



内容简介

本书全面介绍了统计自然语言处理的基本概念、理论方法和最新研究进展，内容包括形式语言与自动机及其在自然语言处理中的应用、语言模型、概率图模型、语料库技术、汉语自动分词与词性标注、句法分析、语义分析、篇章分析、统计机器翻译、语音翻译、文本分类、信息检索与问答系统、自动文摘和信息抽取、口语信息处理与人机对话系统等，既有对基础知识和理论模型的介绍，也有对相关问题的研究背景、实现方法和技术现状的详细阐述。

本书可作为高等院校计算机、信息技术等相关专业的高年级本科生或研究生的教材或参考书，也可供从事自然语言处理、数据挖掘和人工智能等研究的相关人员参考。

目 录

中文信息处理丛书序言	倪光南
序一	高庆狮
序二	冯志伟
第2版前言	
第1版前言	
第1章 绪论	1
1.1 基本概念	1
1.1.1 语言学与语音学	1
1.1.2 自然语言处理	2
1.1.3 关于“理解”的标准	5
1.2 自然语言处理研究的内容和面临的困难	5
1.2.1 自然语言处理研究的内容	5
1.2.2 自然语言处理涉及的几个层次	8
1.2.3 自然语言处理面临的困难	8
1.3 自然语言处理的基本方法及其发展	11
1.3.1 自然语言处理的基本方法	11
1.3.2 自然语言处理的发展	12
1.4 自然语言处理研究现状	15

1.5 本书的内容安排	16
第2章 预备知识	18
2.1 概率论基本概念	18
2.1.1 概率	18
2.1.2 最大似然估计	18
2.1.3 条件概率	19
2.1.4 贝叶斯法则	19
2.1.5 随机变量	20
2.1.6 二项式分布	21
2.1.7 联合概率分布和条件概率分布	21
2.1.8 贝叶斯决策理论	22
2.1.9 期望和方差	22
2.2 信息论基本概念	23
2.2.1 熵	23
2.2.2 联合熵和条件熵	24
2.2.3 互信息	26
2.2.4 相对熵	27
2.2.5 交叉熵	27
2.2.6 困惑度	28
2.2.7 噪声信道模型	28
2.3 支持向量机	30
2.3.1 线性分类	30
2.3.2 线性不可分	31
2.3.3 构造核函数	31
第3章 形式语言与自动机	33
3.1 基本概念	33
3.1.1 图	33
3.1.2 树	33
3.1.3 字符串	34
3.2 形式语言	35
3.2.1 概述	35
3.2.2 形式语法的定义	35
3.2.3 形式语法的类型	36
3.2.4 CFG 识别句子的派生树表示	38
3.3 自动机理论	39
3.3.1 有限自动机	39
3.3.2 正则文法与自动机的关系	40
3.3.3 上下文无关文法与下推自动机	41
3.3.4 图灵机	43
3.3.5 线性界限自动机	44

3.4	自动机在自然语言处理中的应用	45
3.4.1	单词拼写检查	45
3.4.2	单词形态分析	48
3.4.3	词性消歧	49
第4章	语料库与词汇知识库	53
4.1	语料库技术	53
4.1.1	概述	53
4.1.2	语料库语言学的发展	54
4.1.3	语料库的类型	57
4.1.4	汉语语料库建设中的问题	59
4.1.5	典型语料库介绍	60
4.2	词汇知识库	67
4.2.1	WordNet	68
4.2.2	FrameNet	69
4.2.3	EDR	70
4.2.4	北京大学综合型语言知识库	71
4.2.5	知网	73
4.2.6	概念层次网络	77
4.3	语言知识库与本体论	79
第5章	语言模型	83
5.1	n 元语法	83
5.2	语言模型性能评价	85
5.3	数据平滑	86
5.3.1	问题的提出	86
5.3.2	加法平滑方法	87
5.3.3	古德-图灵(Good-Turing)估计法	87
5.3.4	Katz 平滑方法	89
5.3.5	Jelinek-Mercer 平滑方法	90
5.3.6	Witten-Bell 平滑方法	92
5.3.7	绝对减值法	93
5.3.8	Kneser-Ney 平滑方法	93
5.3.9	算法总结	95
5.4	其它平滑方法	97
5.4.1	Church-Gale 平滑方法	97
5.4.2	贝叶斯平滑方法	97
5.4.3	修正的 Kneser-Ney 平滑方法	98
5.5	平滑方法的比较	99
5.6	语言模型自适应方法	100
5.6.1	基于缓存记忆的语言模型	100

