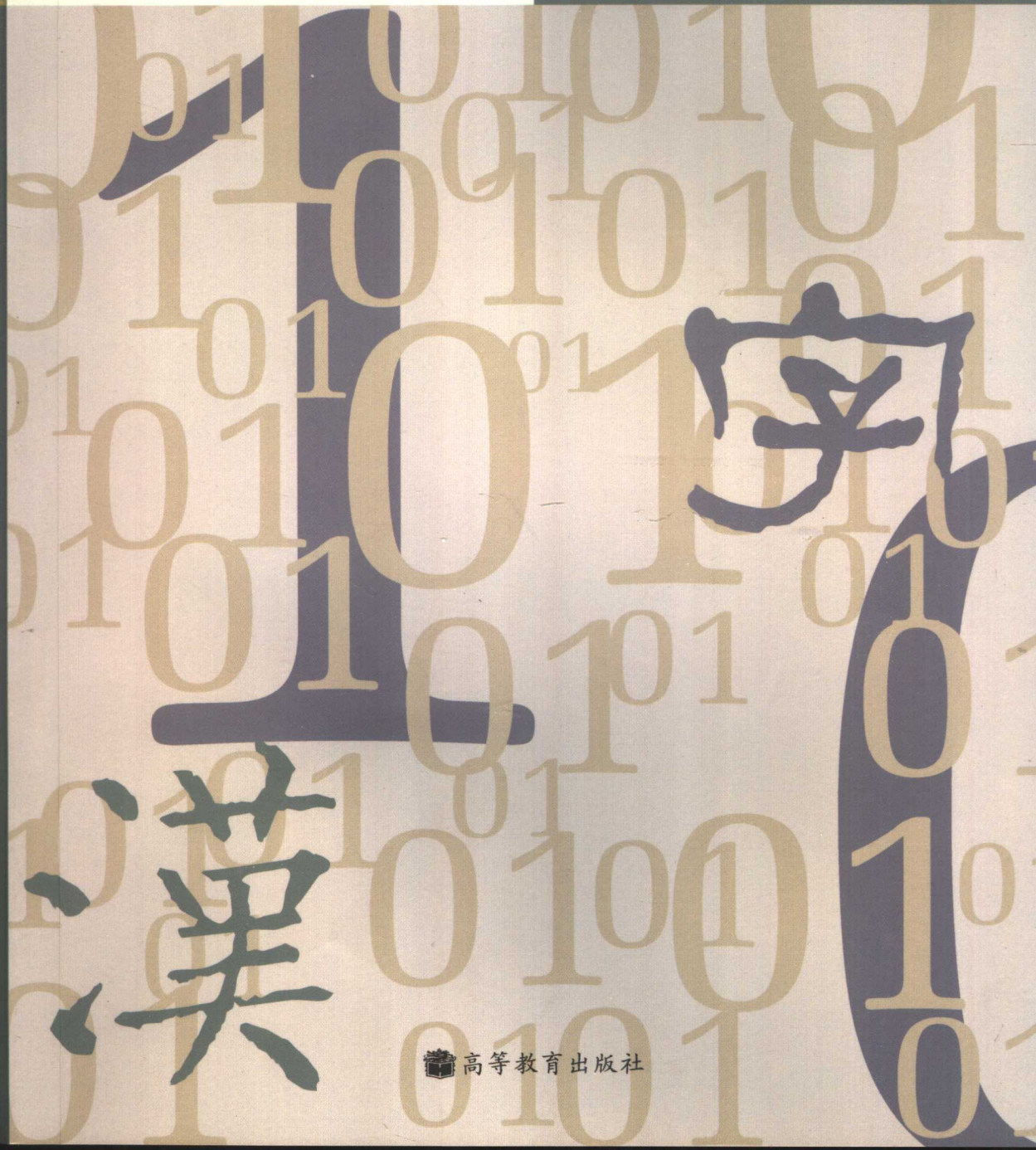
自然语言处理

◇江铭虎　主编

Natural Language Processing



高等教育出版社

自然语言处理

Natural Language Processing

◇江铭虎　主编

高等教育出版社

内容简介

自然语言处理是运用计算机对自然语言进行分析和理解，从而使计算机在某种程度上具有人的语言能力。本书重点介绍了自然语言处理的基本问题、相关方法和重要领域，包括汉语句型分析与分布统计、语料库处理、文本自动分类与检索、文本自动文摘、中文文本自动校对、人机交互技术、汉语盲文翻译和甲骨文信息处理等。本书既有数学理论模型，又有实验论证，从理论到实践，深入浅出，结构合理，概念阐述明确，公式推导简明，易于理解，便于教学。本书可作为中文信息处理专业和计算语言学专业的高年级本科生、研究生的教材或参考书，也可供自然语言处理或计算机信息处理和人工智能领域的相关人员参考。

