中 文 信 息 处 理 丛 书

自然语言理解

——一种让机器懂得人类语言的研究

姚天顺等 著

清华大学出版社
广西科学技术出版社

中文信息处理丛书

自然语言理解

《Natural Language Understanding》

一种让机器懂得人类语言的研究

姚天顺等 著

清华大学出版社
广西科学技术出版社

(京)新登字 158 号

内 容 简 介

本书内容包括引言、汉语的计算机理解、语法分析、语义分析、概念分析、故事表示、词汇集聚理论、特性和公式、词汇功能文法、功能合一文法、语料库语言学等等，并介绍了作者的研究成果：关于汉语理解和汉英机器翻译，即机器词典、词汇语义驱动理论、中间转接语言、目标语言生成、语义关系集、规则描述语言和汉英机译的实例等。

图书在版编目(CIP)数据

自然语言理解：一种让机器懂得人类语言的研究/姚天顺著. —北京：清华大学出版社，1995.7
(中文信息处理)
ISBN 7-302-01911-8

Ⅰ. 自… Ⅱ. 姚… Ⅲ. 自然语言理解 Ⅳ. TP18

中国版本图书馆 CIP 数据核字(95)第 10563 号

出 版 者：清华大学出版社(北京清华大学校内，邮编 100084)
责任编辑：徐培忠
印 刷 者：北京人民文学印刷厂
发 行 者：新华书店总店北京科技发行所
开　　本：787×1092 1/16　印张：22.25　字数：551 千字
版　　次：1995 年 12 月第 1 版　1995 年 12 月第 1 次印刷
书　　号：ISBN 7-302-01911-8/ TP・875
印　　数：0001—1000
定　　价：32.00 元

Natural Language Understanding

——A study of making a machine understand human languages

Yao Tian-Shun et al.

The natural language understanding makes a study of how human languages can be understood by computer. It was about at the initial stage of the com puter coming into the world that an idea came upon some people. If a computer was able to understand human and their writings, it was marvelous that the computer could do what we asked it to do in accordance with its understanding when we use it. In this case, the computer acts as an electronic brain because it could learn to do what we asked. However, it was a kind of fond dream at that time. In order to realize such a dream, people met with technological crisis, and some people thought that it was impossible to do. But nowadays, we have already stepped into the 20th century. Great changes have taken place. Computer has been advancing on function and volume. Theoretical research has made great progress in the processing of natural languages. So the dream comes upon people today again and makes more and more people remake efforts in it. Especially the study of the man-machine interface system of new generation computers and robots puts it into true. Now the study of natural language understanding gradually becomes one of the popular topics of computer sciences.

This book is the result from my lecturing and every effort made by our research group colleagues in the past eight years. We compile the book for three aims. Firstly, this book is both teach-yourself and reference book used to contribute to the faculty engaged and interested in computational linguistics, intelligence computer, man-machine interface, robot phonic dialogue, natural language inquiry of large database, expert system, computer auto write, abstract taking, file and autoclassification, secretarial(document)management system, automatic language processing of large industrial operation, machine translation, computer automational processing of file and corpus concerning arts and social science, man-machine interactive of CAD, CAI and OA and so on. Secondly, it is intended for the computational faculty who are not familian with the natural language processing, but want to be clear about something in man-machine interface while they are designing their systems. In the past year's rising and falling history in this field more and more people don't know about it only because its research content is quite different from other fields of computer science. Lastly, the book, the natural language understanding (NLU), can be used

as a textbook for the postgraduate's course in NLU or computational linguistics (CL), which may take 40 class hours.

In order to meet the demands of the language information processing, the book contains an introduction, computational Chinese understanding, syntax analysis, semantic analysis, concept analysis, story representation, lexical cohension, features and formulae, lexical function grammar, functional unification grammar, corpus linguistics in which people are interested and focused with our attention on the problems about Chinese computer processing. Meanwhile, in the book, we would like to show our contribution: including the theory and method regarding "the method of lexical semantic driven" and how Chinese analysis and machine translation are underway in the method. In preparation for our publication, we also assimilate what we have learned from《Computational Linguistics》and other works. We sincerely hope that the book would mirror the newest contemporary, and we are either in acknowlegment of my collegues and students' help to rewrite and have more materials in it. We are sure that the book may be helpful to those who are willing to go in for this kind of research project.

