重点大学计算机专业系列教材

中文文本信息处理的原理与应用

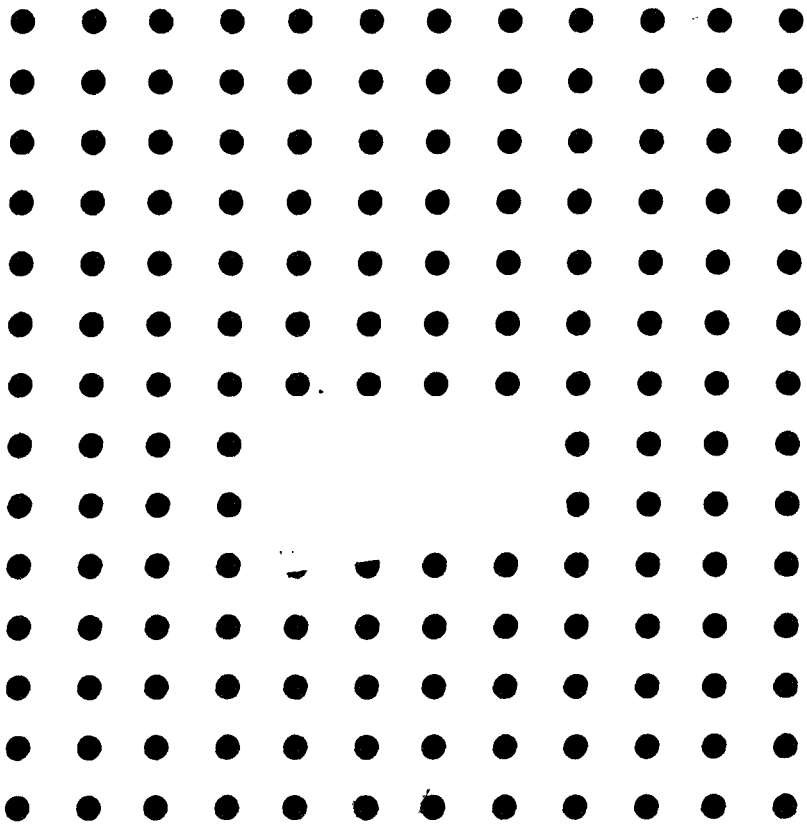
苗夺谦　卫志华　编著

清华大学出版社

重点大学计算机专业系列教材

中文文本信息处理的原理与应用

苗夺谦　卫志华　编著



清华大学出版社
北京

内 容 简 介

本书是一本全面系统地介绍中文文本信息处理的教材，内容丰富，由浅入深地讲述了中文文本信息处理的原理与应用。本书不仅介绍了基于规则的自然语言分析方法，也介绍了基于统计学的方法。全书共分为四大部分，分别是词法分析、语法处理、语义分析和应用与技术。其中前三部分是自然语言处理的基本理论，第一部分针对中文处理中特有的分词问题，介绍了自动分词算法、分词中歧义的消除和未登录词的识别算法，另外还介绍了语料库的相关知识。第二部分和第三部分都是从语法（语义）的表示入手，将自然语言形式化，再给出语法（语义）分析的算法，并针对该过程中的歧义问题给出了一些成熟的解决方案。最后一部分讲述自然语言理解在信息检索、信息抽取、自动文摘和文本分类等领域的应用。本书思路清晰，在每部分及每章的开始都介绍了该部分知识与其他部分之间的关系，以及该部分的知识点之间的关系，以帮助读者从整体上把握中文文本信息处理的思路，并能根据不同的需求或不同的问题选择适当的算法。

本书涉及内容广泛，能满足不同水平读者群的需求，可以作为计算机、信息类高年级本科生的教材，也可作为自然语言处理方向研究生的教材，也非常适合作为自然语言处理应用领域的研究人员和技术人员的参考资料。

图书在版编目（CIP）数据

中文文本信息处理的原理与应用/苗夺谦，卫志华编著. —北京：清华大学出版社，2007.9
（重点大学计算机专业系列教材）
ISBN 978-7-302-15498-3

Ⅰ. 中…　Ⅱ. ①苗… ②卫…　Ⅲ. 汉字信息处理　Ⅳ. TP391.12

中国版本图书馆 CIP 数据核字（2007）第 089972 号

责任编辑：魏江江　林都嘉
责任校对：李建庄
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　**邮　　编**：100084
c-service@tup.tsinghua.edu.cn
社 总 机：010-62770175　**邮购热线**：010-62786544
投稿咨询：010-62772015　**客户服务**：010-62776969
印 刷 者：北京市清华园胶印厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185×260　**印　　张**：21.25　**字　　数**：513 千字
版　　次：2007 年 9 月第 1 版　　**印　　次**：2007 年 9 月第1 次印刷
印　　数：1～4000
定　　价：29.00 元