图书在版编目(CIP)数据

自然语言处理/江铭虎主编. —北京：高等教育出版社，2006.12
ISBN 7-04-010214-7

Ⅰ.自… Ⅱ.江… Ⅲ.自然语言处理-高等学校-教材
Ⅳ.TP391

中国版本图书馆 CIP 数据核字(2004)第 134891 号

策划编辑 李海凤 **责任编辑** 李海凤 **封面设计** 张申申 **责任印制** 韩 刚

出版发行	高等教育出版社	**购书热线**	010-58581118
社 址	北京市西城区德外大街 4 号	**免费咨询**	800-810-0598
邮政编码	100011	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京中科印刷有限公司	**畅想教育**	http://www.widedu.com
开 本	787×960 1/16	**版 次**	2006 年 12 月第 1 版
印 张	27.25	**印 次**	2006 年 12 月第 1 次印刷
字 数	520 000	**定 价**	41.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

物料号 10214-00

序　言

自然语言处理旨在研究利用计算机理解和生成自然语言，它是一门新兴的交叉学科。20世纪中叶，随着计算机的诞生，自然语言处理的研究应运而生。到了20世纪末，由于计算机网络的出现以及与之相应的信息技术的快速发展和广泛应用，人们对用机器处理自然语言的需求越来越迫切，因此自然语言处理在20世纪下半叶成为人们研究和关注的热点，并取得了显著的进展。

这部书较全面地总结了这个时期以来国内外有关（特别是汉语自然语言处理）的研究工作。它的内容包括以下几个方面：第一，自然语言处理的基本问题，包括汉语自动分词、汉语文本自动标注、句法分析、语料库处理等；第二，自然语言处理的相关方法，包括统计学方法、人工神经网络方法、粗集理论方法等；第三，自然语言处理的重要领域，包括文本自动分类与检索、文本自动文摘、中文文本的自动校对、人机交互技术等；第四，两个应用实例，即汉语盲文翻译与甲骨文信息处理。

本书的作者长期从事自然语言处理的教学与科研工作，积累了丰富的知识与经验。书中大部分的内容取自他们的研究成果，也包括了他们对自然语言处理的理解与体会。本书具有以下特点：内容先进，较好地反映了当前国内外的研究现状，特别是汉语处理的研究现状；自然语言处理的基本问题和方法多数是在应用背景下讲述的，特别是第4～7章关于自然语言处理具体领域内容的介绍，其中相当大一部分是作者与研究生们一起从事科学研究工作的成果，因此分析深入，具有较强的实用性。这本书是在清华大学研究生讲义的基础上撰写的，经过了教学的检验，并经作者反复补充与修改写成，从而保证了图书的质量。

由于自然语言本身的复杂性，实现机器对自然语言的理解与自动生成是一项十分艰巨的任务，需要自然科学与人文科学中各种不同学科领域的交叉与结合。目前已经取得的成就离最终的目标还有相当的距离，这种现状正是自然语言处理研究的魅力所在，相信经过不同领域研究人员的共同努力，自然语言处理研究今后一定会取得长足的进步。

这本书可以作为相关学科大学生与研究生的教材，也可以作为希望了解、学习和从事自然语言处理研究工作的研究人员的参考读物。

清华大学计算机系　张　钹

2005年12月20日

前　言

自然语言处理是一项十分庞大而繁复的工程，它是自然科学和社会科学交叉的学科，特别是计算机科学、语言学、逻辑学、心理学、信息科学的交叉学科。自然语言处理的目标是实现计算机对文字信息的自动分析和理解，它立足于实验、理论和计算三大支柱，通过对人脑和语言认知的实现途径进行模拟研究，建立起多层次网络处理模型来阐明人脑语言信息处理系统，以期取得突破性进展。这一研究具有很强的生命力，是当代科学新的生长点，不仅对信息科学、认知语言学和心理学的发展有推动作用，也对国民经济和社会的发展有推动作用。从工程的角度讲，自然语言处理就是要设计出一个能理解和生成自然语言的计算机系统，研究重点在相应各种算法的设计上。开展语言信息处理的研究，可以带动多种学科和技术的发展，尤其能为智能科学的突破性进展贡献力量。

