高等学校计算机专业规划教材

中文信息处理原理及应用（第2版）

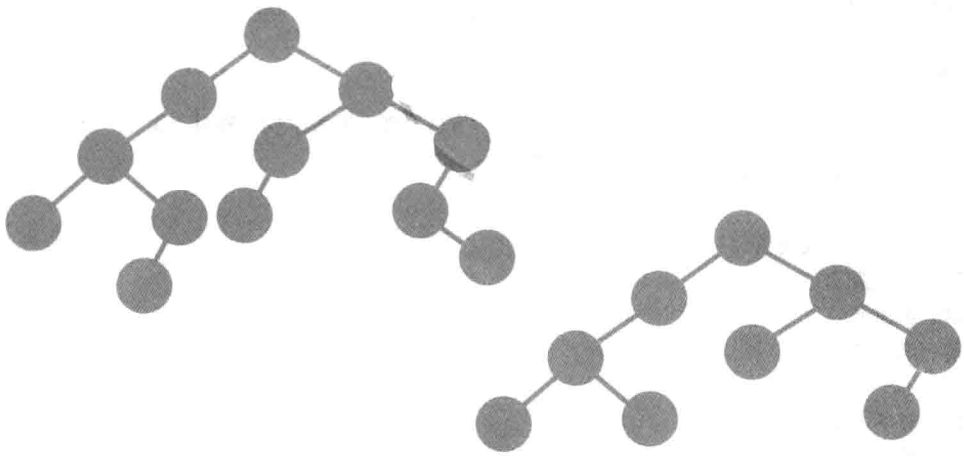
苗夺谦 卫志华 张志飞 编著

清华大学出版社

高等学校计算机专业规划教

中文信息处理原理及应用（第2版）

苗夺谦 卫志华 张志飞 编著



清华大学出版社

内 容 简 介

本书全面系统地介绍了中文文本信息处理技术，由浅入深地讲述了中文文本理解的原理与应用。全书共5个部分：预备知识、词法分析、语法分析、语义分析和应用与技术。预备知识部分介绍了本书所需的数学、语言学和形式语言与自动机方面的理论知识。词法分析、语法分析和语义分析是自然语言处理的基础。词法分析部分针对中文信息处理中特有的分词问题，介绍了若干分词算法以及分词歧义消除和未登录词识别算法。语法分析和语义分析两部分从语法(语义)的表示入手，介绍自然语言的结构化和形式化，给出语法分析和语义分析的常用算法，并针对该过程中的歧义问题给出可行的解决思路。应用与技术部分讲述中文信息处理的应用，尤其是在文本分类、信息检索、问答系统和自动文摘等领域的应用技术。

本书涉及内容广泛，能满足不同层次读者群的需求，可以作为高等学校计算机、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，同时非常适合供自然语言处理应用领域的研究人员和技术人员参考。

图书在版编目(CIP)数据

中文信息处理原理及应用/苗夺谦，卫志华，张志飞编著. —2版. —北京：清华大学出版社，2015
高等学校计算机专业规划教材
ISBN 978-7-302-38950-7

Ⅰ. ①中… Ⅱ. ①苗… ②卫… ③张… Ⅲ. ①中文信息处理—高等学校—教材 Ⅳ. ①TP391.12

中国版本图书馆CIP数据核字(2015)第006101号

责任编辑：龙启铭　战晓雷
封面设计：何凤霞
责任校对：焦丽丽
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954

印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：18.25　　**字　　数**：423千字
版　　次：2007年8月第1版　2015年3月第2版　　**印　　次**：2015年3月第1次印刷
印　　数：1～2000
定　　价：39.00元

产品编号：059286-01

前言

自然语言处理(Natural Language Processing,NLP)就是研究让计算机理解并生成人们日常所使用的语言的理论、方法与技术。其目的在于建立起一种人与机器之间密切而友好的交流,使机器能进行高度的信息传递与认知活动。自然语言处理的研究始于20世纪50年代,经过60多年的发展取得了长足的进展,已经成为人工智能的一个重要研究领域。

目前,在人工智能界,对机器自然语言理解尚无统一的定性准则和量化标准。一般认为,根据著名的图灵测试,至少有以下4条准则可用于判断计算机是否"理解"了某种自然语言:机器能正确回答输入文本中的有关问题;机器有能力生成输入文本的摘要;机器可以用不同的词语和句型来复述输入文本;机器具有将一种自然语言(源语)的输入文本翻译成另一种自然语言(目标语)文本的能力。

