信号与信息处理丛书

计算机自然语言处理

王晓龙　关　毅　等 编著

清华大学出版社

信 号 与 信 息 处 理 丛 书

计算机自然语言处理

王晓龙　关　毅　等 编著

清华大学出版社
北 京

内容简介

计算机自然语言处理技术在我国现代化及信息化建设中起着越来越重要的作用，我国政府已经将它列入“国家中长期科学技术发展纲领”。近年来，语言处理技术，特别是基于国际互联网的中文语言处理技术正在引起我国广大科技工作者的高度重视。

本书既全面阐述了中文语言处理技术的特殊规律，又借鉴了国内外学者在计算语言学领域里的最新成就，还包括了作者的实践经验和体会。

本书可以作为计算机相关专业研究生的专业课教材，也可供相关专业高年级大学生和从事自然语言处理技术研究和应用的科技人员参考。

图书在版编目(CIP)数据

计算机自然语言处理/王晓龙，关毅等编著. —北京：清华大学出版社，2005.4
(信号与信息处理丛书)
ISBN 7-302-10089-6

Ⅰ. 计…　Ⅱ. ①王…　②关…　Ⅲ. 自然语言处理　Ⅳ. TP391

中国版本图书馆 CIP 数据核字(2004)第 130469 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969
组稿编辑：陈国新
文稿编辑：马幸兆
版式设计：肖　米
印 刷 者：北京市清华园胶印厂
装 订 者：三河市李旗庄少明装订厂
发 行 者：新华书店总店北京发行所
开　　本：185×260　**印张**：11.5　**字数**：260 千字
版　　次：2005 年 4 月第 1 版　2005 年 4 月第1 次印刷
书　　号：ISBN 7-302-10089-6/TP・1035
印　　数：1～3000
定　　价：23.00 元

丛书出版说明

FOREWORD

信号与信息处理可以说是信息技术中的核心部分。随着信息科学与技术的飞速发展，随着信息技术深入到各个领域而得到广泛的应用，信号与信息处理也作为前沿技术而发生着重大的变化。编辑出版“信号与信息处理丛书”正是为了反映这种变化，为了加速培养这方面的人才，也为了进一步推动这一领域的发展。本丛书的内容力求能反映信号与信息处理技术的前沿内容，具有高的学术意义与应用价值。入选的书稿可以是创作的专著，也可以是高水平的译作。

这套丛书不仅适合于作研究生教学参考之用，也可作为高校教师与有关领域研究人员学习与工作的参考书。

从历史来看，真正影响着生活的是不断增长的知识与技术的积累和经反复探索所形成的观念。相信这套丛书的出版，会增加正在成长中的信号与信息处理技术的积累，而它对生活的作用则是显而易见的。

李衍达

2004 年 8 月 24 日

前言

FOREWORD

计算机自然语言处理是用计算机通过可计算的方法对自然语言的各级语言单位(字、词、语句、篇章等)进行转换、传输、存储、分析等加工处理的科学,是一门与语言学、计算机科学、数学、心理学、信息论、声学等相联系的交叉性学科。国际互联网技术的飞速发展,极大地推动了信息处理技术的发展,也为信息处理技术不断提出新的需求。语言作为信息的载体,语言处理技术已经日益成为全球信息化和我国社会与经济发展的重要支撑性技术。

本书全面阐述了自然语言处理技术的基本原理和实用方法,全书共分为基础、原理和应用3个篇章。第1章是概论;第2、第3章构成本书的基础篇,论述了自然语言处理技术的数学基础和中文语言处理特有的自动分词技术;第4、第5章构成本书的原理篇,分别论述了基于统计和基于语言学规则的语言处理技术的基本原理;第6～9章构成本书的应用篇,论述了在音字转换、自动文摘、信息检索、手写体识别等应用领域中的实用语言处理方法。