May, 1993

清华大学出版社　广西科学技术出版社
计算机学术著作出版基金

评 审 委 员 会

主 任 委 员　张效祥

副主任委员　周远清　汪成为

委　　员　（按姓氏笔画排列）

王鼎兴　杨芙清　李三立　施伯乐　徐家福
夏培肃　董韫美　张兴强　徐培忠

出 版 说 明

近年来，随着微电子和计算机技术渗透到各个技术领域，人类正在步入一个技术迅猛发展的新时期。这个新时期的主要标志是计算机和信息处理的广泛应用。计算机在改造传统产业，实现管理自动化，促进新兴产业的发展等方面都起着重要作用，它在现代化建设中的战略地位愈来愈明显。计算机科学与其它学科的交叉又产生了许多新学科，推动着科学技术向更广阔的领域发展，正在对人类社会产生深远的影响。

科学技术是第一生产力。计算机科学技术是我国高科技领域的一个重要方面。为了推动我国计算机科学及产业的发展，促进学术交流，使科研成果尽快转化为生产力，清华大学出版社与广西科学技术出版社联合设立了“计算机学术著作基金”，旨在支持和鼓励科技人员，撰写高水平的学术著作，以反映和推广我国在这一领域的最新成果。

计算机学术著作出版基金资助出版的著作范围包括：有重要理论价值或重要应用价值的学术专著；计算机学科前沿探索的论著；推动计算机技术及产业发展的专著；与计算机有关的交叉学科的论著；有较大应用价值的工具书；世界名著的优秀翻译作品。凡经作者本人申请，计算机学术著作出版基金评审委员会评审通过的著作，将由该基金资助出版，出版社将努力做好出版工作。

基金还支持两社列选的国家高科技重点图书和国家教委重点图书规划中计算机学科领域的学术著作的出版。为了做好选题工作，出版社特邀请“中国计算机学会”、“中国中文信息学会”帮助做好组织有关学术著作丛书的列选工作。

热诚希望得到广大计算机界同仁的支持和帮助。

清华大学出版社
广西科学技术出版社
计算机学术著作出版基金办公室

1992 年 4 月

中文信息处理丛书编委会

中文信息处理丛书

序　言

中文信息处理技术在我国现代化及信息化建设中，越来越起着重要的作用，作为一个高新技术的重点，它已经列入国务院批准的“国家中长期科学技术发展纲领”。十几年来，我国的中文信息处理领域里，在技术的研究、产品的开发以及产业的建立等方面都取得了显著的成绩。现在很需要把这些方面的成果加以综合并且提炼出来，以便推广应用，并且作为一个起点，再上一个新台阶。这就是我们组织编写并出版这套中文信息处理丛书的目的。

在这套丛书即将开始出版之际，我愿向读者介绍以下两点：

第一　为什么我们要把中文信息处理技术作为高新技术的一个重点来发展呢？

我们日常工作中的信息，绝大部分是以语言文字作为媒介，传播交换和记载的。因此随着计算机的推广应用，由数据处理、信息处理发展到知识处理，对语言文字的处理的要求的深度和广度越来越高。这个问题在西方国家并不突出。因为计算机从诞生之日开始，就是以处理西方语言为基础的。换言之，他们无须经过呼吁和宣传，随着计算机的推广应用的发展，很自然地都会主动地研究和解决自己国家使用计算机如何不断地适应自己国家的语言文字问题。可惜，我们的汉语与西方语言的差别很大。能够处理西方语言的计算机，面对汉语，却显得无能为力。例如：