INTRODUCTION

出版说明

随着国家信息化步伐的加快和高等教育规模的扩大，社会对计算机专业人才的需求不仅体现在数量的增加上，而且体现在质量要求的提高上，培养具有研究和实践能力的高层次的计算机专业人才已成为许多重点大学计算机专业教育的主要目标。目前，我国共有16个国家重点学科、20个博士点一级学科、28个博士点二级学科集中在教育部部属重点大学，这些高校在计算机教学和科研方面具有一定优势，并且大多以国际著名大学计算机教育为参照系，具有系统完善的教学课程体系、教学实验体系、教学质量保证体系和人才培养评估体系等综合体系，形成了培养一流人才的教学和科研环境。

重点大学计算机学科的教学与科研氛围是培养一流计算机人才的基础，其中专业教材的使用和建设则是这种氛围的重要组成部分，一批具有学科方向特色优势的计算机专业教材作为各重点大学的重点建设项目成果得到肯定。为了展示和发扬各重点大学在计算机专业教育上的优势，特别是专业教材建设上的优势，同时配合各重点大学的计算机学科建设和专业课程教学需要，在教育部相关教学指导委员会专家的建议和各重点大学的大力支持下，清华大学出版社规划并出版本系列教材。本系列教材的建设旨在“汇聚学科精英、引领学科建设、培育专业英才”，同时以教材示范各重点大学的优秀教学理念、教学方法、教学手段和教学内容等。

本系列教材在规划过程中体现了如下一些基本组织原则和特点。

1. 面向学科发展的前沿，适应当前社会对计算机专业高级人才的培养需求。教材内容以基本理论为基础，反映基本理论和原理的综合应用，重视实践和应用环节。

2. 反映教学需要，促进教学发展。教材要能适应多样化的教学需要，正确把握教学内容和课程体系的改革方向。在选择教材内容和编写体系时注意体现素质教育、创新能力与实践能力的培养，为学生知识、能力、素质协调发展创造条件。

3. 实施精品战略，突出重点，保证质量。规划教材建设的重点依然是专业基础课和专业主干课；特别注意选择并安排了一部分原来基础比较好的优秀教材或讲义修订再版，逐步形成精品教材；提倡并鼓励编写体现重点大学计算机专业教学内容和课程体系改革成果的教材。

4. 主张一纲多本，合理配套。专业基础课和专业主干课教材要配套，同一门课程可以有多本具有不同内容特点的教材。处理好教材统一性与多样化的关系；基本教材与辅助教材以及教学参考书的关系；文字教材与软件教材的关系，实现教材系列资源配套。

5. 依靠专家，择优落实。在制订教材规划时要依靠各课程专家在调查研究本课程教材建设现状的基础上提出规划选题。在落实主编人选时，要引入竞争机制，通过申报、评审确定主编。书稿完成后要认真实行审稿程序，确保出书质量。

繁荣教材出版事业，提高教材质量的关键是教师。建立一支高水平的以老带新的教材编写队伍才能保证教材的编写质量，希望有志于教材建设的教师能够加入到我们的编写队伍中来。

教材编委会

FOREWORD

前言

现在是一个在线信息、电子通信和互联网流行的年代,应该看到,商业部门、政府机构以及个人正面对着越来越多与工作、生活密切相关的文本信息,每天都有大量的信息在遍布世界各地的互联网上产生、发布、交换、存储和获取,如何从这些大量文本中挖掘潜在的有使用价值的信息,仍然是一个难题。

采用计算机技术来研究和处理自然语言是20世纪50年代初才开始的。50多年来,这项研究取得了长足的进展,成为了计算机科学中一门重要的新兴学科——自然语言处理(Natural Language Processing,NLP)。现在,语音和语言的计算机处理进入了一个令人振奋的时期,基于互联网的语言处理技术的需求,有力地推动了各种实用的自然语言处理系统的开发。

就处理对象而言,自然语言处理可以分为语音处理和文本语言处理;就人与计算机沟通的方向而言,可以分为自然语言理解(使计算机既能理解自然语言文本的意义)和自然语言生成(计算机以自然语言文本来表达给定的意图、思想等)。中文信息处理是自然语言处理中一个重要的分支,本书着重介绍中文文本信息理解的方法和技术。