参加本书编写工作的有王晓龙、关毅(第1章、第2章、第3章、第8章)、刘秉权(第6章)、林磊(第9章)、陈清才(第2章、第7章)、刘远超(第7章)、赵岩(第3章、第5章)、赵健(第4章、第8章)、肖镜辉(第4章),全书由王晓龙、林磊等进行了统编和审校。由于编著者水平有限,错误和疏漏在所难免,敬请读者批评指正。

作　者

2004年3月

目录

CONTENTS

第1章

CHAPTER 1

引 言

语言是音义结合的词汇和语法的体系，是人类最重要的交际、思维和传递信息的工具。语言随着社会的产生而产生，随着社会的发展而发展，经历了漫长而缓慢的发展过程，成为一种极其复杂的、特殊的、充满了灵活性和不确定性的社会现象。在人类逐步进入信息化社会的今天，语言文字信息的计算机自动处理水平和处理量已成为衡量一个国家是否步入信息社会的重要标准之一。我国政府一直将中文语言处理技术这一学科作为高技术产业化的重点领域。从20世纪80年代开始至今，中文语言处理技术在字处理、词处理等领域均取得了重要的进展，获得了一系列实用化的成果，不仅使中文这一世界最古老的文字之一顺利地搭上了信息时代的火车，而且在文字识别、语音识别、机器翻译等语言处理技术方面与西文相比毫不逊色，在排版印刷等应用方面达到了世界领先的水平。

所谓计算机自然语言处理，是用计算机通过可计算的方法对自然语言的各级语言单位(字、词、语句、篇章等等)所进行的转换、传输、存储、分析等加工处理。本书主要讨论中文语言处理的基本理论和实用方法。这里所说的中文，广义上是指汉语和我国少数民族的语言文字，狭义上是特指汉语文字，包括以文本、图像、声音等形式存在的汉语口语和书面语。本书采用狭义的定义。中文语言处理通常是指以计算机为工具，采用可计算的方法对中文信息所进行的自动加工处理。从技术路线上，可以分为基于统计的语言处理技术和基于语言学规则的语言处理技术两大类。前者从大规模真实语料库中获得各级语言单位上的统计信息，并依据较低级语言单位上的统计信息，用相关的统计推理技术计算较高级语言单位上的统计信息；后者通过对语言学知识的形式化，形式化规则的算法化，以及算法实现等步骤将语言学知识转化为计算机可以处理的形式。按语言处理技术处理对象的不同，语言处理技术可以划分为字处理技术、词处理技术、语句处理技术、篇章处理技术等。按照语言处理技术的应用领域，语言处理技术可以划分为应用基础技术、应用技术两大类。本书主要介绍在词、语句、篇章等语言结构单位上引入语言学规则的统计语言处理方法的基本原理和应用。

自然语言处理技术是一门与语言学、计算机科学、数学、心理学、信息论、声学相联系的交叉性学科，与自然科学和社会科学的许多主要学科都有千丝万缕的联系，其中，又与语言学、计算机科学和数学的关系最为密切。在更加细微的层面上，与自然语言处理技术密切相关的学科有计算语言学、智能化人机接口、自然语言理解，等等。其中，计算语言学是现代语言学的一大分支，它是用计算机理解、生成和处理自然语言，即它的研究范围不仅涵盖语言信息的处理，还包括语言信息的理解和生成。智能化人机接口侧重于语言信息处理的应用研究，即运用语言处理技术改善人机交互的方式、手段和途径。自然语言理解则是人工智能的一个分支，其研究重点侧重于对经过深度加工处理的语言信息的理解，相当于语言处理技术在较高级语言单位上的应用基础研究。

从20世纪50年代初俄汉机器翻译系统诞生算起，中文语言处理技术的发展历史已经有50多年了。从20世纪80年代初期开始，随着计算机技术的普及和发展，中文语言处理技术在应用需求的推动下进入了一个快速发展的时期，在字处理、词处理、句处理、篇章处理等技术方面取得了一系列基础研究和应用研究的标志性成果。

