



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、隐马尔可夫模型、语料库技术、汉语自动分词与词性标注、句法分析、词义消歧、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

序一
序二
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	4
1.2 自然语言处理研究的内容和面临的困难	4
1.2.1 自然语言处理研究的内容	4
1.2.2 自然语言处理涉及的几个层次	5
1.2.3 自然语言处理面临的困难	6
1.3 自然语言处理的基本方法及其发展	8
1.3.1 自然语言处理的基本方法	8
1.3.2 自然语言处理的发展	9
1.4 自然语言处理研究现状	12
第2章 预备知识	14
2.1 概率论基本概念	14
2.1.1 概率	14
2.1.2 最大似然估计	14
2.1.3 条件概率	15
2.1.4 贝叶斯法则	15
2.1.5 随机变量	16

2.1.6 二项式分布	17
2.1.7 联合概率分布和条件概率分布	17
2.1.8 贝叶斯决策理论	17
2.1.9 期望和方差	18
2.2 信息论基本概念	19
2.2.1 熵	19
2.2.2 联合熵和条件熵	19
2.2.3 互信息	21
2.2.4 相对熵	22
2.2.5 交叉熵	22
2.2.6 困惑度	23
2.2.7 噪声信道模型	25
2.3 支持向量机	25
2.3.1 线性分类	25
2.3.2 线性不可分	26
2.3.3 构造核函数	26
第3章 形式语言与自动机	28
3.1 基本概念	28
3.1.1 图	28
3.1.2 树	28
3.1.3 字符串	29
3.2 形式语言	30
3.2.1 概述	30
3.2.2 形式语法的定义	30
3.2.3 形式语法的类型	31
3.2.4 CFG 识别句子的派生树表示	33
3.3 自动机理论	34
3.3.1 有限自动机	34
3.3.2 正则文法与自动机的关系	36
3.3.3 上下文无关文法与下推自动机	37
3.3.4 图灵机	38
3.3.5 线性界限自动机	39
3.4 自动机在自然语言处理中的应用	40
3.4.1 单词拼写检查	40
3.4.2 单词形态分析	43
3.4.3 词性消歧	44
第4章 语料库与词汇知识库	48
4.1 语料库技术	48
4.1.1 概述	48
4.1.2 语料库语言学的发展	49

4.1.3	语料库的类型	52
4.1.4	典型语料库介绍	54
4.1.5	汉语语料库建设中的问题	60
4.2	词汇知识库	62
4.2.1	WordNet	62
4.2.2	FrameNet	64
4.2.3	EDR	64
4.2.4	知网	66
4.2.5	概念层次网络	70
4.3	语言知识库建设中的本体论	71
第5章	语言模型	74
5.1	n 元语法	74
5.2	语言模型性能评价	77
5.3	数据平滑	77
5.3.1	问题的提出	77
5.3.2	加法平滑方法	78
5.3.3	古德-图灵(Good-Turing)估计法	79
5.3.4	Katz 平滑方法	79
5.3.5	Jelinek-Mercer 平滑方法	81
5.3.6	Witten-Bell 平滑方法	82
5.3.7	绝对减值法	83
5.3.8	Kneser-Ney 平滑方法	84
5.3.9	算法总结	86
5.4	其它平滑方法	87
5.4.1	Church-Gale 平滑方法	87
5.4.2	贝叶斯平滑方法	88
5.4.3	修正的 Kneser-Ney 平滑方法	88
5.5	平滑方法的比较	90
5.6	语言模型自适应方法	90
5.6.1	基于缓存记忆的语言模型	91
5.6.2	基于混合方法的语言模型	92
5.6.3	基于最大熵的语言模型	92
第6章	隐马尔柯夫模型	94
6.1	马尔柯夫模型	94
6.2	隐马尔柯夫模型的构成	96
6.3	前后向算法及参数估计	97
6.3.1	求解观察序列的概率	97
6.3.2	维特比算法	101