本书所述内容覆盖了自然语言处理的主要范围。第1章论述自然语言处理的意义、历史与现状以及自然语言处理的方法、特点和规律。第2章围绕语言分析的几个层次，论述汉语自动分词、自动标注、句法语义分析和语料库处理等自然语言处理中一些基本问题。第3章阐明语言统计和语言分析的关系，介绍系统设计的思想和技术。第4章介绍文本自动分类与检索的有关理论和方法，论述粗集和模糊理论在分类与检索中的设计与实现。第5章介绍自动文摘的理论和方法，论述自动文摘中篇章结构的语法、语义分析以及基于词语和概念统计的自动文摘的设计与实现。第6章介绍中文文本自动校对的设计思想，使用统计模型、歧义消解、模糊匹配等方法，提出了基于实例、统计和规则的综合语义校对策略。第7章论述人机对话中的语音识别与合成、自然语言生成和对话系统的句法语义分析的理论、方法和系统实现。第8章论述汉语盲文翻译和甲骨文信息处理的理论和方法。其中，第3章的汉语句型分析与分布统计、第4章4.6节的模糊分类系统设计的基本思想、第5章的文本自动文摘和第6章的中文文本自动校对等内容，是清华大学计算语言学实验室罗振声教授和他的研究生郑碧霞（第3章初稿作者）、李蕊（第4章4.6节初稿作者）、郭玉箐（第5章初稿作者）、季姮（第5章初稿作者）、骆卫华（第6章初稿作者）、龚小谨（第6章初稿作者）的科研工作成果，经罗教授同意将其编入本书。罗教授多年来对我的关心、帮助和支持，使本书的编写工作得以顺利完成。在此，对罗振声教授及其研究生表示衷心的感谢。本书中的第1章、第2

章、第7章由我和研究生王彬完成，第4章的4.1～4.5节由我和研究生盛晓炜完成，第8章8.1节由我和本科生谭刚完成，第8章8.2节由我和研究生王俊珍、蔡慧颖完成，书中的大部分内容是清华大学研究生多年使用的教学讲义，大多是清华大学计算语言学实验室多年来科研工作的总结，既有数学理论模型，又有实验论证。全书由我进行编辑整理并统稿。

清华大学计算机系的张钹教授（中国科学院院士）、北京语言大学计算机系的宋柔教授和清华大学智能技术与系统国家重点实验室的苑春法教授对本书的初稿提出了很诚恳的建议和意见，张钹院士还为本书作序，我在此表示衷心的感谢。书中各章末列举了主要的参考文献，在此对所引参考文献中的作者和出版机构也表示感谢。

本书的相关工作得到了国家自然科学基金、清华大学985基础研究基金、教育部优秀青年教师资助计划、清华大学研究生精品课建设、清华大学认知科学创新基地项目、中国科学院自动化研究所模式识别国家重点实验室开放基金的支持，本书的出版得到了高等教育出版社的大力支持和帮助，在此一并表示衷心的感谢。

本书体系完整，条理清楚，便于教学和自学，可以作为自然语言处理和人工智能领域研究生和高年级本科生的教学参考书。

由于时间限制，本书难免存在疏漏和不足之处，欢迎各位专家和读者批评指正。

江铭虎

2006年8月20日于清华园

目　录

第1章 概 论

1.1 自然语言处理研究的意义、历史与现状

1.1.1 自然语言处理研究的意义

自然语言处理(natural language processing,英文缩写为NLP),是研究如何利用计算机来理解和处理自然语言的,即把计算机作为语言研究的工具,在计算机技术的支持下对语言信息进行定量化的研究,又被称为自然语言理解(natural language understanding,英文缩写为NLU)或计算语言学(computational linguistics)。自然语言处理主要用于说明方法,侧重于工程;而计算语言学这个术语主要用于说明理论。从理论研究的角度看,语言科学是人文科学与自然科学之间的桥梁,而自然语言处理正是这两大科学相结合的一个有代表性的交叉学科。开展语言信息处理的研究,可以带动多种学科和技术的发展,尤其能够为智能科学的突破性进展贡献力量。

自然语言处理是一项十分庞大而繁复的工程,它是自然科学和社会科学交叉的学科,特别是计算机科学、语言学、逻辑学和心理学的交叉学科。自然语言处理的目的是实现计算机对语言信息的自动分析和理解,它立足于实验、理论和计算三大支柱,通过以对人脑及语言认知的实现途径进行模拟研究,建立起多层次网络处理模型来阐明人脑语言信息处理系统,以期取得突破性的进展。它的研究具有很强的生命力,是当代科学新的生长点,这不仅对信息科学,而且对认知语言学、心理学,以及对国民经济和社会的发展都会起到推动作用。自然语言处理的研究不可能一步就达到对大规模真实文本的完善处理,必须逐层逐步地加以分析和解决,各层次的研究既相互独立,又有着十分密切的联系,对每一层次的研究,都应考虑更高层次的研究需要。