美国计算机科学家Bill Manaris给自然语言处理作了如下定义。

自然语言处理是研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力和语言应用的模型,建立计算框架来实现该语言模型,提出相应的方法不断地完善该语言模型,根据该语言模型设计各种实用系统,并探讨这些实用系统的评测技术。

由此定义可以看出,计算机对自然语言的研究和处理一般可以经过以下的过程。

1. 把需要研究的问题加以形式化,使之能以一定的数学形式严密地表示出来,这种形式化的表示能反映该问题的语言学特征;

2. 把这种严密的数学形式表示成算法,使之可以在计算机上实现形

式化；

3. 根据算法编写计算机程序，使之能在计算机上实现；

4. 对所建立的自然语言处理系统进行评测，使之不断地改进质量和性能，以满足实际应用的要求。

从上面的处理过程可以看到，自然语言处理需要计算机科学、数学、语言学、信息论等多个领域的知识。单就文本理解研究而言，建立自然语言处理模型也需要不同层面的知识：

1. 形态学知识：描述语素的结合规律，说明语素是怎样形成单词的；

2. 词汇学知识：描述词汇系统的规律，说明单词本身固有的语义特征和语法特征；

3. 句法学知识：描述单词（或词组）之间的结构规则，说明单词（或词组）怎样形成句子；

4. 语义学知识：描述句子各成分之间的语义关系，这样的语义关系是与情景无关的，说明怎样从构成句子的各个成分推导出整个句子的语义；

5. 语用学知识：描述与情景有关的情景语义，说明怎样推导出句子具有的与周围环境话语有关的各种含义；

6. 外部世界的常识：描述关于语言使用者和使用环境的一般性常识，例如，语言使用者的信念和目的，说明怎样推导出这样的信念和目的的内在结构。

根据不同的应用需求，所需的知识层面有所不同，本书主要介绍词法、语法及语义处理的理论与技术。

需求是推动技术发展和进步的源泉，由于信息时代的强大信息需求，“搜索”被称为“第四桶金”，而作为搜索的关键技术之一的“自然语言理解”技术也必将会显示出它强劲的发展态势。无论作为在校的计算机和信息学专业学生，还是工作在中文理解领域的工程技术人员，都需要夯实自然语言理解的基础。为此，我们希望在计算机专业和信息学专业的高年级本科生和研究生中普及自然语言理解领域的理论和技术。本书正是应这样的需求而编写的，从内容的安排来看，具有如下特点。

1. 内容丰富：随着大规模语料库的建设，基于统计学的方法在自然语言处理的各个层面发挥了越来越重要的作用，尤其是 20 世纪 80 年代以来，基于规则的传统方法逐步让位于基于统计学的方法。但是近些年来的研究表明，单纯运用基于统计学的方法并不能解决深层次的自然语言处理问题。因此，很多学者提倡将规则引导的深层分析和具有鲁棒性的基于统计学的浅层分析结合起来。基于这样的需求，本书介绍了在词法分析、语法处理和语义分析各阶段涉及的基于规则的方法和基于统计学的方法，以利于读者更好地将这两种研究思路结合起来。

2. 理论与应用相结合：正如我们前面所述，需求是推动技术发展的源泉，本书除介绍中文文本信息处理的基本理论外，还给出了它的重要应用，如文本分类、信息检索、信息抽取和自动文摘。这样做的好处是，在介绍理论的同时，可以提供一个背景来理解和模拟特定领域中的应用问题，作为教材，可以为学生提供相应的实验素材。

3. 本地化：目前许多教材都是针对英文的理解而编写的，编者将把研究对象锁定在中文文本处理，系统讲述了中文文本理解的基本理论，列举其研究成果，使之更加适合作

为中文文本理解的工具书。在本书参考文献中给出了许多中文信息处理的资源，读者可以由此找到所需的中文语料和工具集。

本书首先可以作为计算机专业、信息学专业高年级本科生和研究生的教材，由于本书内容丰富，涉及面广，因此，也可以作为自然语言处理领域专业人员的参考书。

由于编者知识水平有限，错误在所难免，诚心地希望广大读者提出批评和指正。

编　者

CONTENTS

目录

第一部分　词法分析

第二部分 语法处理

第三部分 语义分析

第四部分 应用与技术

CHAPTER 1

概　论　第1章

1.1 自然语言处理与中文信息处理

1.1.1 自然语言处理

语言是人类思维的载体，是人际交流的重要工具，也是人们生活中不可缺少的组成部分。自然语言是指人类语言集团的本族语，如汉语、英语、日语等。自然语言是相对于人造语言而言的。人造语言是指世界语或计算机的各种程序设计语。在人类历史上以语言文字形式记载和流传的知识占到知识总量的80%以上。就计算机的应用而言，据统计用于数学计算的仅占10%，用于过程控制的不到5%，其余85%左右都是用于语言文字的信息处理。在信息化社会中，语言信息处理的技术水平和每年处理的信息总量已成为衡量一个国家现代化水平的重要标志之一。

