普通高等教育“十一五”国家级规划教材

中文信息处理教程

Chinese Information Processing Tutorial

宋继华　杨尔弘　王强军　编著

高等教育出版社
HIGHER EDUCATION PRESS

普通高等教育“十一五”国家级规划教材

中文信息处理教程

Zhongwen Xinxi Chuli Jiaocheng

宋继华　杨尔弘　王强军　编著

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

本书是普通高等教育“十一五”国家级规划教材，是编者依据自身的教学实践，在学习、吸收、借鉴前辈经验的基础上归纳、提炼而成的中文信息处理教材。书中比较系统地介绍了本学科领域的基本原理、方法和应用技术。

全书共8章，按中文信息处理的语言单位层级——汉字、词语、句子、篇章依次展开。内容包括：导论、汉字的信息处理、词的信息处理、句子的信息处理、句子语义表达与分析、中文信息处理的基础资源、文本分析与处理、中文信息处理评测。

本书的特色是简明、实用，逻辑性强，可读性好，注重引导学生动手解决实际问题。每章都附有习题，并给出了大部分习题的参考答案。

本书可作为高等学校计算机、信息管理等专业本科生和研究生的教材，也可供从事中文信息处理研究和应用的科技工作者参考。

图书在版编目(CIP)数据

中文信息处理教程/宋继华，杨尔弘，王强军编著. —北京：高等教育出版社，2011.6

ISBN 978-7-04-031896-8

Ⅰ. ①中… Ⅱ. ①宋…②杨…③王… Ⅲ. ①汉字信息处理-高等学校-教材 Ⅳ. ①TP391.12

中国版本图书馆CIP数据核字(2011)第113063号

策划编辑 武林晓　　责任编辑 武林晓　　封面设计 张 志　　版式设计 范晓红

插图绘制 尹 莉　　责任校对 殷 然　　责任印制 张福涛

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京七色印务有限公司
开　　本 787mm×1092mm 1/16
印　　张 18.5
字　　数 370千字
购书热线 010-58581118
咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
　　http://www.hep.com.cn
网上订购 http://www.landraco.com
　　http://www.landraco.com.cn
版　　次 2011年6月第1版
印　　次 2011年6月第1次印刷
定　　价 29.00元

物料号 31896-00

前　言

我们所处的时代是以互联网为主要标志的海量信息时代。海量信息时代的主要特征之一，就是数字化信息的有效利用正在持续不断地、乐此不疲地向信息技术提出各种新的、超越人类固有思维和想象的需求和挑战。尽管身处多媒体世界，尽管身陷海量信息时代，依然有超过90%的信息是以语言文字作为载体，进行着输入、输出、存储、加工、传输、交换、检索、提取、过滤、管理等处理工作。换言之，语言文字是负载和传递信息的重要物质基础！因此，语言信息的自动处理无疑是当今科技研究、开发应用的关键内容之一。事关我们国家和民族的语言信息处理——中文信息处理，理所当然地就成为我国信息科学技术发展的一个制高点。

中文信息处理（包括对汉语及少数民族语言的信息处理）在我国信息领域科技进步与产业发展中，一直占据着特殊位置。从历史上看，汉字激光照排、汉卡、汉语拼音方案、汉字编码输入（诸如五笔字型）、手写输入等等，这些具有鲜明时代特征的自主创新成果，造就了中文信息处理产业的诞生、成长和成熟，并卓有成效地推动了我国信息化产业的发展。时至今日，面对海量数据的飞速发展，面对我国网民数量的持续增长，怎样架构良好的人机交互环境，如何提供有效的信息获取工具，无疑是社会和时代向中文信息处理提出的更多、更新、更高的需求，而中文信息处理自身也的确赢得了前所未有的机遇并面临严峻的挑战！因此可以有这样的定论：无论是基础理论研究还是应用领域的突破，只有中文信息处理自身得以高度发展，才能客观上保证我国国民充分、有效地利用海量信息。