在字处理技术方面，提出了信息处理用的汉字机内码。汉字机内码定义了汉字在计算机内部的存储方式，目前有中国国家标准GB2312—80；港澳台地区普遍使用的台湾《通用汉字标准交换码》，其地区标准号为CNS11643；国家信息产业部和质量技术监督局发布的《信息技术和信息交换用汉字编码字符集、基本集的扩充》，其国家标准号为GB18030—2000；中日韩三国根据UCS标准共同制定的《CJK统一汉字编码字符集》，其国际标准号为ISO/IEC10646，我国的国家标准号为GB13000—90。

汉字的输入码（或称汉字外码）提供了汉字输入的途径。目前主要的汉字输入码有五笔字型、拼音码等。通过建立输入码与汉字机内码的一一对应关系，使汉字输入到计算机中。

汉字字形库（或称汉字字形码、汉字发生器编码）存放如汉字的宋体、黑体、楷体等各种字体的点阵或曲线矢量字形信息，通过专门的处理程序把要输出的汉字转换成对应的汉字字形后在显示器、打印机上输出。汉字库中还包含了汉字放大、缩小、斜体、粗体等字体变化的信息。

汉字处理的应用技术主要包括汉字排版（如北大方正的激光照排系统）、印刷体汉字识别和联机手写汉字识别技术，对这些技术目前已经出台了一些标准，例如中国电子技术标准化研究所、北京汉王科技有限公司等单位起草的《联机手写汉字识别技术要求与测试规程》。

在词处理技术方面，词是自然语言中最小的有意义的构成单位，是自然语言处理中最基本的研究对象，也是其他研究的先行和基础。汉语不同于印欧语，它不把空格作为词的分隔标志，所以制定分词规范非常必要。目前有国家制定的《信息处理用现代汉语分词规范即自动分词方法》[1]（中华人民共和国国家标准GB13715），其主导的判别标准就是“使用频繁，结合紧密”。当然，在实际应用时，还要考虑文化、语感等多种因素才能得到正确的分词结果。对此，山西大学的刘开瑛[2]等、北京语言文化大学的宋柔[3]等学者进行了长期的研究，取得了很多研究成果。

词处理主要包括分词、词性标注、词义消歧三项内容。常用的分词方法包括正向最大

匹配、反向最大匹配以及基于词网格的统计方法。分词的主要难点在于歧义消解和新词识别。由于汉语本身的复杂性，目前这两个问题并没有得到根本性的解决。清华大学孙茂松、黄昌宁[4,5]，东北大学姚天顺等学者对此均进行了深入研究，取得了长足的进展。词性标注常用的方法是基于隐马尔科夫模型的词性标注方法。常用的词性标注方法包括基于词典知识库的方法，还有一些常用的基于统计的分类方法，包括贝叶斯方法和最大熵模型。分词和词性标注是所有中文语言处理应用技术的基础，广泛地应用于机器翻译、信息检索等各个领域。由中国科学院刘群、白硕等开发的ICTCLAS分词和词性标注系统，以及由哈尔滨工业大学自然语言处理研究室开发的ICSU分词和词性标注系统均取得了较高的分词和词性标注精度。