• 西方语言为拼音文字，而汉语是表意文字。西文字符只有 20 余个，而汉语文字仅常用的就有六、七千个，总数超过五万。这是一个根本性的问题。仅这一个差异就引起了处理汉语的计算机与处理西方语言的计算机一系列的差异，需要我们自己去解决。包括键盘输入、汉字打印与显示、内部代码、汉字识别、程序语言的数据类型、数据库的检索和排序等等。

• 西方的书面语言，词与词之间有空格。而汉语的词与词之间无空格。于是词的切分问题就成了计算机处理汉语的首要问题。

• 西方语言的同音词很少，而汉语的同音词很多。例如，JI 音汉字就有一百多个。辨析同音词就成了汉语语音处理的关键。

• 西方语言多有形态变化（例如：多数、少数，过去、现在，男、女等等），而汉语缺少形态变化。计算机对汉语的处理（例如，机器翻译、人机接口等）无法利用形态，只能在语法、语义上找出路。

• 汉语的语法尚未形成规范化，而且人们习惯于非规范化的语法。于是语义的研究的重要性比西方语言重要得多。例如，“吃饭”“吃大碗”和“吃食堂”的理解只能靠语义来解决。

• 汉语的自动（计算机）处理是多学科和跨学科的研究工作，特别需要计算机科学与语言学的密切结合，而且要依靠长期积累的语言学的研究成果。但我国语言学界多着重汉语教学，对象是人，而不是机器，因此对其丰硕的研究成果要经过改造、深化、量化，甚至要从头开始。要清醒地认识到它的艰巨性，要持续不懈地抓下去。

以上只是几个突出的问题。还有很多其它问题，不再赘述。这些语言上的特点造成了计算机处理汉语的很多障碍，每前进一步都会遇到新问题，使我们不得不花费自己很多力量去解决。

再就计算机的发展趋势而言，计算机产业面临转型期，多媒体和笔记本式计算机将成为热门产品。这些产品的核心技术无不与中文信息处理技术有关。因此，加强中文信息处理的研究更为必要。

第二　中文信息处理技术包括哪些科目呢？

大体上包括下列一些科目：

- 词的切分和频率统计
- 汉语句型和短语的研究及频率统计
- 汉语语义的研究
- 键盘和非键盘汉字输入技术及处理系统
- 汉语语料库的开发及应用
- 汉字的机器代码，程序设计语言的数据类型
- 汉语开放系统的接口规范
- 语声输入与合成
- 汉字识别
- 字形生成
- 汉语分析及理解
- 汉语生成
- 人机接口
- 机器翻译
- 情报检索
- 自动标引和抽词，自动文摘
- 全文检索
- 电子印刷出版系统
- 汉语辅助教学
- 电子词典

以上这些科目，有些是基础研究，有些是技术研究，也有些可以直接转化为产品。这些科目的分类并非学科分类，不过是按照编者本人日常接触的项目，把它们罗列出来而已。其分类的科学性、正确性和完整性尚待商榷。必须指出，有些基础性研究虽然看不到直接的经济效益，但它的研究成果则是其它研究工作所必需，而且要先行。

到目前为止，在上述这些项目中，有些已经产业化，例如电子印刷出版和少数几个汉字输入系统；有些项目已经商品化，正向产业化迈进；很多项目已经实用化。但每个领域都有很多问题等待我们去解决。今后的工作只能加强，不能削弱，使我们中文信息处理的每个领域，每个项目都沿着实用化、商品化和产业化的道路奋勇前进。我相信我们这套丛书必将在促进中文信息处理技术的发展方面发挥它应有的作用。这套丛书大约十册左右，将在“八五”期间陆续出版。