2011年IBM公司的一台名为"沃森"(Watson)的超级计算机在美国最有影响的智力竞猜节目《危险边缘》(Jeopardy)的比赛中,以3倍的巨大分数优势力压另两位参赛选手肯·詹宁斯和布拉德·鲁特,夺得这场人机大战的冠军。《纽约时报》评论称,这场比赛对于IBM公司来说,不是为了成为公众瞩目的焦点或是赢得100万美元的奖金,而是希望证明IBM公司在人工智能领域已经迈出了一大步,这些智能机器将来可以理解人类语言并做出反应,甚至不可避免地将取代一部分人。"沃森"的胜利又一次掀起了自然语言处理研究的热潮。除IBM公司外,Microsoft、Google、Apple、Facebook和百度等国内外著名的IT公司都投入巨大的人力、物力与财力开展自然语言处理的研究与开发工作。

需求是推动技术发展和进步的源泉。现在是一个在线信息、电子通信和互联网流行的年代,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,而如何从这些大量文本中挖掘潜在的有用信息,是一个很有价值的课题。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。因此,对自然语言理解的研究方兴未艾,具有广阔的发展前景。

全世界六分之一以上的人口使用中文,作为自然语言处理的重要分支,中文信息处理越来越受到工业界和学术界的重视。一些国际著名的IT公

司都成立专门针对中文信息处理的研究机构，很多高校开设了中文信息处理课程。

在工业界强大需求的推动下，无论作为在校的计算机和信息学专业学生，还是工作在中文处理领域的工程技术人员，都需要夯实中文信息处理的基础。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处理，尤其是中文信息处理的理论和技术。本书正是应这样的需求而编写的，我们希望本书的出版有助于相关人员更全面、更准确地掌握中文信息处理已有的研究成果和最新进展，最终有助于促进我国中文信息处理的快速发展。

本书是在我校和多家兄弟院校多年教学实践基础上凝炼而成的，目前是上海市精品课程“中文信息处理”的主讲教材。本书主要介绍中文分词、语法分析和语义分析的理论与技术，以及中文信息处理技术在文本分类、信息检索、问答系统和自动文摘等应用领域的建模方法。

自然语言处理包括语言的理解和语言的生成过程，应当是一个层次化的过程，分为词法分析（词法层）、句法分析（语法层）、语义分析（语义层）、语言生成（语用层）。它们密切协作才能完成好自然语言的处理。与本书第1版相比，第2版继续沿用4个主要部分的组织结构，即词法分析、语法分析、语义分析、应用与技术，但是在内容和形式上有所不同。在内容方面：第2版在概论之后增加了自然语言处理涉及的一些预备知识，如概率论、随机过程、信息论、形式语言与自动机以及语言学知识；在词法分析部分，增加了若干当前较为流行的统计分词方法；在语法分析部分，增加了当前较热门的依存句法分析；在语义分析部分，增加了概念标注；在应用与技术部分，将信息抽取替换为问答系统。同时，删除了一些在中文自然语言理解中应用相对少的算法，如在语法分析部分删除了移进-归约句法分析器；语料库构建也不再作为单独的一章，而是放在第5章作为一节。在形式方面：在每部分介绍前增加了知识结构图，有利于读者从整体上把握内容，特别需要重点掌握加粗显示的知识点；在介绍具体方法时给出例子以帮助理解，如分词方法、句法分析方法等；此外，将涉及的相关评测独立为一节做专门介绍，并对历届评测的信息及时更新，鼓励读者参与以提高动手能力。

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

由于作者知识水平有限，错误在所难免，诚望广大读者提出批评和指正。

作　者

于上海同济嘉园

2015年1月

目录

第 1 部分　预 备 知 识

第 1 章　概论　/3

1.1　自然语言处理与中文信息处理 …… 3
1.1.1　自然语言处理 …… 3
1.1.2　中文信息处理 …… 4
1.2　研究内容 …… 6
1.3　应用领域 …… 8

第 2 章　预备知识　/9

2.1　数学基础 …… 9
2.1.1　概率论 …… 9
2.1.2　随机过程 …… 11
2.1.3　信息论 …… 18
2.1.4　形式语言与自动机 …… 21
2.2　语言学基础 …… 24
2.2.1　计算语言学概述 …… 24
2.2.2　语素和词 …… 24
2.2.3　句法与篇章语法 …… 25
2.2.4　词义与句义 …… 27