由于自然语言处理的对象是人类自然形成的极其复杂的语言现象,所以这门学科极具艰巨性。事实上,这门学科自20世纪40年代产生以来,经历了十分曲折的发展历程;然而随着信息社会的到来,自然语言处理在机器翻译、信息检索、人机交互等信息处理领域有着广泛的应用前景,这是这门学科的实用价值所在。

1.1.2 自然语言处理研究的发展历程

1. 国外研究现状

自然语言处理是运用计算机对自然语言进行分析和理解，从而使计算机在某种程度上具有像人的语言处理能力。国外关于自然语言处理方面的研究起步较早，一些卓有成就的语言学家、逻辑学家和心理学家都在自然语言处理中的语法、句法及语义分析方面提出了一系列较为系统的理论方法。可以认为，自然语言处理的研究始于机器翻译。1954 年初，美国乔治敦大学在国际商用机器公司（IBM）的帮助下，在 IBM－701 上进行了第一次机器翻译实验。此后，机器翻译成了自然语言处理的重要研究课题。

到了 20 世纪 60 年代，随着一些新的人工智能方法的提出和 Chomsky 等人在语言理论上的突破，人工智能学者开发了一批新的语言处理系统。这些早期的自然语言处理没有成熟的语言句法分析，采用的主要技术是模式识别中的句法匹配，而且只能达到英语的受限领域的有限目标。

在 20 世纪 70 年代，出现了一些有名的自然语言处理系统。如 W. Woods 在 1972 年设计了自然语言信息检索系统 LUNAR，并在此系统中提出了著名的扩充转移网络（augmented transition network，英文缩写为 ATN）。SHRDLU 是 T. Winogard 于 1972 年在美国麻省理工学院的人工智能实验室开发出来的一个自然语言理解系统，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所需的知识）、一个问题求解器（可以为执行命令和寻找问题答案作出安排），是一个句法、语义和推理的组合系统。MARGI 是由 R. Schank 于 1975 年在斯坦福大学人工智能实验室建立起来的一个系统，其目的是提供一个自然语言理解的模型，该系统是根据 Schank 早年提出的概念从属理论建立的，系统由概念分析器、推理器和篇章生成 3 个模块组成。

20 世纪 80 年代，各种新的语法体系应运而生。如 Gazder 的广义短语结构语法（generalied phrase structure grammar）、Bresnan 与 Kaplan 的词汇功能语法（lexical functional grammar）、M. Kay 的功能合一语法（functional unification grammar）等。由于新的语法体系运用了复杂特征集与功能合一技术，使得自然语言处理能力较以前采用单一标记的处理大大增强。

20 世纪 90 年代，在国际上掀起了语料库语言学（corpus linguistics）的研究热潮。语料库语言学研究机读（自然语言）文本的采集、存储、检索、统计、语法标注、句法和语义分析，以及具有上述功能的语料库在语言定量分析、词典编撰、作品风

格分析和机器翻译等领域的应用，为自然语言处理的研究提供了新思路。

在语料库建设方面，英美各国相继开发了超大型语料库，如来自15个英语国家的跨国英语语料库（每个库带语法标注的词有100万条）、1亿词级的英国国家语料库、2亿词级的Birmingham英语语料库、1亿词级的美国计算语言学语料库、赫尔辛基的千万词级历史英语语料库，以及上亿词级的法语语料库。在加工的深度上，也从词性标注扩展到句法和语义的标注。如：英国兰开斯特大学3000万词的Longman/Lancaster英语口语语料库，有语法和句法标注和节奏韵律标注，该学校的Lancaster Treebank项目，自1986年陆续开发了功能不同的词性自动标注工具；美国宾夕法尼亚大学的Penn TreeBank是一个经过句法标注的语料库，通过吸收和改造一些现有的词性标注工具和句法分析器，形成了一个完整的语料库加工处理系统。由于该系统具有功能强大、操作简单的语料校对工具，从而大大提高了人工校对的效率。近年来不断有超大规模经过加工的熟语料库问世。我国也从最初的收集生语料到现在建有一定规模的熟语料库，并由此发展了汉语的自动分词技术、词性标注技术和句法分析技术，这些技术的突破也有赖于加工更加精细的汉语语料库。未来语料库的规模将是百亿词级，语料库的应用前景也将越来越广阔。