语言信息处理的对象不仅包括语言，也包括记录和承载这种语言历史印迹、脉络以及发展进程的文字。而无论是语言还是文字，又都表现出其特有和固有的形、音、义等属性。因此对其进行持续、深入、细致的研究，且坚持从计算机处理的角度分析、提取、描述、形式化、模型化这些对象的特点，进而设计、实现自动处理这些对象的算法和系统，构建辅助人类进行信息获取、信息交流的工具，自然而然地成为语言信息处理研究的基础内容。而这些研究不但需要语言学、计算机科学、数学、思维科学等学科已有和新创的理论与方法的支持，而且亟须这些学科的深度交融！也正因为如此，语言信息处理本质上属于多学科交叉、具有多边缘性质的学科。

语言信息处理的对象决定了处理技术的应用。汉字自身的历史积淀及其独一无二的字形特点、汉语的无形态标记、连续书写等属性，决定了汉语在存储、处理、输入/输出等方面都需要有独特的技术支持。因此，具体到汉语信息处理，除了借鉴已有的语言信息处理技术外，更需面向汉字和汉语本体探究其特有的处理技术。同时，鉴于语言符号和语言意义与知识结构、思维方式有着密不可分的关系，若要有效地进行语言信息处理，自然离不开语言使用场景之外的知识。因此语言知识资源的建设也是语言信息处理研究中重要的内容。

本书的出发点是：梳理汉语信息处理研究的内容和技术，以汉语组成元素（汉字、词语、句子、篇章等）的处理技术和语言资源建设作为全书的组织线索，以关键内容和知识结构图作为章节内容的导入，以知识节点逐层展开的逻辑和模式将语言信息处理的技术呈现给读者，力求将中文信息处理中独有的技术和语言信息处理中通用的技术具体落实到汉语实例上，以实现集理论和实践为一体的目标，进而使本套教材成为高年级本科生和研究生名副其实地进入本领域学习、研究和发展的入门教材。

本书编写过程中，自身参考或向读者推荐了诸多在领域内有价值的网站，除个别官方网站地址已在书中直接标注之外，其他网站的地址则由编者汇集成文件，有需要者请在中国高校计算机课程网（http://computer.cncourse.com）下载。

本书的编者来自北京师范大学、北京语言大学和河北大学，自20世纪80年代后期始，上述3所大学在中文信息处理领域就陆续取得过具有时代特征的成果。本书的编写框架、思路以及理念也均来自上述3所院校本科“中文信息处理”课程以及研究生“自然语言处理”、“语言信息处理”等课程的实际需求和教学积淀。

本书编写过程中，北京师范大学语言与文字资源研究中心彭炜明博士、胡佳佳博士以及吴华琼、段磊、何静、王雪芝、韩芳、康明吉、刘天益等硕士研究生，在框架研讨、资料收集以及内容撰写等方面做了大量的、扎扎实实的工作。中国传媒大学国家语言监测与研究中心有声媒体语言分中心主任、博士生导师侯敏教授、中北大学刘冬明老师也对本书初稿给予了重要的修改建议！同时，本书参考了前修与时贤的众多文献及其研究成果，限于篇幅，不能一一列举，在此一并表示谢意！从这个意义上来讲，本书应该是集体智慧的结晶。

本书作为专业方向的教材，是引玉之砖。然而，囿于编者在学术素养、学科积淀以及领域认知等方面的局限和不足，难免挂一漏万和存在差错，敬请读者对错误、疏漏之处给予中肯的批评和指正，我们将诚心以待、认真接受！