6.3.3 HMM 的参数估计	102
第7章 汉语自动分词与词性标注	105
7.1 汉语自动分词中的基本问题	105
7.1.1 汉语分词规范问题	105
7.1.2 歧义切分问题	106
7.1.3 未登录词问题	108
7.2 基本分词方法	109
7.2.1 基于统计语言模型的分词方法	109
7.2.2 N-最短路径方法	111
7.2.3 基于 HMM 的分词方法	114
7.2.4 基于三元统计模型的分词与词性标注一体化方法	115
7.2.5 由字构词的汉语分词方法	117
7.2.6 方法比较	118
7.3 未登录词处理方法概述	120
7.4 基于多特征的命名实体识别模型	122
7.4.1 模型描述	122
7.4.2 词形和词性上下文模型	123
7.4.3 实体模型	124
7.4.4 专家知识	128
7.4.5 模型训练	128
7.4.6 测试结果	129
7.5 词性标注	130
7.5.1 概述	130
7.5.2 基于统计模型的词性标注方法	131
7.5.3 基于规则的词性标注方法	134
7.5.4 统计方法与规则方法相结合的词性标注方法	136
7.5.5 词性标注中的生词处理方法	138
7.6 词性标注的一致性检查与自动校对	139
7.6.1 词性标注一致性检查方法	139
7.6.2 词性标注自动校对方法	141
7.7 汉语分词与词性标注系统评测	143
第8章 句法分析	147
8.1 概述	147
8.1.1 基本概念	147
8.1.2 语法形式化	147
8.1.3. 基本方法	148
8.2 统计句法分析	150
8.2.1 语法驱动的分析方法	151
8.2.2 数据驱动的分析方法	158

8.2.3 其他分析方法	159
8.3 句法分析器评测	160
8.4 汉语句法结构特点	163
8.5 层次化汉语长句结构分析	165
8.5.1 标点符号在句法分析中的作用	165
8.5.2 层次化汉语长句结构分析的思路	166
8.5.3 汉语标点符号的分类	167
8.5.4 句法规则提取方法	168
8.5.5 HP 分析方法	169
8.5.6 实验	171
8.6 浅层句法分析	173
8.6.1 概述	173
8.6.2 Base NP 识别问题	174
8.6.3 基于支持向量机的 Base NP 识别方法	175
8.6.4 基于 WINNOW 的 Base NP 识别方法	177
8.6.5 基于条件随机场的 Base NP 识别方法	179
8.7 依存语法理论与依存句法分析	181
8.7.1 依存语法理论	181
8.7.2 依存句法分析	183
第9章 语义消歧	190
9.1 概述	190
9.2 有监督的词义消歧方法	191
9.2.1 基于互信息的消歧方法	191
9.2.2 基于贝叶斯分类器的消歧方法	193
9.3 基于词典的词义消歧方法	194
9.3.1 基于语义定义的消歧方法	194
9.3.2 基于义类辞典的消歧方法	195
9.3.3 基于双语词典的消歧方法	195
9.3.4 Yarowsky 算法及其相关研究	196
9.4 无监督的词义消歧方法	197
9.5 词义消歧系统评测	199
第10章 统计机器翻译	201
10.1 机器翻译概述	202
10.1.1 机器翻译的发展	202
10.1.2 机器翻译方法	202
10.1.3 机器翻译研究现状	204
10.2 基于噪声信道模型的统计机器翻译原理	205