在语言知识库的构建方面，美国普林斯顿大学智能科学实验室开发的英语词汇语义数据库WordNet于1990年在因特网上公布之后引起了广泛的关注，许多研究人员用WordNet来进行英语语料库的语义标注和词义辨识。继WordNet之后，美国FrameNet（框架网络）工程启动，该工程由美国国家科学基金支持，由加州国际计算机科学研究所和加州大学伯克利分校语言学系联合开发。在欧洲，荷兰的阿姆斯特丹大学于1996年开始开发荷兰语、意大利语、西班牙语的多语种词汇、语义数据库EuroWordNet。这些工程的开发为语义知识体系的构建进行了有益的探索。

2. 国内研究现状

与国外的研究相比，我国自然语言处理研究所面临的一个重要难题就是如何结合汉语自身的特点选取有效的形式理论和研究方法对汉语进行分析处理。

汉语的种种特点使我们的自然语言理解无法直接套用西方现有的语法和语义结构体系，这使得汉语自然语言理解研究工作困难重重。令人欣慰的是近几年国内自然语言处理的研究取得了很大的成绩，无论是汉语书面语的自动分词、汉语电子词典、汉语机读语料库、机器翻译、汉语人机交互、汉语文献检索等应用研究，还是结合汉语、汉字特点探索自然语言处理基础理论的研究，都取得了可喜的成果。

对自然语言处理的理论研究是从以词形分析为主的早期阶段以及注重语义分析的中期阶段发展到了基于语料库统计方法的近期阶段。国内众多学者都为此作

出了孜孜不倦的探索和努力。就理论发展来说，国内自然语言处理的研究借鉴了国外的各种理论，提出了一系列符合汉语特点的语言分析方法和语言表示理论。早期的系统大都是基于转换生成语法和扩充转移网络，在语义分析方面大多是采用汉语格语法理论，并专门研究了汉语的各种信息在语义网络中的表示方法。近年来中国科学院声学研究所黄曾阳先生提出了概念层次网络理论（hierarchical network of concept），它是面向自然语言理解的理论框架，以语义表达为基础，并以一种概念化、层次化和网络化的形式来实现对知识的表达。

随着计算语言学研究的深入和汉语自然语言信息处理应用系统的开发，学术界开始感觉到建设语言知识基础工程的迫切性。从20世纪80年代中后期以来，学术界投入了许多力量来进行这方面的建设工作。北京大学的《现代汉语语法信息词典》就是这方面的突出成果，它直接建立在词组本位语法体系的框架上，实现了对5万多汉语通用词语的句法属性特征的全面系统的描述，是第一部大规模的以复杂特征集的形式化方式表述汉语语法的机用词典。此外，还有《信息处理用汉语语义词典》、《现代汉语述语动词机器词典》等机用词典。这两部语义词典在语义场、格语法等语义学理论的框架下对汉语的实词（特别是动词）进行具体的语义属性刻画；清华大学的《汉语语素数据库》近几年得以建成和应用；董振东先生的知网（HowNet）于1999年在因特网上公布，这是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识性知识库。这些词典为汉语自然语言应用系统知识库的开发提供了一个基础，它们在目前的一些应用系统中发挥着实际的效用。

在语料库语言学方面，近年来国内对汉语语料库加工从单级处理发展到多级处理的研究，并在自动分词、词性标注和依存关系标注等方面取得了可喜的成果；开发和积累了许多有用的处理工具，逐渐形成了一个较完整的汉语语料库多级加工处理系统；建立了数个有一定规模的汉语语料库，在汉语语料的标注和利用带标记的语料来自动获取语言知识等领域取得了不少的成果。如：中国社会科学院文学研究所拥有千万词级汉语语料库；中国台湾地区“中央研究院”也有近千万词级的古代、近代、现代汉语语料库；清华大学拥有5000万字的ZW大型通用汉语语料库，利用英汉双语平行语料进行自动机器翻译的研究，并用来对现代汉语述语动词电子词典进行研究；北京交通大学信息科学研究所对新华社《人民日报》、《计算机世界报》的真实语料进行加工处理，并用于语音识别、音字转换和语言理解。

概括说来，目前国内自然语言处理的研究课题主要有：自然语言的句法和语义分析、语料库建设和语料加工技术、基于语料库的语言分析方法、机器翻译系统及其评测方法、文本分析与生成、机用词典、文本检索、自动文摘、文本校对、文字识