编者　于北京

2011年4月15日

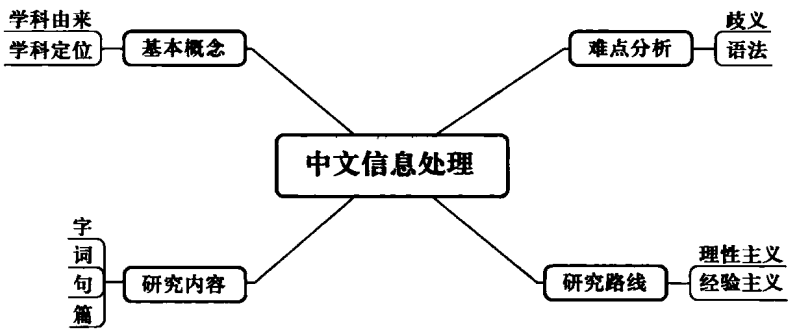
目　录

第1章 导论

本章概览

关键问题	阅读要点
何为“图灵测试”？	基本概念/学科由来
机器处理中文信息都要做哪些工作？	研究内容
为什么中文信息处理要研究语法？	难点分析/语法

知识结构图



1.1 基本概念

1.1.1 学科由来

从历史上第一台计算机诞生之日起，人类就没有停止过对机器智能的开发。机器智能也叫人工智能，就其本质而言，是对人类思维过程的模拟。而语言能力无疑是其中最重要的一个方面，因此，自然语言的信息处理从一开始就成为人工智能学科领域中首要的研究课题。这里举两个典型的例子。

一是人机对话。在早期关于机器智能的大讨论中，图灵就把人机之间的自然语言对话作为衡量机器是否具有智能的判断标准。1950 年 10 月，图灵在其具有划时代意义的论文《计算机器与智能》中提出了“机器能思考吗？”的问题，并为此设计了著名的“图灵测试”。

- 一个封闭的小屋。

- 屋外一个人。
- 屋内依次进入一个人和一台计算机。
- 屋外的人并不知道屋内是人还是计算机。
- 屋外的人向屋内的人和计算机提出各种问题。
- 屋外的人根据回答来判断屋内是人还是计算机。
- 如果判断不出来，那么可以认为计算机具有了智能。

计算机要能通过图灵测试，就必须具备一定的语言理解能力和语言生成能力。

二是机器翻译。几乎在计算机诞生的同时，人们就有了利用机器来进行语言自动翻译的想法。1954 年，IBM-701 型计算机首次完成了英俄机器翻译试验，向公众和科学界展示了机器翻译的可行性，从而拉开了机器翻译研究的序幕。如果说图灵测试还只是停留在计算机科学理论层面的探讨，那么机器翻译则是具体的语言工程问题了。

早期的机器翻译只依靠简单的单词对译规则，译文结果很难让人满意。很快人们便认识到自然语言的复杂性，1966 年 11 月，美国科学院发布的 ALPAC 报告全面否定了机器翻译的可行性，并建议停止对机器翻译项目的资金支持。虽然机器翻译转入低潮，但一门新的独立学科——计算语言学（Computational Linguistics）也从此确立了。

计算语言学是通过建立形式化的计算模型来分析、理解和处理自然语言的学科。随着科学技术的发展和各国信息交流的日益频繁，特别是由于互联网的兴起，信息处理中的语言信息比重日渐突出。在巨大的市场需求推动下，计算语言学获得了前所未有的蓬勃发展，应用范围也扩展到信息检索、信息抽取、文本挖掘等多个领域。由于学科的技术性和工程性进一步彰显，计算语言学也更多地被称为“语言信息处理”或“自然语言处理”（Natural Language Processing，NLP）。

中文信息处理（Chinese Information Processing）即中文的语言信息处理，而中文是指以汉语为主体，包括各少数民族语言在内的所有中华民族语言。本书以汉语为主要讨论对象，以下如不加说明，中文均指汉语而言。

1.1.2 学科定位

中文信息处理以汉语为研究对象，因此必然会涉及汉语的语言学知识。实际上，计算语言学本身就是计算机科学与语言学的一个交叉学科，在计算机科学中属于人工智能子学科，在语言学中属于应用语言学子学科，如图 1-1 所示。