在语句处理方面，汉语语句处理技术是近年来中文信息处理研究领域的一个热点，也是一个难点。它是建立在汉字编码、汉语词语切分、汉语词法分析等基础上的一项技术，同时也是汉语篇章理解的基础。汉语的语句级处理技术主要包括句法分析、语句的语义分析等研究内容。吴蔚天[6]等在设计汉英机器翻译系统Sino Trans时提出的汉语完全语法树模型，虽然在一定程度上推动了用计算机分析汉语句法结构研究的发展，但在直接发掘汉语语言知识，揭示汉语的语言成分组合规律方面并没有多大进展。由于汉语语句处理是以语句作为研究对象，因此无论是在汉语语法研究还是汉语计算的数学模型上都存在相当大的难度，目前主要针对汉语短语展开研究，采用的方法主要是基于语法规则的方法、基于数学统计的方法，以及规则与统计相结合的方法。《现代汉语语法信息词典详解》[7]以朱德熙[8]先生提出的词组本位语法体系作为设置各项语法范畴的理论基础，马希文[9]的《从计算语言学角度看语法研究》，冯志伟[10]的《计算语言学对理论语言学的挑战》和他提出的潜在歧义结构论，白硕[11]的《语言学知识的计算机辅助发现》，罗振声[12]等对汉语句型的自动分析和分布统计的研究，以及北京大学计算语言学研究所的俞士汶、詹卫东[13]在基于规则的汉语短语分析方面所做的研究工作，都代表了基于规则的汉语句法分析的主流。在汉语语句语义的理解方面，值得关注的是国内学者黄曾阳[14]积多年研究心得，提出面向整个自然语言理解的理论框架——概念层次网络理论(HNC)，对传统的基于句法知识的语言表述及处理模式提出了挑战，代之以语义表达为基础，以对汉语进行理解。近年来，随着国外对统计语言学研究的兴起，国内的学者也针对汉语的统计语言模型展开了大量的研究。清华大学的黄昌宁[15]，哈尔滨工业大学的王晓龙[16]，赵铁军[17]，微软亚洲研究院的自然语言理解小组，中国科学院计算研究所的白硕、刘群[18]等学者和机构对汉语的统计语言模型进行了深入的研究和探讨，对汉语的*N*-gram模型、HMM模型进行了完善，并应用到了音字转换、机器翻译、句法分析等研究方面，取得了丰硕的成果。随着对汉语语句分析技术研究的不断深入，一些系统也随之投入到了现实应用当中，其中最引人瞩目的就是哈尔滨工业大学王晓龙等学者与微软公司联合开发的微软拼音输入法(MSPY)，这种语句级的输入法较以前的单词级的输入法在性能上有了很大的提高。此外，还有黑马中文自动校对系统、安徽中科大讯飞信息科技有限公司开发的TTS(text to speech)系统、Word中的中文自动更正功能等，都采用了汉语语句或者短语处理技术。

在篇章处理技术方面，当前对于篇章级的分析主要集中于研究一篇文章的话语结构，即进行话语结构分析。话语结构分析主要是指跨越语句本身的多个语句、段落之间在结

构或语义上的相互关系的分析。话语结构分析通常包括两种类型：一种是基于语法结构的衔接性(cohesion)分析，另一种是基于语义之间的连贯性分析(coherence)。前者主要是指在结构与形式上的衔接，而后者则指语义上的连贯性。前者讨论的范围比较趋向于文本呈现出来的表面结构如何彼此串联，例如，文本中间是如何运用适当的连接词或副词来串联句子，或者在文法层次上，句子和句子之间是如何依赖同样的主题词以及类似的句法结构来串联彼此；后者趋向于语义层面上更抽象的一致性，即文本实体之间是否基于相同的主题进行讨论。在实际的自然语言处理中很难将它们区分开，因为判断语法层的连接需要通过人脑中的认知模型来完成，而对人脑中认知模型的判断又是以文本中的语法结构为基础的。也就是说，人们在判断文本实体间的语义连贯性的时候往往会参照实体之间的语法衔接关系，而作者在撰写文章的时候也往往会借助形式上的语法衔接关系来反映语义上的连贯性，所以很难将这两者区分开来。

作者在撰写文章的时候，总会赋予一篇文章一定的语法或者语义结构。通过对这种结构的分析，可以帮助我们更好地理解全文的内容，从而更准确地完成相应的自然语言处理任务。当前话语结构分析主要应用于自动文摘领域，如哈尔滨工业大学的王开铸[19]等开发的 HIT—863 系统、王晓龙等开发的 InsunAbs[7] 系统、上海交通大学的王永成等开发的"OA"系统，都不同程度地进行了文章的话语结构分析。此外，在清华大学的罗振声、北京邮电大学的钟义信等的自动文摘研究中，话语结构分析也是一个重要的研究内容。除了应用于文摘系统，话语结构分析也可应用于信息抽取研究，如东北大学的姚天顺[20]等关于信息抽取研究中所进行的语段层分析。