别、智能型汉字输入方法和人机交互接口等。我国的自然语言处理工作者们吸取借鉴国外同行的理论方法和最新研究成果，并结合汉语的特点开展和加强基础理论的研究，从信息处理的角度进行汉语的研究，向自然语言处理的本土化发展。自然语言处理研究将向语法、语义、语用和语境各方面的综合研究方向迈进。

1.2 自然语言处理研究的方法、特点和规律

1.2.1 理性主义与经验主义

自然语言处理研究的方法有很多种，但从方法论上可将其归入理性主义与经验主义两大研究方法。理性主义研究方法通常根据一套规则或程序，将自然语言理解为符号结构——该结构的意义可以从结构中符号的意义推导出来；而经验主义的研究方法主要是统计学的方法与神经网络学习方法。

1. 理性主义方法

理性主义方法认为人的很大一部分语言知识是与生俱来的，即是由遗传决定的。受 Chomsky 内在语言官能(innate language faculty)学说的影响，自然语言处理学界很多人信奉理性主义。在一个典型的自然语言处理系统中，句法分析器按照人所设定的自然语言语法把输入句分析为句法结构，再根据一套语义规则把语法符号结构映射到语义符号结构。自然语言处理系统中的规则集通常是先验的，即是由人设计好以后赋予机器的，因此这是一种典型的理性主义的方法。

2. 经验主义方法

经验主义方法认为人的知识只有通过感官传入，再通过一些简单的联想和概括操作才能获得，人不可能天生拥有一套有关语言的原则和处理方法。表现在自然语言处理中，许多研究尝试从大量的语言数据中获取语言的结构知识，从而开辟了基于语料库语言学这种经验主义的研究方法。统计学方法试图建立统计性的语言处理模型，并由语料库中的训练数据来估计统计模型中的参数。神经网络方法继承了人工智能中的连结主义传统，从给定训练数据之间的输入－输出关系，由机器通过学习来获得神经元之间的连结强度，以反映输入与输出状态之间的映射关系。简而言之，理性主义强调基于规则的方法，经验主义强调基于学习的方法。

语料库语言学是经验主义研究方法的代表。它针对规则模型的困境，以语料库和统计模型为基础，从语料库中存储的大规模真实文本中直接获取多种语言知识。由于采用此法获得的语言知识覆盖面宽，因此能大大改进语言处理的质量。

语料库是指由大量真实自然语言文本组成的集合，它集中了语言的具体应用的大量实例，通过利用一定的自动学习算法，主要是概率统计技术，从中获取带有相应概率的语言使用规则。语料库并不是大量文本的简单堆积，要从语料库中真正获取语言知识，就必须对库存语料进行词法、句法、语义等各个层次的加工，语料库的结构决定了整个系统知识获取的能力和效率。目前对文本语料库的研究已逐渐形成完整的理论体系，称为语料库语言学。它的崛起，使从大规模真实文本中发现并总结自然语言的各种语言事实和语法规律成为可能。语料库作为一种知识源，提供了大量的语言事实，如词的用法、词与词之间的搭配等，它是一种更真实、更客观的语言使用资料。从语料库中获取的知识主要有：①词频统计数据，包括单个词的出现频率和邻接词的同现频率；②语料经过词性标注后某一标注串出现的频率；③词与词之间的联结关系、依存关系等。

近几年来，全球范围内的自然语言处理学界兴起了对大规模语料库的研究兴趣。这主要是因为计算机产业和信息处理技术的迅速发展，计算机的存储能力和运算速度大大提高，使得在计算机中存储大量的文本和对文本方便快速地扫描、检索成为可能；因特网上的电子文本数量与日俱增，可以比较容易地获得大量语料。另外语音识别领域在20世纪70年代开始逐渐采用概率模型替代原来的基于规则的识别手段，概率模型的参数是通过对大量语声语料进行统计训练得来的。概率模型的识别效果大大优于使用规则的方法，这给自然语言处理领域对文本语料的研究提供了有益的借鉴。研究文本语料库的目的是从语料中发现包含在其中的语言使用规律。从而利用这些规律对库内或库外文本进行句法语义分析。利用语料库研究语言规律包括：收集语料、建设语料库、对语料库进行多层次标注加工、从标注过的语料中获取多层次的语言学知识，以及语料库语言学的应用研究等。