任何自然语言都可看作是表达意义的符号系统，而符号分视觉符号（书面语）和听觉符号（口语）。根据符号性质的差异可以将中文信息处理的对象分为四类[詹卫东，2008]，如表 1-1 所示。

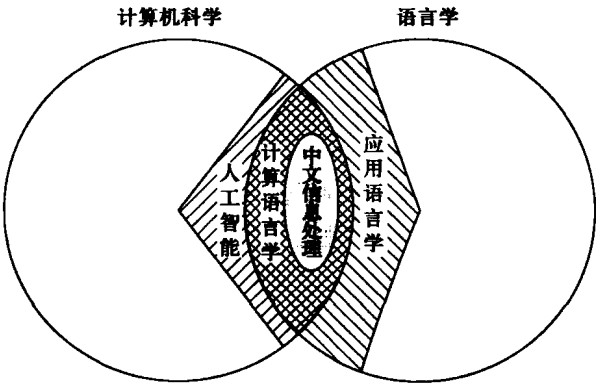


图 1-1 中文信息处理的学科地位

表 1-1 中文信息处理对象按符号性质分类

	书面文本[视觉符号]	口语语音[听觉符号]
处理符号的形式	汉字输入、存储、输出 篇章版式分解与生成	语音信号采集、 波形特征抽取、波形生成
处理符号的意义 [自然语言理解]	文本理解、机器翻译、信息检索、…… 文本生成、文本摘要、问答系统、……	语音识别、口语翻译、…… 语音合成、口语问答、……

其中，符号形式的处理在中文信息处理中相对成熟，且已经得到广泛运用。而符号意义的处理，即所谓自然语言理解，则是当前中文信息处理的核心和难点。在自然语言理解中，语音符号一般都要转化为文本符号进行处理，所以本书也主要以文本信息的处理为主。

1.2 研究内容

汉语的语言单位从小到大依次为语素、词、短语、句、段落和篇章，而记载汉语的书面符号为汉字。形式上，语素基本对应汉字（有少数例外），短语的分析参考句子，段落篇章对应整个文本，因此，按照汉语语言单位的层级可将中文信息处理的研究内容分为：汉字的信息处理、词的信息处理、句子的信息处理和文本（语篇）的信息处理。自然语言理解所关注的语言意义（语义）在各个层级都有反映，而句子是能够表达一个完全意思的语言单位，所以语义的表达与分析通常都选择在句子这一级展开。另外，各级处理过程都离不开基础数据资源的支持，而各类技术评测是推动中文信息处理不断向前发展的重要动力。综上所述，本书的内容安排如下：

（1）汉字的信息处理。主要包括汉字的输入、输出和编码，是进行中文信息处理

的前提条件。

（2）词的信息处理。汉语中词是连写的，词与词之间没有空格，也没有其他形态标记，因此在对汉语理解前需要先从输入的字符串系列中分解出单词并确定词性、词义，这一过程称为词法分析。

（3）句子的信息处理。对经过词法分析的单词/词性序列进行句法结构的剖析（parse），输出有效的句法树或短语片段，这一过程称为句法分析。

（4）句子语义表达与分析。语义的形式化是自然语言理解中最关键，也是最难的一步。目前句义处理尚处于理论探索阶段，还很不成熟，因此本书只介绍较有影响的几个语义框架：Fillmore 的格语法理论、Schank 的概念依存理论、黄曾阳的 HNC 理论，以及面向语义 Web 的领域本体技术。

（5）中文信息处理基础资源。可分为两类：语料库和知识库。基础资源的建设和应用是中文信息处理中一个重要研究课题，特别需要计算机学科和语言学科的通力协作。

（6）文本分析与处理。包括文本分类、信息检索、问答系统、信息抽取、自动文摘和机器翻译等，是中文信息处理中最高层级的应用，也是最活跃的研究区域。

（7）中文信息处理评测。技术评测不但为中文信息处理的研究提供了一个公共的展示平台，更重要的是，它能起到一定的技术导向作用。可以通过评测来关注中文信息处理研究的最新动向。