第 2 部分　词 法 分 析

第 3 章　自动分词概述　/33

3.1　自动分词 …… 33
3.1.1　分词规范 …… 33
3.1.2　自动分词的研究内容及意义 …… 34
3.1.3　自动分词方法 …… 34
3.2　分词歧义问题 …… 35
3.3　未登录词问题 …… 37
3.4　自动分词评测 …… 39

第4章 基于词典的分词方法 /43

4.1 分词词典 …… 43
4.1.1 关于分词词典的构造 …… 43
4.1.2 基于词属性的分词词典 …… 44
4.1.3 基于逐字二分的分词词典 …… 45
4.2 机械分词方法 …… 46
4.2.1 正向最大匹配算法 …… 46
4.2.2 逆向最大匹配算法 …… 47
4.2.3 邻近匹配算法 …… 48
4.2.4 最短路径匹配算法 …… 49
4.3 基于规则的分词方法 …… 51
4.3.1 分词预处理中的规则 …… 51
4.3.2 分词规则 …… 52
4.4 中文姓名切分 …… 54
4.4.1 切分姓名中的当用资源 …… 54
4.4.2 同源对表、互斥对表及其操作 …… 57
4.4.3 姓名左右边界的确定 …… 57
4.4.4 屏蔽与恢复 …… 58
4.4.5 同源对表和互斥对表的校正规则 …… 58
4.4.6 概率再筛选 …… 59
4.4.7 中文姓名切分系统 …… 59

第5章 基于语料库的分词方法 /61

5.1 语料库 …… 61
5.1.1 语料库概述 …… 61
5.1.2 语料库加工规范 …… 65
5.1.3 现代汉语语料库构建实例 …… 71
5.2 基于统计的分词方法 …… 72
5.2.1 统计分词概述 …… 72
5.2.2 统计分词消歧 …… 73
5.2.3 统计未登录词获取 …… 76
5.2.4 统计分词模型 …… 83
5.3 基于机器学习的分词方法 …… 85
5.3.1 最大熵分词 …… 86
5.3.2 条件随机场分词 …… 88

第2部分习题 /91

第3部分 语法分析

第6章 自动词性标注 /95

6.1 词性标注概述 …… 95
6.1.1 词性标注 …… 95
6.1.2 词性标记规范 …… 96
6.1.3 词性消歧 …… 97
6.1.4 词性标注评测 …… 98
6.2 基于统计的词性标注方法 …… 99
6.2.1 统计模型的训练 …… 99
6.2.2 马尔可夫模型标注方法 …… 100
6.2.3 隐马尔可夫模型标注方法 …… 103
6.3 基于规则的词性标注方法 …… 106
6.3.1 按兼类词搭配关系构造的规则 …… 106
6.3.2 按词语结构获取的规则 …… 107
6.4 其他标注方法 …… 108
6.4.1 基于规则和统计相结合的标注方法 …… 108
6.4.2 基于条件随机场的词性标注方法 …… 109
6.4.3 词性标注中的未登录词处理方法 …… 109

第7章 语法表示方法 /110

7.1 语法表示概述 …… 110
7.2 形式语法描述 …… 110
7.2.1 重写规则 …… 110
7.2.2 转移网络 …… 112
7.3 短语结构语法 …… 113
7.4 依存语法 …… 115

第8章 句法分析方法 /117

8.1 句法分析概述 …… 117
8.1.1 句法分析 …… 117
8.1.2 结构歧义 …… 118
8.1.3 句法分析评测 …… 119
8.2 基于规则的句法分析方法 …… 120
8.2.1 自顶向下句法分析 …… 121
8.2.2 自底向上句法分析 …… 122
8.2.3 线图句法分析 …… 124

8.2.4 转移网络句法分析…… 126
8.3 基于统计的句法分析方法 …… 129
8.3.1 概率上下文无关文法分析…… 129
8.3.2 依存句法分析…… 137

第3部分习题 /142

第4部分 语义分析

第9章 概念标注 /147

9.1 概念标注概述 …… 147
9.2 语言知识库 …… 148
9.3 概念标注方法 …… 150

第10章 语义表示 /154

10.1 语义表示概述…… 154
10.2 语义逻辑表示法…… 155
10.2.1 一阶谓词演算…… 155
10.2.2 基本逻辑形式语言…… 157
10.2.3 逻辑形式中的歧义表示…… 159
10.2.4 论旨角色…… 160
10.3 语义网络表示法…… 161
10.4 语义框架表示法…… 162