在这样的社会需求下，自然语言处理或称计算语言学，作为语言信息处理技术的一个高层次的重要方向，一直是人工智能领域所关注的核心课题之一。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。这又是一门非常复杂的学科，还涉及数学、语言学、逻辑学和心理学等多个研究领域。

用自然语言与计算机进行通信，这是人们长期以来所追求的。因为它既有明显的实际意义，同时也有重要的理论意义：人们可以用自己最习惯的语言来使用计算机，而无须再花大量的时间和精力去学习不自然和不习惯的各种计算机语言；人们也可通过它进一步了解人类的语言能力和智能的机制。

实现人机间自然语言通信意味着要使计算机既能理解自然语言文本的意义，也能以自然语言文本来表达给定的意图、思想等。前者称为自然语言理解，后者称为自然语言生成。因此，自然语言处理大体包括了自然语言理解和自然语言生成两个部分。本书重点讲述与自然语言理解相关的理论和方法。

首先，需要明确“理解”的含义是什么？

正如什么叫“智能”一样，对于“理解”这个术语也存在着各式各样的认识。然而在人工智能界或者语言信息处理领域中，人们普遍认为可以采用著名的图灵(Turing)试验来

判断计算机是否“理解”了某种自然语言，具体的判别准则至少有如下四条。

（1）回答问题：机器能正确地回答输入文本中的有关问题；

（2）文摘生成：机器有能力产生输入文本的摘要；

（3）释义：机器能用不同的词语和句型来复述其输入文本；

（4）翻译：机器具有把一种语言（源语）翻译成为另一种语言（目标语）的能力。

要做到让计算机真正地理解人的语言并非易事，因为，一个自然语言系统必须使用相当多的语言自身结构的知识，包括什么是词、词如何组成句子、词的意义是什么、词的意义对于句子的意义有什么影响等。然而，如果不考虑构成人类智能的其他方面的因素——人类的一般世界知识和人类的推理能力，就不可能完全揭示人类的语言行为。比如，一个人要回答问题或参与对话，他不仅要会使用这种语言，而且还要知道所讨论话题的背景知识和谈话所处的场景。因此，要让计算机具有这种能力，就需要从语言学知识角度出发构造关于语言理解和生成的计算模型，并且这些模型还要在特定领域背景下表现良好。

从语言学角度，可以归纳出与自然语言理解有关的一些知识。

（1）语音和音韵知识：研究词语与其发音如何关联。这种知识对于基于语音的系统是至关重要的。

（2）词语形态学知识：研究词语如何由被称为词素的更基本的意义单位构成。词素是语言中一种最基本的意义单位，在西文中，比如前后缀，在汉语中，比如偏旁部首。

（3）句法知识：研究词语是如何排列以组成正确的句子，并决定每个单词在句子中所充当的结构角色，以及短语之间的构成关系。

（4）语义知识：研究词语的意义以及在句子中词语意义是如何相互结合以形成句子意义的。这是上下文无关的意义研究，即一个句子在不考虑其上下文的情况下所具有的意义。

（5）语用知识：研究句子如何在不同情形下被使用，以及这种使用如何影响句子的解释。

（6）篇章知识：研究在前面句子的影响下，下面的句子如何解释，主要包括代词指代的解释和信息中所包含的时态解释等。

（7）世界知识：包括了语言所处的背景知识，这种知识对于语言的理解和使用是必需的。

以上几方面的语言知识代表了自然语言理解的不同层面，事实上，一般的自然语言处理系统都会涉及其中的多个层面。

正如人们在人工智能领域和计算机领域中最常用的思维方式，在这里，首要的任务是将研究对象在计算机中表示出来，即如何将上述的不同层面的语言学知识在计算机中表示出来，在此基础上，才能完成文本意义的计算，也就是文本意义的解释（理解）。然而，语言并不像数学公式那样严格，相反，语言中存在广泛的歧义性，比如：在词的层面就有一词多义和多个词同义的问题，一个句子在不同语言环境中也有不同的含义，至于对篇章的理解就更加仁者见仁，智者见智了。可见，在每个层面的语言表示和解释中都涉及歧义消解，因此，歧义消解是自然语言理解的一个基本问题。