1.3 难点分析

1.3.1 歧义

歧义在自然语言中是普遍存在的，汉语的歧义可以分为以下几类。

（1）语音歧义。汉字为单音节文字，这样必然存在大量同音字。比如，著名语言学家赵元任先生曾用单一音节“shi”叙述了一个结构完整的故事。

施氏食狮史

石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮，十时，适十狮适市，是时，适施氏适市，施氏视是十狮，拭矢试，使是十狮逝世，适石室，石室湿，氏使侍拭石室。石室拭，始食是十狮尸，始识是十狮尸，实十石狮尸，试释是事。

同音字现象将给语音识别带来极大困难，而在进行拼音标注时又需要面对多音字现象，比如：快乐（le4）/音乐（yue4）。

（2）分词歧义。比如：

a）他/说/的/确实/在理。

他/说/的确/实在/理。

b）潘基文/秘书长

潘基/文秘/书/长

大部分的分词歧义可以根据短语或句中上下文判断，但有些涉及更大范围语义和结构的情况则很难解决，如："乒乓球拍卖完了"、"美国会通过某种法案制裁伊拉克"。

（3）结构歧义。比如：

a）[咬死猎人]的狗

咬死[猎人的狗]

b）[张三和李四]的朋友

张三和[李四的朋友]

（4）词义歧义。比如：

他说："她这个人真有意思（funny）"。她说："他这个人怪有意思的（funny）"。于是人们以为他们有了意思（wish），并让他向她意思意思（express）。他火了："我根本没有那个意思（thought）"！她也生气了："你们这么说是什么意思（intention）"？事后有人说："真有意思（funny）"。也有人说："真没意思（nonsense）"。

（5）语用歧义。比如：

这下可好了！（有希望了）

这下可好了！（完蛋了）

自然语言理解的主要目标就是求得语言表达中的意义，歧义的分析和消解是本学科需要解决的最主要问题，也是研究的难点之一。

1.3.2 语法

英语等西方语言的歧义处理很大程度上可以利用其语法形态上的变化，性、数、格、时态、语态等都在语言的表层形式上有所反映。而汉语是一种形态不发达的孤立语，这些直接的线索对于汉语分析来说并不适用。例如：

- 他跑着回来告诉我们这个消息。

这句共有"跑着"、"回来"、"告诉"三个动词，究竟哪一个是中心动词，由于无形态标志，很难加以区分。而与此句相应的英语句子为：

- He came running back to tell us the news.

由于 to 加动词原形构成不定式（目的状语），动词原形的词尾加 ing 构成现在分词（方式状语），所以计算机很容易确定该句的中心动词为 came（过去时）。中心动词的确立对于句法分析来说是至关重要的，因此，汉语的句法分析精度一直要低于英语。

汉语形态缺乏的影响其实远不止句法的分析过程，它甚至对语法体系本身也有一定的影响。比如词语的归类问题，由于同一词充当不同句法成分时没有形态变化，导致词形、词类、成分之间的对应很不平衡。例如：

- 他在教育儿子。
- 这次教育很成功。
- 教育的效果很明显。

三句中的“教育”分别处在不同句法成分中，但基本意义是一样的，如果归入同一类，则存在所谓“类无定职”的问题（成分与词类不对应）；如果归入不同类，则又有所谓“词无定类”问题（词形与词类不对应）。这两种情况对句法分析来说都是不利的，“类无定职”则句法系统规则混乱，而“词无定类”则词库无法定义。这本身就是一个两难的选择，处理起来还需要分辨兼类和活用等不同情况，因此变得非常复杂。比如上例就是“教育”的兼类，而下面的例句则为词类活用，需要做不同处理。