5.6.2 基于混合方法的语言模型	101
5.6.3 基于最大熵的语言模型	102
第6章 概率图模型	104
6.1 概述	104
6.2 贝叶斯网	106
6.3 马尔柯夫模型	108
6.4 隐马尔柯夫模型	110
6.4.1 求解观察序列的概率	111
6.4.2 维特比算法	115
6.4.3 HMM 的参数估计	116
6.5 层次化隐马尔柯夫模型	119
6.6 马尔柯夫网络	120
6.7 最大熵模型	122
6.7.1 最大熵原理	122
6.7.2 最大熵模型的参数训练	124
6.8 最大熵马尔柯夫模型	125
6.9 条件随机场	127
第7章 自动分词、命名实体识别与词性标注	129
7.1 汉语自动分词中的基本问题	129
7.1.1 汉语分词规范问题	129
7.1.2 歧义切分问题	130
7.1.3 未登录词问题	132
7.2 汉语分词方法	135
7.2.1 N -最短路径方法	135
7.2.2 基于词的 n 元语法模型的分词方法	138
7.2.3 由字构词的汉语分词方法	140
7.2.4 基于词感知机算法的汉语分词方法	142
7.2.5 基于字的生成模型和区分式模型相结合的汉语分词方法	144
7.2.6 其他分词方法	146
7.2.7 分词方法比较	147
7.3 命名实体识别	150
7.3.1 方法概述	150
7.3.2 基于 CRF 的命名实体识别方法	152
7.3.3 基于多特征的命名实体识别方法	154
7.4 维吾尔语人名识别方法研究	162
7.5 词性标注	164

7.5.1 概述	164
7.5.2 基于统计模型的词性标注方法	165
7.5.3 基于规则的词性标注方法	168
7.5.4 统计方法与规则方法相结合的词性标注方法	170
7.5.5 词性标注中的生词处理方法	172
7.6 词性标注的一致性检查与自动校对	173
7.6.1 词性标注一致性检查方法	173
7.6.2 词性标注自动校对方法	175
7.7 关于技术评测	177
第8章 句法分析	179
8.1 句法结构分析概述	179
8.1.1 基本概念	179
8.1.2 语法形式化	180
8.1.3 基本方法	181
8.2 基于PCFG的基本分析方法.....	184
8.2.1 PCFG.....	184
8.2.2 面向 PCFG 的内向外向算法.....	185
8.2.3 选择句子的最佳结构.....	187
8.2.4 PCFG 的概率参数估计.....	188
8.2.5 分析实例.....	190
8.3 词汇化的短语结构分析器.....	192
8.4 非词汇化句法分析器.....	196
8.5 其他相关研究	199
8.5.1 PCFG 方法的改进	199
8.5.2 数据驱动的分析方法	200
8.5.3 语义信息的利用	202
8.6 短语结构分析器性能评价	202
8.6.1 评价指标.....	202
8.6.2 短语结构分析器性能比较.....	204
8.7 层次化汉语长句结构分析.....	207
8.7.1 标点符号在句法分析中的作用.....	208
8.7.2 层次化汉语长句结构分析的思路.....	209
8.7.3 汉语标点符号的分类.....	210
8.7.4 句法规则提取方法.....	211
8.7.5 HP 分析算法.....	211
8.8 浅层句法分析.....	214
8.8.1 概述.....	214
8.8.2 基本名词短语的定义.....	215
8.8.3 基于SVM的base NP 识别方法.....	216