1.2.2　汉语语言处理的方法

汉语是由汉字组成，汉字是一种象形文字（表意体系的文字），不同于印欧语言的表音体系的文字。其一，汉字不是由字母来表示音素或音节的，而是由不同笔画（通过点、曲线等方式）的组合构成汉语的单音节语素（表意符号）来描绘不同的事物和意象；其二，汉字不是直接表音的，这使得汉字在一定程度上具有超时空性。几千年来，汉字的语音变化很大，但字形从古至今是一脉相承的，汉字所记录的语素的意义却变化不大。对于一个生僻汉字或古文字，虽然我们可能不了解其读音，但文字的意义却能在大脑中反映出来，这一点在现代的西方人看来几乎是不可想像的。从信息处理的角度分析，汉语语言与西方语言相比有如下一些特点：汉语是大字符集的语言，形态上的区别特征少，没有词形变化，句型变化繁多，存在着大量

同形歧义和同音歧义现象，一词多义现象比西方语言更为严重。正是这些特点使得在进行汉语语言处理时不能直接套用西方语言的处理方法，有很多特殊的问题需要考虑。近年来通过采用先进的 fMRI 技术对印欧语系和汉语的名词、动词和兼类词的脑神经成像分析发现：印欧语系的词在形态上的区别特征多，有词形变化，这三类词脑神经成像在不同的区域；而这三类汉语词在脑神经成像的区域却是交叠的。这证明了汉语信息处理不能直接套用西方语言的处理方法。我们采用人工神经网络模拟技术，对这三类汉语词（共 68 个）进行了词性和句法语义标注，分别采用 50 维句法特征、132 维语义特征、29 维句法加 35 维语义的混合特征、8 维句法加 132 维语义的混合特征进行自组织特征映射分析，发现单独采用语义特征对这三类汉语词进行的自组织特征映射与脑神经成像的分布区域（相互交叠）相吻合。单独采用句法特征对这三类汉语词进行的自组织特征映射的分布区域是相互分开，这是由于句法的类别属性特征明显，而汉语的句法理论均是从印欧语系的句法理论移植过来的，一直沿用至今，如果我们采用印欧语系的句法理论进行汉语自然语言理解处理时，很多问题没法解决也没法解释。通过采用句法和语义的混合特征对这三类汉语词进行自组织特征映射，人们发现，当语义特征的维数远远多于句法特征的维数时，各类词的分布区域趋于相互交叠，当语义特征与句法特征的维数相差不多时，各类词的分布区域趋于相互分开。此项研究表明，对于母语是汉语的人来说，人们在日常交流中对汉语（或汉语的词）进行理解时，主要是从语句的意义或词的内容含义上进行理解，不太注重句法和语法因素。很多没有学过句法和语法理论的人，也可以毫无困难地进行交流，表达他们的思想。

正是由于人脑对汉语的学习和处理方式有别于西方语言，计算机对自然汉语的处理也是有别于对西方语言的处理，本书主要是针对汉语信息处理，因此本书的研究方法有其自身的特点，是国外同类教材无法替代的。

在对汉语的自然语言进行处理时，首先要了解汉语的特点。

汉语的认知理论、模型不同于印欧语，在分析方法上，后者在词汇、语法、语用、语境诸层面上有明显的特征区别，相互之间又有对应关系。但汉语则不同，各层面之间很难划分经纬，词法和句法之间没有明显的界限。汉语由于没有明显的自然形态界限可以作为分词依据，对词没有一致认可的定义。汉语缺乏严格意义上的形态变化（形态标记）。同一词作为不同的语法成分出现时，词的形态保持不变。汉语中词的语法功能较复杂，词性与相应的句中语法成分不能一一对应。汉语信息处理的关键问题在于名词和动词两类词的语法和语义。汉语短语作为语言的一个层次，占有十分重要的位置，与印欧语言不同，汉语句子的构造原则跟短语的构造原则基本一致，因而导致了许多句法上的歧义结构，如果仅用词性标记的同现信

息很难确定其具体结构，无疑给汉语的分析和处理增加了困难。如能把各类短语的结构和功能分析清楚了，那么句子的结构实际上也就分析清楚了。

具体到汉语的词和短语上，由于汉语中有意义的、可以自由运用的最小单位是词，因此它是汉语研究中非常重要的语言单位，也是计算机汉语信息处理中汉语语言理解、机器翻译等研究的基本单位。词是句法分析也是语义分析的最小单位。汉语短语是由两个或两个以上的词，按照一定的语法规则组成，表达一定意义的语言单位。短语比词大，可是又不成为一个句子。短语依据其语法结构并结合其语法功能，分为以下结构类型：附加、重叠、数量、同位、联合、偏正、主谓、述宾、述补、连动和兼语结构。名词短语是构成语言的重要组成部分，是传递信息不可缺少的基本单位。在自然语言处理领域，名词短语的正确分析对于机器翻译、文献检索以及句法分析等更具有重要意义。

