中文信息处理丛书

自然语言理解

——种让机器懂得人类语言的研究

（第 圆版）

姚天顺摇朱靖波
张摇琍摇杨摇莹摇编著

清 华 大 学 出 版 社

（京）新登字 员缘愿号

内摇容摇简摇介

摇摇自然语言理解是人工智能的一个重要分支，主要研究如何利用计算机来理解和生成自然语言。本书重点介绍了自然语言理解所涉及的各个方面，包括语法分析、语义分析、概念分析、语料库语言学、词汇语义驱动、中间语言、宰燥罪藻藏词汇树邻接文法、链接文法、基于语段的机器翻译方法、内识别与文本过滤、机器翻译的评测等，既有对基础知识的介绍，又有对最新研究进展的综述，同时还结合了作者多年的研究成果。

本书可作为高等院校计算机、人工智能等专业的高年级本科生或研究生的教材及教学参考书，也可供从事中文信息处理、数据挖掘以及人工智能研究的相关人员参考。

书摇摇名：自然语言理解
作摇摇者：姚天顺摇朱靖波摇张琍摇杨莹摇编著
出 版 者：清华大学出版社（北京清华大学学研大厦，邮编 员园园园愿源
课题：
责任编辑：薛慧
封面设计：
印 刷 者：北京昌平环球印刷厂
发 行 者：新华书店总店北京发行所
开摇摇本：苑愿苑伊员园怨圆 员辕员远摇印张：猿圆摇字数：苑园缘千字
版摇摇次：圆园园圆年 员园月第 圆版摇圆园园圆年 员园月第 员次印刷
书摇摇号：陨杂月晕 苑鄄猿园圆鄄园缘猿远苑鄄源辕栽孕· 猿园苑园
印摇摇数：园园园员～ 园园园园
定摇摇价：园园援园园元

中文信息处理丛书编委会

中文信息处理丛书

序摇摇言

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”。我国的中文信息处理事业正在不断向前推进,在技术研究、产品开发以及产业化发展等方面都取得了显著的成绩。现在有必要把这些方面的成果加以综合、提炼,以便更好地推广应用,并且作为一个起点,再上一个新台阶。中文信息处理在从汉字信息处理进入汉语信息处理之后,在从单机信息处理进入网络信息处理之后,已经面临着新的更大的挑战和机遇,需要我们重新对中文信息处理进行全面的审视与整合,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。

在这套丛书出版之际,我愿向读者介绍以下几点:

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?

语言文字是信息的首要载体。我们日常工作中的信息,绝大部分是以语言文字表达、记载、传播和交换的。因此随着计算机和因特网的推广应用,由数据处理、信息处理发展到知识处理,对语言文字处理要求的深度和广度越来越高,可以认为一个国家的语言文字的信息处理水平和处理量基本上代表了这个国家进入信息社会的程度,其语言文字信息的处理能力直接关系到它在网络社会和网络经济中的国际竞争能力。目前,网络社会和网络经济正以我们难以预料的速度在全世界发展,其阻碍发展的首要瓶颈问题就是自然语言的处理问题。网络社会也是人类社会,网络经济也是人类经济,需要以自然语言作为社会交际工具,一旦基于网络的自然语言处理问题得到突破,网络社会和网络经济将会突飞猛进。我们要在下一个世纪成为世界强国,就不能不把语言文字信息处理技术作为高新技术的一个重点来发展。在世界一流高新技术企业纷纷在中国设立“中国研究院”,争先把“中文信息处理”作为研究的重中之重的时候,我们当然要抢占中文信息处理这个高新技术发展的制高点。

第二,中文信息处理与印欧语系的语言信息处理的不同之处是什么?

计算机从诞生之日开始,就是以处理印欧语系的语言为基础的。换言之,计算机对于印欧语系的自然语言处理具有较好的支撑能力,计算机的推广应用在语言文字信息处理方面受到的阻力较小。我们的汉语却与印欧语系的语言差别很大,能够处理那些语言的计算机,面对汉语汉字,却显得无能为力。例如:

· 印欧语系为拼音文字,所使用的字符仅二十余个,而汉语是意音文字,常用的汉字就有六七千个,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印欧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。

- 印欧语系的书写,词与词之间有空格,而书面汉语的词与词之间无空格,于是词的机器自动切分问题就成了计算机处理汉语的首要问题。
- 印欧语系的同音词较少,而汉语的同音词较多。例如,仅在《现代汉语词典》中 [illegible]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。
- 印欧语系多有形态变化 (例如 :复数、单数,过去、现在,阴性、阳性等等) ,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) 无法利用形态变化,只能在句法、语义上找出路。
- 汉语的语法研究尚未形成规范化,而且人们习惯于约定俗成的语法。于是语义研究显得尤其重要。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、认知科学等学科的密切结合,而且要依靠长期积累的语言学的研究成果。但我国语言学界过去的研究多着重汉语教学,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、形式化,甚至要从头开始。要清醒地认识到面向机器的汉语研究的艰巨性,要持续不懈地抓下去。

以上只是几个突出的问题,还有一些其他问题,不再赘述。这些语言上的特点造成了计算机处理汉语的众多障碍,每前进一步都会遇到新问题,我们不得不花费比印欧语系的信息处理多得多的力量去解决。

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成为热门产品,计算机从单机进入网络,网上的汉语信息处理正在成为强势和主流,并对语言文字的信息处理提出新的要求。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,取得网络化的自然语言处理的突破更为必要。

第三, 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
- 汉语分析及篇章理解
- 汉语生成
- 人机接口

- 汉外外汉机器翻译
- 信息检索
- 自动标引和抽词,自动文摘、文本自动分类与网站自动分类、信息自动提取与知识挖掘
- 全文检索
- 电子印刷出版系统
- 汉语辅助教学与现代远程教学
- 电子词典等