8.8.4 基于 WINNOWN 的 base NP 识别方法·····	217
8.8.5 基于 CRF 的 base NP 识别方法·····	219
8.9 依存句法理论简介·····	220
8.10 依存句法分析·····	223
8.10.1 概述·····	223
8.10.2 生存式依存分析方法·····	224
8.10.3 判别式依存分析方法·····	226
8.10.4 确定性依存分析方法·····	228
8.10.5 其他相关研究·····	231
8.10.6 基于序列标注的分层式依存分析方法·····	233
8.11 依存分析器性能评价·····	235
8.11.1 评价指标·····	235
8.11.2 依存分析性能比较·····	236
8.12 短语结构与依存结构之间的联系·····	240
第9章 语义分析 ·····	244
9.1 语义消歧概述 ·····	244
9.2 有监督的词义消歧方法 ·····	245
9.2.1 基于互信息的消歧方法 ·····	245
9.2.2 基于贝叶斯分类器的消歧方法 ·····	247
9.2.3 基于最大熵的消歧方法 ·····	248
9.3 基于词典的词义消歧方法 ·····	249
9.3.1 基于词典语义定义的消歧方法 ·····	249
9.3.2 基于义类辞典的消歧方法 ·····	250
9.3.3 基于双语词典的消歧方法 ·····	250
9.3.4 Yarowsky 算法及其相关研究 ·····	251
9.4 无监督的词义消歧方法 ·····	252
9.5 词义消歧系统评价 ·····	254
9.6 语义角色标注概述 ·····	255
9.7 语义角色标注基本方法 ·····	257
9.7.1 自动语义角色标注的基本流程 ·····	257
9.7.2 基于短语结构树的语义角色标注方法·····	257
9.7.3 基于依存关系树的语义角色标注方法·····	259
9.7.4 基于语块的语义角色标注方法·····	261
9.7.5 语义角色标注的融合方法·····	262
9.8 语义角色标注的领域适应性问题·····	264
9.9 双语联合语义角色标注方法 ·····	267
9.9.1 基本思路·····	267

9.9.2 系统实现·····	269
9.9.3 实验·····	272
第10章 篇章分析·····	276
10.1 基本概念·····	276
10.2 基本理论·····	277
10.2.1 言语行为理论·····	278
10.2.2 中心理论·····	279
10.2.3 修辞结构理论·····	281
10.2.4 脉络理论·····	283
10.2.5 篇章表示理论·····	284
10.3 篇章衔接性研究·····	286
10.3.1 基于指代消解的衔接性相关研究·····	286
10.3.2 基于词汇衔接的衔接性相关研究·····	289
10.4 基于 HMM 的词对位模型·····	290
10.4.1 基于信息性的连贯性相关研究·····	290
10.4.2 基于意图性的连贯性相关研究·····	292
10.5 篇章标注语料库·····	293
10.6 关于汉语篇章分析·····	294
第11章 统计机器翻译·····	297
11.1 机器翻译概述·····	298
11.1.1 机器翻译的发展·····	298
11.1.2 机器翻译方法·····	298
11.1.3 机器翻译研究现状·····	300
11.2 基于噪声信道模型的统计机器翻译原理·····	301
11.3 IBM 的五个翻译模型·····	304
11.3.1 模型 1·····	304
11.3.2 模型 2·····	307
11.3.3 模型 3·····	308
11.3.4 模型 4·····	313
11.3.5 模型 5·····	315
11.4 基于 HMM 的词对位模型·····	317
11.5 基于短语的翻译模型·····	319
11.5.1 模型演变·····	319
11.5.2 短语对抽取方法·····	321
11.6 基于柱搜索的解码算法·····	325
11.7 基于最大熵的翻译框架·····	329
11.7.1 模型介绍·····	329

11.7.2 对位模型与最大近似	331
11.7.3 对位模板	332
11.7.4 特征函数	332
11.7.5 参数训练	333
11.8 基于层次短语的统计翻译模型	333
11.8.1 概述	333
11.8.2 模型描述	335
11.8.3 参数训练	336
11.8.4 解码方法	337
11.9 树翻译模型	339
11.9.1 树到树的翻译模型	339
11.9.2 树到串的翻译模型	342
11.9.3 串到树的翻译模型	345
11.10 树模型的相关改进	349
11.10.1 源语言句法增强的串到树的翻译模型	349
11.10.2 基于无监督树结构的翻译模型	351
11.11 句法模型解码算法	354
11.12 基于谓词论元结构转换的翻译模型	355
11.13 各种翻译模型的分析	358
11.14 集外词翻译	361
11.14.1 数字和时间表示的识别与翻译	362
11.14.2 命名实体翻译	363
11.14.3 普通集外词的翻译	370
11.15 统计翻译系统实现	371
11.16 系统融合	374
11.16.1 句子级系统融合	374
11.16.2 短语级系统融合	375
11.16.3 词汇级系统融合	376
11.16.4 构建混淆网络的词对齐方法	379
11.17 译文质量评估方法	383
11.17.1 概述	383
11.17.2 技术指标	384
11.17.3 相关评测	392
11.17.4 有关自动评测方法的评测	396
第12章 语音翻译	399
12.1 语音翻译基本原理和特点	399
12.1.1 语音翻译基本原理	399
12.1.2 语音翻译的特点	400