最后，感谢“计算机学术著作出版基金评审委员会”把出版中文信息处理丛书列入了“八五”出版计划。感谢清华大学出版社和广西科学技术出版社给予出版基金的支持。

中国中文信息学会理事长　陈力为

1992 年 5 月　于北京

序

语言学是一门古老的科学，是一个民族相互交际的最重要工具。长期以来都是以手工方式进行研究的。然而进入本世纪20年代以来，语言学在现代科学体系中的地位有了急剧的变化。人们认为语言是哲学和人文科学发展的突破口，是社会科学、自然科学与思维科学的接合部，成了一门带头的科学。所以会发生这种变化，固然由于人们对语言所具有的文化本原性，也是和当前科学技术发展的影响密切相关的。到了50年代，一门新的利用计算机研究语言的学问，计算语言学（即自然语言理解）问世了。它不但极大地推动了语言学本身的发展，而且形成了一门深入到人类活动的各个领域、具有广泛应用价值的语言工程学。本书就是一本介绍这门学科的少有的好书。

自然语言理解真正成为一种实用的学科，是60年代以后的事。1962年国际上成立了计算语言学协会，有关的研究走上了有组织阶段，并形成一门以计算语言学理论为基础的语言工程学科。它广泛地应用于智能计算机人机接口；机器人语音对话；电话翻译系统；大型数据库自然语言查询；专家系统自然语言接口；CAD，CAI和OA的人机交互系统；计算机自动书写，摘要提取，文档自动分类和文书管理系统；大型工业操作过程的自动化语言；机器翻译和机助翻译；自然语言语音通讯；文学与社会科学的文档和语料计算机自动处理；……等等。它成为了当前最热门的研究课题之一。

但是对于这样一门重要的学科，比较深入地介绍这方面的专业书籍却十分缺乏，介绍汉语理解方面的就更少了。该书作者把自己八年来从事计算语言方面的研究和研究生教学过程中的经验编写成书，从多方面收集了该领域当代最重要的理论和方法，包括有形式语言和短语结构文法、上下文无关文法、转换文法、扩充的上下文无关文法、语义网络、SUSY的命题逻辑、概念依从理论、故事表示、集聚理论、特性与集合、词汇功能文法、合一文法、语料库语言学等等，并特别注意汉语的计算机处理问题。与此同时，他们还把自己的关于计算机的汉语理解，以及汉语理解的“词汇语义驱动”理论和方法介绍给大家。这是十分难能可贵的。该书的最后，还介绍了如何利用这种方法实现汉语分析和机器翻译等等。确实是一本极为需要的书籍。相信它的出版，必将为中文信息处理和计算语言学的理论和技术在我国的普及推广，发挥积极的作用。

陈力为

1994年2月26日

前　言

有关自然语言理解(Natural Language Understanding,NLU)方面的书籍很多,但是中文书却很少,除了刘开瑛和郭炳炎等先生的一本介绍自然语言处理(Natural Language Processing,NLP)之外,还没见到其他国家出版社正式出版的书籍。学术界的很多朋友希望我把这几年关于"自然语言理解"的研究生讲稿和我们的工作整理出来,供大家参考。这本书就是由于这样一个原因出版的。

自然语言理解是研究计算机如何理解人类语言的学问。大约在计算机问世的初期,人们就想,如果计算机能够理解人的话,懂得人们写的是什么。那么我们使用计算机时,只要告诉它要做什么,它就按理解的去做,那就太好了,太容易使用了。这使得计算机真的像个电脑,让它做什么它就懂得做什么。但是在当时的条件下,这只是一种梦想。为了实现这一梦想,甚至还出现过技术上的危机,认为这样的梦想是不可能的。到了20世纪的今天,情况有很大的变化,计算机的功能、容量和速度都有几个数量级的提高,自然语言处理的理论研究方面也有巨大进展。因此人们又想起了这个梦想,很多人再度为此努力奋斗,特别是新一代计算机和机器人关于人机接口系统的研究,使得梦想逐渐变成现实。自然语言理解的研究(也称为计算语言学)正成为计算机科学界热门课题之一。