经过 20 余年的艰苦奋斗，在语言学家、计算机专家的共同努力下，中文语言处理技术从无到有，取得了非常丰硕的成果。当前，在基础研究方面，中文语言处理技术正在跨越汉语自动语法分析的难关。在应用研究方面，基于国际互联网的语言处理技术，如文本分类、信息提取、自动问答、基于内容的信息检索等正在成为新的研究热点。但是，在国际互联网技术飞速发展，中文网页信息急剧膨胀，以及中文已经成为仅次于英语的世界第二大网页信息语种的今天，中文语言处理的发展速度仍然缓慢，特别是随着各项基础研究和应用研究的纵深发展，许多长期积累的困难和问题日益突出，已经成为中文语言处理技术继续发展的障碍。这些困难和问题主要表现在：长期以来，汉语语言学研究基本上是面向汉语教学的，能够直接面向计算机的形式化语言学的研究成果数量较少，而且缺乏统一的标准；中文语言处理研究力量分散，存在着低水平重复的现象；科学公正的评测机制尚未建立起来。中文语言处理技术的发展期待着语言学家和计算机专家打破门户之见，协同一致，兼收并蓄，早日实现资源和成果共享，为进一步提高我国的语言处理水平而共同奋斗。

本书作者长期从事中文语言处理的研究工作，20 多年来，在课题组全体成员的共同努力下，在语句输入、语料库加工、自动文摘、信息检索等领域做了一些探索性的工作，本书就是对这些工作的总结。我们希望通过本书，使读者在掌握中文语言处理技术的基本原理和主要方法的同时，认识到研究这门科学的高度复杂性和艰巨性，并且希望越来越多的人加入到这一艰苦、繁琐而又充满情趣的科学研究队伍中来。

参考文献

1 刘源等. 信息处理用现代汉语分词规范即自动分词方法. 北京:清华大学出版社,广西:广西科学技术出版社,1994
2 刘开瑛. 现代汉语自动分词系统中几个问题的讨论. 计算机开发与应用,1998
3 宋柔. 关于分词规范的探讨. 语言文字应用,1997(3)
4 孙茂松,黄昌宁,邹嘉彦等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展,1997,34(5):332～339
5 黄昌宁. 中文信息处理中的分词问题. 语言文字应用,1997(1):71～78
6 吴蔚天,罗建林. 汉语计算语言学—— 汉语形式语法和形式分析. 北京:电子工业出版社,1994:83,155～164
7 俞士汶. 现代汉语语法信息词典详解. 北京:清华大学出版社,1996
8 朱德熙. 语法问答. 北京:商务印书馆,1993
9 马希文. 从计算语言学角度看语法研究. 国外语言学,1989(3)
10 冯志伟. 计算语言学对理论语言学的挑战. 语言文字应用,1992(1)
11 白硕. 语言学知识的计算机辅助发现. 北京:科学出版社,1995
12 罗振声,郑碧霞. 汉语句型自动分析和分布统计算法与策略的研究. 中文信息学报,1994(2)
13 詹卫东. 面向中文信息处理的现代汉语短语结构规则研究. 博士论文,1999
14 黄曾阳. HNC(概念层次网络)理论. 北京:清华大学出版社,1998
15 黄昌宁. 关于处理大规模真实文本的谈话. 语言文字应用,1993(2)
16 关毅,王晓龙,张凯. 基于统计与规则相结合的汉语计算语言模型及其在语音识别中的应用. 高技术通讯,1998,8(4):16～20
17 赵铁军等. 机器翻译原理. 哈尔滨:哈尔滨工业大学出版社,2000
18 刘群,詹卫东,常宝宝等. 一个汉英机器翻译系统的计算模型与语言模型. 见:吴泉源,钱跃良主编. 智能计算机接口与应用进展. 北京:电子工业出版社,1997,253～258
19 刘挺,王开铸. 基于篇章多级依存结构的自动文摘研究. 计算机研究与发展,1999,36(4)
20 姚天顺. 自然语言理解——一种让机器懂得人类语言的研究. 北京:清华大学出版社,2002