第11章 语义分析 /166

11.1 语义分析概述…… 166
11.2 基于语义特征的语义分析…… 167
11.2.1 组合理论…… 167
11.2.2 λ表达式与语义解释 …… 168
11.2.3 带语义解释的简单语法和词典…… 170
11.2.4 语义角色…… 172
11.2.5 特征合一的语义解释…… 173
11.3 基于语法关系的语义分析…… 176
11.4 基于模板匹配的语义分析…… 179
11.5 语义消歧…… 183
11.5.1 语义消歧概述…… 183
11.5.2 基于规则的语义消歧…… 184
11.5.3 基于统计的语义消歧…… 193

第 4 部分习题　/199

第 5 部分　应用与技术

第 12 章　文本分类　/203

12.1　文本分类概述…… 203
12.1.1　自动文本分类定义…… 203
12.1.2　文本分类任务的特点…… 204
12.1.3　文本分类基本实现途径…… 204
12.1.4　文本分类的组成…… 205
12.1.5　文本分类的应用领域…… 206
12.1.6　国内外研究现状…… 207
12.2　文本分类方法…… 208
12.2.1　文本表示与文本特征选择…… 208
12.2.2　分类器设计…… 211
12.2.3　分类器的阈值选择…… 215
12.3　文本分类评测…… 216
12.3.1　单类赋值…… 216
12.3.2　多类排序…… 218

第 13 章　信息检索　/219

13.1　信息检索概述…… 219
13.1.1　信息检索的对象和任务…… 219
13.1.2　信息检索的评测…… 220
13.1.3　信息检索模型…… 220
13.1.4　中文信息检索的特点…… 222
13.2　基于统计的信息检索模型…… 222
13.2.1　布尔模型及其扩展…… 222
13.2.2　向量空间模型…… 224
13.2.3　概率模型…… 232
13.3　基于语义的信息检索…… 239
13.3.1　基于 NLP 的方法 …… 239
13.3.2　潜在语义索引…… 241
13.3.3　基于神经网络的信息检索…… 246
13.4　信息检索技术评测…… 247
13.4.1　文本检索会议…… 247
13.4.2　亚洲语言信息检索评测会议…… 248
13.4.3　863 信息检索评测项目 …… 248

13.5 Web信息检索 …… 249
13.5.1 Web信息检索的特点 …… 249
13.5.2 搜索引擎 …… 251

第14章 问答系统 /258

14.1 问答系统概述 …… 258
14.1.1 问答系统的发展 …… 258
14.1.2 问答系统的定义 …… 259
14.1.3 问答系统的研究趋势 …… 260
14.2 关键技术 …… 260
14.2.1 关键词抽取 …… 261
14.2.2 关键词扩展 …… 263
14.3 问答系统评测 …… 264
14.4 Watson问答系统 …… 265

第15章 自动文摘 /267

15.1 自动文摘概述 …… 267
15.1.1 文摘的定义 …… 267
15.1.2 文摘的分类 …… 268
15.1.3 自动文摘的意义 …… 269
15.2 自动文摘的方法 …… 269
15.2.1 基于统计的自动文摘 …… 269
15.2.2 基于理解的自动文摘 …… 270
15.2.3 基于信息抽取的自动文摘方法 …… 271
15.2.4 基于结构的自动文摘 …… 271
15.3 自动文摘系统评测 …… 272
15.3.1 内部评价 …… 272
15.3.2 外部评价 …… 273
15.4 自动文摘系统 …… 273