本书是为了这样三种需要而编写的:首先是对自然语言理解感兴趣,从事计算语言学、智能计算机人机接口、机器人语音对话、大型数据库自然语言查询、专家系统、计算机自动书写、摘要提取、文档自动分类和文书管理系统、大型工业操作过程的自动化语言、机器翻译与机助翻译、文学与社会科学的文档和语料计算机自动处理、CAD,CAI和OA的人机接口等研究工作感兴趣或已从事这些方面工作的专业工作者服务的,它是一本自学书和参考书;第二是为那些不了解自然语言理解的计算机专业人员服务的。由于这个领域几起几落的历史,它的研究内容与计算机科学的其它领域研究的内容有很多不同,很多人对它不了解,它可为在设计自己的系统时需要人机接口方面知识的人服务。最后,本书又可用作为讲授"自然语言理解"这门40学时的研究生课程的教材。

书中内容考虑到语言信息处理的需要,包括有引言、汉语的计算机理解、语法分析、语义分析、概念分析、故事表示、集聚理论、特性和公式、词汇功能文法、功能合一文法、语料库语言学和机器词典等等。较为深入地涉及当代计算语言学最感兴趣的理论和方法,并特别注意汉语的计算机处理问题。与此同时,在第十二到十五章,以及三个附录,把我们关于汉语理解和汉英机器翻译的工作介绍给大家,也就是关于机器词典、词汇语义驱动理论、中间转接语言、目标语言生成、语义关系集、规则描述语言和汉英机译的实例等。当然,我们的工作虽已基本完成,但还在不断的完善,任重而道远,有很多不成熟的地方,请批评指正,并希望得到学术界的认可。

所有这些内容都是我八年来讲课和我们研究组同志们共同努力工作的结果。在书稿的形成和后来的不断修改过程中,重新整理了我的讲稿,较多地吸收了近期《国际计算语言学

学报》、《中文信息学报》和同行们著作里的一些有意义的内容，希望尽可能反映当代最新国内外的学术思想和内容。同事们和学生们也作了不少补充，甚至重写。相信对有志于研究这方面工作的同行们会有一点帮助的。

参加本书初稿编写工作的有：张桂平（第 4 章）、滕永林（第 5 章）、鲍志斌（第 6 章）、李渝生（第 7 章）、寇育新（第 8 章）、李晶姣、周强、郭宏蕾（第 11 章）、唐泓英（第 12，13 章）、刘东立（第 12，14 章）、王宝库（第 13 章，附录 2）、卞世力（第 15 章）等。刘东立，唐泓英，王宝库，李晶姣等还参加了本书最后的复校工作。书中的全部文字都是由马波录入的，在此表示感谢。

本书在编写过程中，自始至终都得到陈力为院士在精神鼓励和学术指导方面的巨大帮助，并给以热情推荐，使本书得以顺利出版。书中有关介绍我们自己的工作部分，也是由于多年来一直得到国家自然科学基金委员会、863 国家高技术智能计算机专家组和国家教委博士点基金等等的资助。没有他们长期的支持，本书也是不可能出版的。在此一并表示感谢。

书中的内容虽然经过仔细校对，但错误和不当之处难免，请批评指正。

姚天顺

于东北大学

1993 年 5 月

目　录

第一章　引言 …… 1
1.1　NLU 是人工智能领域的一个分支 …… 1
1.2　知识处理问题 …… 1
1.3　自然语言理解及其研究内容 …… 3
1.4　自然语言理解的研究近况 …… 5

第二章　汉语的计算机理解 …… 8
2.1　引言 …… 8
2.2　汉语理解中的特殊问题 …… 8

第三章　语法分析 …… 14
3.1　语法分析的任务 …… 14
3.2　短语结构语言 …… 14
3.3　早期系统：上下文无关分析器 …… 19
3.4　转换分析器：第一类系统 …… 28
3.5　扩充的上下文无关分析系统 …… 37