总之，汉语的特点可归纳为：汉语词性缺少外部的形态标志，组词成句不靠形态变化，而是靠语序和虚词；语素组成词，词组成短语或句子，都用一套共同的规则；汉语动词没有“时”的变化；助词是汉语特有的一类词，并有丰富的量词。这些特点也使在进行汉语信息处理时要面临以下几个问题。

1. 汉语的歧义问题

自然语言分为形式和内容两部分，在语言形式上，表现为语法；在语言内容上表示为语义，若语言形式完全决定了语言的意义，则机器对语言的理解无二义，汉语是口语和书面语十分发达的自然语言，在汉语的生成中，难免会出现许多歧义现象。歧义是自然语言中的普遍现象，它是区别于人工语言的最大特点，对歧义问题的研究是任何一种自然语言处理体系都无法回避的问题。通常歧义分为语音歧义、词汇歧义、句法歧义和语境歧义等。句法歧义受到的关注最多，它是句子表达的意义与句法形式之间的矛盾，同一句法形式含有多个不同的意义时就产生了句法歧义，自然语言中这种类型的歧义最为常见，它不是单独出现的，而是同特定的句法结构有联系，大部分分词歧义可通过句法分析得到解决。词性的歧义着眼于同一词属于不同的语法范畴，词义歧义是一个词有多种词义，词性的歧义考虑的是词在语法上的多义问题，而词义歧义则主要研究词在语义方面的多义问题。排除歧义是一项非常艰难的工作，是一个多层次的处理过程，没有一种方法可以完全解决这个问题，在排歧过程中，需综合利用语法和语义知识，由于语法规则库的完备性受到限制，字典中语义属性的标注难免有不完善之处，此外歧义的处理还常常需要利用上下文信息，因而对排歧处理不能令人满意。目前还没有比较完善的系统化方法解决语境歧义。

2. 汉语语法兼类现象(词的同形异类现象)

同一形式的词具有两种或两种以上语法功能类别的现象称为兼类现象。如汉语中"连"这个词兼有副词、介词、动词、名词和量词五种词性。兼类词在词典词条中所占的比重虽不高,但出现的频率却很高,而且越是常用的词,其兼类现象越严重。

3. 分词

计算机进行汉语信息处理时,其核心是对词的处理,首先遇到的是词的切分问题。由于汉语句子中词与词之间无空格,必须把句中各词正确地切分出来,才能正确理解和处理汉语句子。因此分词是语料库加工的首要环节,它直接影响着后续处理的性能,特别是由于语料内容的广泛性,其中会出现大量的专有名词,如何正确识别和切分这些专有名词是对分词系统提出的一个研究课题。

在分词问题基本解决后,汉语理解研究又转向了语句的句法分析、语义分析和篇章理解等基础理论工作方面。

4. 词性标注

作为大规模语料库自动加工的关键步骤之一,词性标注的任务是在建立句法结构树之前首先应明确文本中所有语法兼类词在具体使用场合中所属的词性,通常利用不同词性在文本中出现的概率性知识和局部组合规则,在分词基础上进行加工,以利于其后的句法标注和语义标注。词性标注系统使用的标注集不局限于传统语法规定的词性范畴,还包括习用语、语素、标点等非词标记。明确兼类词在句中的词性是决定其语法功能、分析句法结构的先决条件,而词性又是在一定的句法环境中确定下来的。词性标注和句法标注两者形成循环嵌套关系。为打破这种嵌套,首先把词性标注问题限制在词法平面内处理,在不涉及句法结构的条件下消除大多数的兼类,再上升至高一级的句法平面解决其余的少数兼类现象。

在词法平面内现有的词性标注法有:基于规则的方法、基于统计的方法、基于神经网络的方法和规则与统计相结合的混合方法。

基于统计的词性标注法具有较好的鲁棒性,处理未知词和对开放语料的推广能力强。它的全部知识是通过对大规模语料库的参数训练自动得到的,因此可获得很好的一致性和较高的覆盖率,并可将一些不确定的知识客观地进行定量化。

规则法处理标注是利用上下文框架规则描述在特定的语境下一个兼类词到底应标注什么标记,这里的语境包括词语信息、词性信息、某些词语的特征信息。规则处理是根据特征词、词性组合和上下文关系消除兼类。规则对兼类现象的覆盖率和规则处理的正确率是相互制约的。一条规则覆盖的兼类现象越多,它处理的