12.2 语音翻译研究现状	401
12.3 C-STAR、A-STAR 和 U-STAR	404
11.3.1 C-STAR 概况	404
11.3.2 A-STAR 和 U-STAR	405
12.4 系统与项目介绍	406
12.5 口语翻译方法	411
12.5.1 基于对话行为分析的口语翻译方法	412
12.5.2 基于句子类型的口语翻译方法	413
第13章 文本分类与情感分类	416
13.1 文本分类概述	416
13.2 文本表示	417
13.3 文本特征选择方法	419
13.3.1 基于文档频率的特征提取法	419
13.3.2 信息增益法	420
13.3.3 χ^2 统计量	420
13.3.4 互信息法	421
13.4 特征权重计算方法	422
13.5 分类器设计	424
13.5.1 朴素贝叶斯分类器	424
13.5.2 基于支持向量机的分类器	425
13.5.3 k -最近邻法	426
13.5.4 基于神经网络的分类器	426
13.5.5 线性最小平方拟合法	426
13.5.6 决策树分类器	427
13.5.7 模糊分类器	427
13.5.8 Rocchio 分类器	427
13.5.9 基于投票的分类方法	428
13.6 文本分类性能评测	428
13.6.1 评测指标	428
13.6.2 相关评测	430
13.7 情感分类	431
第14章 信息检索与问答系统	434
14.1 信息检索概要	434
14.1.1 背景概述	434
14.1.2 基本方法和模型	435
14.1.3 倒排索引	439
14.1.4 文档排序	440

14.2 隐含语义标引模型	440
14.2.1 隐含语义标引模型	440
14.2.2 概率隐含语义标引模型	441
14.2.3 弱指导的统计隐含语义标引模型	443
14.3 检索系统评测	445
14.3.1 检索系统评测指标	445
14.3.2 信息检索活动评测	446
14.4 问答系统	448
14.4.1 概述	448
14.4.2 系统构成	450
14.4.3 基本方法	451
14.4.4 QA 系统评测	453
第15章 自动文摘与信息抽取	455
15.1 自动文摘技术概要	455
15.2 多文档摘要	456
15.2.1 问题与方法	456
15.2.2 文摘评测	458
15.3 信息抽取	460
15.3.1 概述	460
15.3.2 传统的信息抽取技术	461
15.3.3 开放式信息抽取	463
15.4 情感信息抽取	467
15.5 情感分析技术评测	468
第16章 口语信息处理与人机对话系统	471
16.1 汉语口语现象分析	471
16.1.1 概述	471
16.1.2 口语语言现象分析	472
16.1.3 冗余现象分析	474
16.1.4 重复现象分析	475
16.2 口语句子情感信息分析	476
16.2.1 情感词汇分类	476
16.2.2 口语句子情感信息分析	477
16.3 面向中间表示的口语解析方法	479
16.3.1 概述	479
16.3.2 中间表示格式	480
16.3.3 基于规则和 HMM 的统计解析方法	481
16.3.4 基于语义决策树的口语解析方法	486
16.4 基于 MDP 的对话行为识别	487

16.5 基于中间表示的口语生成方法	488
16.5.1 基本思路	488
16.5.2 微观规划器	489
16.5.3 表层生成器	490
16.6 人机对话系统	491
16.6.1 系统组成	491
16.6.2 相关研究	492
参考文献	495
自然语言处理及其相关领域的国际会议	551
名词术语索引	553