第四章　语义分析 …… 50
4.1　表示语义的逻辑语言 …… 50
4.2　语义网络 …… 53
4.3　用于语言表示的命题语言 …… 62

第五章　概念分析 …… 82
5.1　概念依从理论 …… 82
5.2　概念分析 …… 92
5.3　概念记忆和推理 …… 98
5.4　概念生成 …… 105

第六章　故事表示 …… 110
6.1　脚本 …… 110
6.2　规划 …… 111
6.3　目标 …… 113
6.4　脚本表示 …… 119

6.5 规划表示 …… 120
6.6 宏观与微观事件表示 …… 124
6.7 一个故事 …… 127

第七章 词汇集聚理论 …… 133
7.1 词的集聚性 …… 133
7.2 义类词库和词汇集聚 …… 135
7.3 寻找词汇链 …… 137
7.4 利用词汇链确定文本结构 …… 141

第八章 特性和公式 …… 154
8.1 特性结构 …… 154
8.2 特征结构的公理化和一阶逻辑公式 …… 158
8.3 结论 …… 165

第九章 词汇功能文法 …… 166
9.1 引言 …… 166
9.2 功能文法 …… 167
9.3 LFG 的两个语法层次结构 …… 169
9.4 功能合格条件 …… 174
9.5 LFG 理论的进一步的内容 …… 176

第十章 功能合一文法 …… 181
10.1 引言 …… 181
10.2 功能描述 …… 182
10.3 合一运算 …… 184
10.4 句子的功能描述 …… 188
10.5 简单的合一文法 …… 191

第十一章 语料库语言学 …… 193
11.1 引言 …… 193
11.2 国外语料库简介 …… 194
11.3 统计学的基本知识及其自然语言处理 …… 198
11.4 汉语语料库词类的自动标注 …… 201
11.5 用于语料库分析的词汇语义技术 …… 207

第十二章 机器词典 …… 215
12.1 基本概念 …… 215
12.2 电子词典的结构 …… 216

12.3 基本词典…… 217
12.4 词典的内部结构…… 225
12.5 词语搭配及搭配词典…… 228
12.6 开发电子词典的集成系统…… 231

第十三章 词汇语义驱动…… 233
13.1 概述…… 233
13.2 复杂特征集…… 234
13.3 词汇语义驱动…… 237

第十四章 中间转接语言的表示法…… 248
14.1 引言…… 248
14.2 基本概念…… 249
14.3 中间转接语言表示法…… 253

第十五章 生成器中的词汇语义驱动方法…… 260
15.1 生成器中的复杂特征集…… 260
15.2 词项位…… 263
15.3 英文生成中的合一与扩展运算…… 264
15.4 英文生成的词汇语义描述…… 268

附录 1 语义关系…… 277
附录 2 规则描述语言…… 282
附录 3 一个汉英机译实例…… 299
附录 4 汉语和英语的词性和下位词性…… 307

汉英专业名词对照表…… 313
英汉专业名词对照表…… 321
参考文献…… 329

Contens

Chapter 1 Introduction ····· 1
1.1 NLU is a Branch of AI ····· 1
1.2 The Problems of Knowledge Representation ····· 1
1.3 Natural Language Understanding and Its Contents ····· 3
1.4 Recent Developments in NLU Research ····· 5

Chapter 2 Chinese Machine Understanding ····· 8
2.1 Introduction ····· 8
2.2 The Special Problems in Chinese Understanding ····· 8

Chapter 3 Syntatic Analysis ····· 14
3.1 Task of Syntatic Analysis ····· 14
3.2 Phrase Structure Language ····· 14
3.3 Early System:Context Free Analyzer ····· 19
3.4 Translational Analyzer:First Class System ····· 28
3.5 Augment Context-Free Analysis System ····· 37

Chapter 4 Semantic Analysis ····· 50
4.1 Logical Language for Text Representation ····· 50
4.2 Semantic Network ····· 53
4.3 Propositional Language for Text Representation ····· 62

Chapter 5 Conceptual Analysis and Generation ····· 82
5.1 Conceptual Dependency ····· 82
5.2 Conceptual Analysis ····· 92
5.3 Conceptual Memory and Inference ····· 98
5.4 Conceptual Generation ····· 105

Chapter 6 Represantation of Stories ····· 110
6.1 Script ····· 110
6.2 Plan ····· 111
6.3 Goal ····· 113
6.4 Representation of Scripts ····· 119