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他研究工作所必需的,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。这套丛书将陆续出版。

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。

中国工程院院士 陈力为
[illegible]年 [illegible]月于北京序
[illegible]年 [illegible]月于北京修订

序摇摇言

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 圆园世纪 圆园年代以来，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 缘园年代，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，那是 远园年代以后的事。员怨远圆年国际上成立了计算语言学协会，使得研究走上了有组织的阶段，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；悦粤阅、悦粤酝和 悦粤陨的人机交互系统；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；自然语言语音通信；国际互联网上的信息分类、浏览、过滤，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、命题逻辑语言、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，他们还把自己关于计算机的汉语理解，以及汉语理解的“概率词汇语义驱动”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

员怨怨苑年员圆月远日

目摇摇录

第 员版前言

有关自然语言理解（[illegible]）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（[illegible]）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语言，懂得人们写的是什么，那么，在我们使用计算机时，只要告诉它要做什么，它就按理解的去做，那就太好了。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 圆园世纪的今天，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，自然语言理解的理论研究有了很大进展。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、[illegible]和[illegible]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；第二个是为那些不了解自然语言处理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他领域研究的内容有很多不同，很多人对它不了解，所以，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，本书又可用作讲授“自然语言理解”或“计算语言学”这门 源园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、特性与公式、词汇功能文法、功能合一文法、语料库语言学和机器词典等。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 员园章到 员缘章，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。

员源

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的结果。在书稿的形成和后来不断修改过程中，又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忙的。

参加本书初稿的编写工作的有（以章节为序）姚天顺（第员圆猿源缘苑愿员圆章）、张桂平（第源章）、滕永林（第缘章）、鲍志斌（第远章）、李渝生（第苑章）、寇育新（第愿章）、李晶皎、周强、郭宏蕾（第员员章）、唐泓英（第员圆员猿章）、刘东立（第员圆员源章）、王宝库（第员缘章，附录圆）、卞世力（第员远章）等。最后，本人对全书进行了整体编辑和一致性修改。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的鼓励和学术指导方面的巨大帮助，以及热情推荐，致使本书得以顺利出版。书中介绍的我们自己的工作部分，也多年来一直得到国家自然科学基金委员会、愿猿猿国家高技术智能计算机专家组和国家教委博士点基金会等项目的资助。没有他们长期的支持，本书是不可能出版的。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

员怨怨猿年 缘月

第 圆版前言

本书第 员版到现在已经七年了。一本计算机的专业书籍这么多年还能再版，是一件幸事。这是因为当年写书时，特别注重基本内容和基本概念的介绍，加上很多学校把它当作研究生的教科书和选修课的教学参考书，再加上清华大学出版社编辑部的支持，所以就再版了。这是要十分感谢大家的。但是，这本书毕竟还是过去了七八年。社会在发展，学科也在发展。特别是国际互联网的出现，经济全球化了，科学研究和社会也将与世界密切相关。很多过去没有的，或者不十分重要的东西现在变得重要了。例如网上的信息获取、交叉语言信息检索、文本主题识别、多国语机器翻译等、自然语言查询，特别是反黄反黑的信息过滤，全世界都在注意，成为信息安全的重大课题。网络语言信息处理的研究受到了前所未有的重视。

这些年来，原来非常看好的电子商务和网站公司，首先从美国开始大幅度地暂时低落，接着影响全世界。原因可能有好多种，如果仅从技术上考虑，网络信息的识别、理解和处理也是一个重要因素。例如我们还没有办法在网上作这样的查询，找一家质量好而又便宜的生产某产品的公司。又例如，我们现在还没有办法从现有的 缉种不同手册中，自动找出所有手册中提出的问题和解决的办法，以支持电子商务的主动服务。都还没有一个好的解决办法。

但是，这些年有没有进步呢？有的，大家还是在努力，有不少进步，总结起来大致有这样四个方面的变化和发展：

基于规则的理论和方法取得不断的改进和发展；语料库语言学的迅速兴起并取得一系列的成果；大容量基于结构化的带标树库的建立及其应用；大粒度的基于语块的语言分析和处理方法的出现等。

因此，我们在再版本书时，除了保留原有基本内容外，着力于添加上述变化了的东西。希望能引起读者思考和找到解决问题的方法。在这同时也修改了原书不易理解和出错的部分。这样，本书就在原书基础上整体性地作了改进，具体的包括：

删掉了电子词典那一章；仔细地修改了引言、语法分析、语义分析、概念分析、语料库语言学、词汇语义驱动、中间语言等 苑章；添加了 宰煽[illegible]词汇树邻接文法、链接文法、基于语段的机器翻译方法、内容识别与文本过滤以及机器翻译的评测等 远章。总加起来，修改的部分总共有 员袁章之多，大部分内容都有了变化。另外在每章后面，为了便于自学和研究生教学，添加了思考题。参考文献也改成按章排列。

本书在再版过程中，特别要感谢责任编辑薛慧女士和其他编辑们，他们的精心审阅、修改和建议，为提高本书的质量起着十分重要的作用。

再版书的修改仍由姚天顺负责，主要的参与者有朱靖波、张琍、杨莹等。具体的是（按章序排列）：张琍（第五，六 章，附录 耘、朱靖波（第十三，十八章）、王大禹（第十一

章)、吕学强(第十二,十七章)、李沐(第十四章)、杨莹(第十九章)和战学刚等。参加部分章节修改的还有陈文亮、刘东立、孙杰、林鸿飞、麻志毅、李珩、张跃、高维君和宋鸿岩等。在此一并表示感谢,他们的辛苦为本书增色不少。

姚天顺

于东北大学