- 这件事情他做得很男人。
- 我们都被就业了。

最后是汉语的语序和省略现象。在语言的组织上，汉语的意合性格外突出，而语序的限制也就相对宽松，例如：

- 我吃了苹果。
- 苹果我吃了。
- 我苹果吃了。

除了语序之外，省略现象也是对句法系统的一大冲击。只要语义清晰，主要成分、虚助词等都可以省略而不影响表达。例如：

- 半夜里，（）忽然醒来，（）才觉得寒气逼人，刺入肌骨，（）浑身打着战。
- 平时多流汗，战时少流血。→（只有）平时多流汗，（才能）战时少流血。

句法分析要求有一个相对稳定的语法体系（规则系统和词汇系统），而汉语语法的灵活性却使得中文信息处理的复杂性比一般自然语言处理都要高。如何进行有效的语法分析是本学科的又一难题。

1.4 研究路线

在自然语言理解的研究方法上，历来有理性主义（Rationalist）和经验主义（Empiricist）两种路线之争。理性主义以规则方法为主，而经验主义以统计方法（机器学习）为主。它们的区别其实是在哲学层面对语言本质的不同认识引起的。从语言学派上讲，理性主义属于结构主义阵营，而经验主义属于功能主义阵营。

理性主义认为，只有探明人类语言理解的内部机制，才有可能真正实现机器的自

然语言理解。因此，它更强调语言学知识，特别是抽象的语言规则系统，而把实际语料当作对其系统的验证。

经验主义注意到，任何规则都有例外，自然语言的复杂性仅靠语言学家内省出来的规则系统是无法驾驭的，因为人类至今尚未透析自身理解语言的内在机理。因此从实用角度出发，它只追求机器在某一具体任务中的表现能和人一致，是一种任务驱动的研究方法。大规模语料库的应用是经验主义的研究基础。经验主义认为语言知识和规律都蕴藏在大规模真实文本中，人工归纳带有个人主观性，且总不免有所遗漏，不如让机器结合实际任务从语料库统计，按需而取。虽然语料库也不可能穷尽所有语言实例，但只要语料规模扩大到可以涵盖大多数情况时就可以实现任务目标。而巨大存储量和快速检索能力正是计算机的强项，这就为采用统计方法的经验主义路线创造了条件。

机器翻译一直以来就是自然语言理解的最前沿课题，可看作是计算语言学发展的风向标。早期的机器翻译走的是理性主义路线，在 ALPAC 报告后基本陷入停滞阶段。大规模电子语料库的建成使得经验主义后来居上，从 20 世纪 90 年代起统计方法就开始成为机器翻译的主流。随着互联网时代的到来，网络语言一方面为统计提供了海量的数据，另一方面也不断冲击着自然语言的固有规则系统，使得统计方法更加占据上风。

除了机器翻译，从目前中文信息处理的其他各项评测中也可以看出，基于统计的方法已经全面超越了基于规则的方法。但在统计方法取得各种成绩的同时，也应该看到其背后的种种弊端，当语料规模增长到一定程度时，其精度必然会达到一个极限。因为概率总是“照顾多数、忽略少数”，那么处于少数的那部分语言现象就必然无法顾及。而且从本质上说，统计模型在自然语言理解时只是出于对现有语料的一种概率拟合，它并没有真正理解语言。任务驱动性质决定了模型只是适应于语言表层形式的计算，很难走向语义理解。

从人机交互的角度来看，自然语言理解其实是一个人机相互适应的过程。采用统计方法建立语言模型是让自然语言适应机器学习，而规则方法则是试图让机器向人学习语言理解机制，当然这个机制本身目前尚在探索中。统计方法代表的是工程路线，规则方法代表的是科学路线，可以分别比作软件工程中的黑盒测试和白盒测试。当然，目前白盒测试还欠缺各种条件，权宜之计下行之有效的只能是黑盒测试，但如何实现从黑盒到白盒的转变是自然语言处理工作者应深思的问题，这也就是为什么理性主义仍有其存在的必要。

1.5 习题

1.“名实之辨”：上网搜索如下几个名词，并根据搜索结果比较它们的异同。