第5部分习题 /275

附录A 北京大学计算语言学研究所汉语词性标注标记集 /276

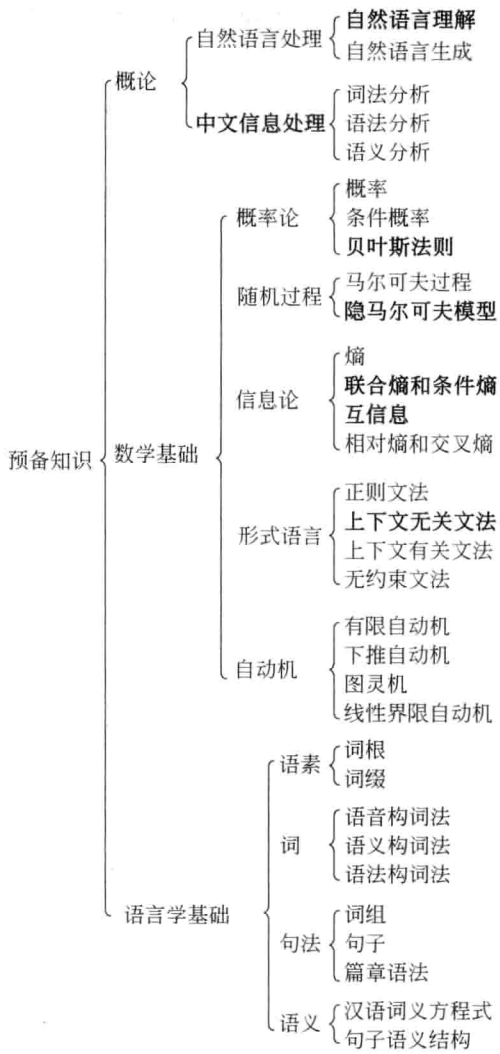
附录B 哈尔滨工业大学CDT依存句法标注体系 /278

参考文献 /279

第 1 部分

预 备 知 识

知识结构图



自然语言处理研究人机自然语言交互的理论与方法，主要包括自然语言理解和自然语言生成两个部分，前者要求计算机理解自然语言文本的意义，后者要求计算机用自然语言文本表达特定意图。本书围绕自然语言理解进行介绍。

中文信息处理是自然语言处理中以汉语作为研究对象的研究分支。其研究内容包括词法分析、语法分析和语义分析，三者层次由低到高，逐步递进。中文信息处理中存在着大量的歧义现象，如分词歧义、词义歧义、结构歧义等，但是一般情况下可以根据相应的语境和场景的规定进行解决。

自然语言处理是一门融计算机科学、语言学、数学于一体的科学。本部分除了绪论外，主要介绍数学和语言学等方面的预备知识。数学方面涉及概率论、随机过程、信息论、形式语言和自动机；语言学方法涉及语素、词、句法、语义等。特别是数学基础知识，时刻贯穿于自然语言处理的各个环节。例如，概率论和信息论常用于分词、词性标注、句法分析等以及歧义消解，随机过程以隐马尔可夫模型为代表用于词性标注等序列标注问题，形式语言和自动机常用于语义分析。

第1章 概　论

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言。在人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上。就计算机的应用而言，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(Natural Language Processing，NLP)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沟通的各种理论和方法。这也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是人们长期以来所追求的目标。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，而无需再花大量的时间和精力去学习不是很符合自然和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。

正如对于“智能”的定义一样，对于“理解”这个术语也存在着各式各样的认识。然而在人工智能领域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4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。

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的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以下与自然语言理解有关的知识。

（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，如西文中的前后缀、汉语中的偏旁部首。

（3）句法知识：研究词语如何排列以组成正确的句子，并决定每个词语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，以及该使用如何影响句子的解释。

（6）篇章知识：研究在前面的句子影响下，后面的句子如何解释，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。

以上几方面的语言学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。

按照人工智能领域和计算机领域中最常用的思维方式，这里的首要任务就是将研究对象在计算机中表示出来，即如何将上述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”。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。

1.1.2 中文信息处理

中文信息处理以汉语作为研究对象，是自然语言处理的重要研究方向。与西文相比，中文处理的主要障碍体现为以下3个问题：第一，输入问题，汉字不是拼音文字，而是像

形文字或音形结合的文字；第二，分词问题，多数中文句子是一长串连续的汉字（而不是以空格或其他分隔标记分开的单词），并且词汇缺少明显的形态变化；进而引发第三个问题，即语法分析问题。

随着CPU从8位发展到16位，又从16位发展到32位，朱邦复等开发了仓颉码，解决了汉语的输入问题，使得计算机在使用汉语的人群中得到普及。这种发展使得更大、更完善的字符集得到应用，并且可以充分利用计算语言学的研究成果进行处理。原先计算机输入是由专门的操作员用烦琐的4位电报码或者使用笨拙的多层大键盘，现在则被少于50个键的QWERT键盘取代。

