然语言接口还把这由推理系统产生的意义结构翻译成语言，这些结构可以对应于用户问题的回答或系统对用户所提的问题.

推理系统还是接口的两个主要辅助类型的信息源. 正如贮存应用领域的知识结构一样，它可以就用户输入的概念形式向接口提出"期望"，而接口可以利用它来对"意义"作出判断. 由于接口和推理机总是共享一部词典，所以推理系统能够回答接口有关领域的问题.

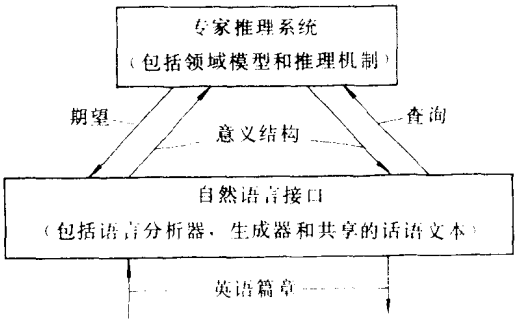

图 1.3 一个会话系统的通用框图

为了满足这些性能，这个系统的结构应是什么样呢？图 1.4 示出了一个简单而适用的模型. 这个模型是以软件工程的结构模块化方法设计的，因此对系统的实现乃至理解和维护都是比较容易的. 同时，它还采用了面向目标的设计思想，使系统的所有性能都