目前，第二个障碍在一定程度上得到克服，但仍然存在。这是因为切分连续的汉字文本比切分英语或德语文本要困难得多。例如，英语中书写人名时首字母要使用大写字母，而汉语中人名没有类似的标记；又如，德语中的复合名词连写成一个字串，而汉语中字词不分，都连写。然而，以下几个方面的发展使得中文分词水平不断提升：

（1）大规模语言数据库的发展；

（2）其他语言资源的建立，例如劳动密集型的词汇、语义资源；

（3）计算机处理能力的迅速提高。这种突破还表现在大量应用系统的诞生，包括语音产品、基于因特网和个人计算机的搜索引擎以及查询和翻译的半自动系统的开发等。

第三个障碍仍是主要的拦路虎，根本原因在于对汉语进行词性标注可以利用的格式标记太少甚至没有，这使得必要的语法分析变得困难。

事实上，无论是中文分词还是语法分析，首要的困难就在于歧义问题，即从下一层次向上一层次转变中存在着歧义和多义现象，比如形式上一样的一段字符串，在不同的场景或不同的语境下，可以理解成不同的词串、词组串等，并有不同的意义。通过下面的一些例子，可以更深入地理解中文信息处理的困难。

（1）分词歧义：即由字到词时的歧义现象，这是中文信息处理中独有的，首先需要解决的问题。例如：

他/的/确切/地址/在/这儿。
他/的确/切/地址/在/这儿。

南京市/长江/大桥。
南京/市长/江大桥。

（2）词义歧义：即一词多义或者一词多词性现象，这在所有的自然语言中都存在，中文也不例外，甚至更多。例如：

汉语**学习**十分重要。（“学习”为名词）
他们努力**学习**汉语。（“学习”为动词）

红花。（“红”指“红色的”）
红军。（“红”指“革命的”）

（3）结构歧义：即由词组成词组乃至句子时，由于其组成的词或词组间可能存在不

同的语法或语义关系而出现的(潜在)歧义现象。例如：

[衣服的袖子]和[口袋]
衣服的[袖子和口袋]

[桌子和椅子]的腿
桌子和[椅子的腿]

(4) 语用歧义：即短语或句子的意义可以有多种，而其真正的意义需要根据语言使用的情景确定。例如：

他说不清楚。(他不了解某事)
他说不清楚。(他无法表达清楚某事)

从以上介绍中不难看出中文信息处理中存在着大量的歧义现象。一般情况下，它们中的大多数都可以根据相应的语境和场景的规定而得到解决。也就是说，从总体上说，并不存在歧义。这也就是日常交流时通常并不感到自然语言歧义的原因。

另一方面，由于歧义所影响的范围不仅仅是词汇级，而且往往影响到句子级甚至语境级，因此，虽然力图在分词阶段消除歧义切分现象，但是在很多情况下，也不得不在后续的词性标注、语法分析和语义分析阶段继续处理歧义问题。消解歧义需要极其大量的知识以及推理。如何将这些知识较完整地加以收集和整理，又如何找到合适的形式，将它们存入计算机系统，以及如何有效地利用它们来消除歧义，都是工作量庞大且十分困难的工作。这不是少数人短时期内可以完成的，还有待长期的、系统的工作。

1.2 研究内容

中文信息处理的研究内容按照层次由低到高可以分为词法分析、语法分析和语义分析。

(1) 词法分析：即由字到词的分析过程。针对中文书写中词与词之间没有空格的问题，将汉字(包括标点符号)组成的字符串自动地划分出词汇，也就是自动分词。词法分析是中文信息处理的基础，后续的词性标注、语法分析和语义分析等工作都依赖于词法分析的结果。例如，对于图1.1所示的文本片段，经过自动分词后，形成如图1.2所示的分析结果。

富士山是日本的一座活火山
山峰终年积雪，云雾围绕，只有在空气干燥的秋冬两季，才能看清它的全貌。
多变的气候，更为它增添了神秘的色彩，甚至使它孕育了许多美丽的神话。
富士山的景色，四季不同。
春天，山顶还戴着雪帽子，山腰的雪却融化了，细碎的小花开遍山坡，远看像一片紫色的海洋；夏天，残雪与山花倒映湖中，充满诗情画意；秋天，满山红叶与雪影辉映，像个娇羞的姑娘；冬天则是纯白的一片，庄严而圣洁。

图1.1 未经处理的文本片段

(摘自北京大学计算语言所语料库)