10.3 IBM 的五个翻译模型	208
10.3.1 模型 1	209
10.3.2 模型 2	212
10.3.3 模型过度	214
10.3.4 模型 3	216
10.3.5 模型 4	220
10.3.6 模型 5	223
10.4 基于 HMM 的词对位模型	225
10.5 基于结构的对位模型	226
10.6 基于反向转换文法的翻译模型	229
10.7 基于有限状态转换机的翻译模型	235
10.7.1 加权的有限状态中心转换机	235
10.7.2 依存转换模型	236
10.7.3 转换算法	238
10.7.4 训练方法	239
10.8 基于句法的翻译模型	242
10.9 基于短语的翻译模型	246
10.9.1 层次化短语对位方法	246
10.9.2 基于短语的联合概率翻译模型	247
10.9.3 基于短语的翻译模型	248
10.9.4 一体化短语分割与对位算法 (ISA)	252
10.9.5 改进的基于 HMM 的短语对获取方法	254
10.10 基于层次短语的统计翻译模型	257
10.10.1 概述	257
10.10.2 模型描述	258
10.10.3 参数训练	260
10.10.4 解码方法	261
10.11 基于语块的翻译模型	262
10.11.1 基于语块的翻译模型结构	263
10.11.2 参数估计	265
10.11.3 解码	266
10.11.4 方法讨论	266
10.12 基于最大熵的翻译模型	267
10.12.1 模型介绍	267
10.12.2 对位模型与最大近似	269
10.12.3 对位模板	270
10.12.4 特征函数	270
10.12.5 参数训练	271
10.13 树到树的翻译模型	272

10.14 树到串的翻译模型	276
10.15 各种翻译模型的分析	279
10.16 解码算法	282
10.14.1 基于栈的解码算法	282
10.14.2 基于 A* 搜索的解码算法	285
10.14.3 贪心爬山解码算法	287
10.14.4 基于动态规划的解码算法	290
10.14.5 Pharaoh 解码器	298
10.14.6 双向搜索算法	302
10.17 统计翻译系统实现	304
10.18 译文质量评估方法	306
10.16.1 概述	306
10.16.2 技术指标	307
10.16.3 评测方法与系统现状	315
10.19 代表系统简介	319
第11章 语音翻译	323
11.1 语音翻译基本原理和特点	323
11.1.1 语音翻译基本原理	323
11.1.2 语音翻译的特点	324
11.2 语音翻译研究现状	325
11.3 C-STAR 组织	329
11.3.1 C-STAR 概况	329
11.3.2 C-STAR 翻译框架	330
11.4 系统与项目介绍	331
第12章 文本分类	340
12.1 概述	340
12.2 文本表示	341
12.3 文本特征选择方法	343
12.3.1 基于文档频率的特征提取法 (DF)	343
12.3.2 信息增益法 (IG)	344
12.3.3 χ^2 统计量 (CHI)	344
12.3.4 互信息法 (MI)	345
12.4 特征权重计算方法	346
12.5 分类器设计	348
12.5.1 朴素贝叶斯分类器	348

12.5.2	基于支持向量机的分类器	349
12.5.3	k -最近邻法	349
12.5.4	基于神经网络的分类器	350
12.5.5	线性最小平方拟合法	350
12.5.6	决策树分类器	350
12.5.7	模糊分类器	351
12.5.8	Rocchio 分类器	351
12.5.9	基于投票的分类方法	352
12.6	文本分类器性能评估方法	352
12.6.1	正确率、召回率和 F-Measure	352
12.6.2	微平均和宏平均	353
第13章	信息检索与问答系统	354
13.1	信息检索概要	354
13.1.1	背景概述	354
13.1.2	基本方法和模型	355
13.1.3	倒排索引	359
13.1.4	文档排序	360
13.2	隐含语义标引模型	360
13.2.1	隐含语义标引模型	360
13.2.2	概率隐含语义标引模型	364
13.2.3	弱指导的统计隐含语义标引模型	366
13.3	检索系统评测与技术现状	368
13.3.1	检索系统评测指标	368
13.3.2	信息检索技术现状	369
13.4	搜索引擎技术	370
13.4.1	搜索引擎核心技术的演进	371
13.4.2	搜索引擎的通用化与专业化	372
13.5	问答系统	373
13.5.1	基本概念	373
13.5.2	系统构成	374
13.5.3	基本方法	375
13.5.4	系统评测与技术现状	376
第14章	自动文摘与信息抽取	379
14.1	自动文摘技术概要	379
14.2	多文档摘要	380
14.2.1	基本方法和问题	380
14.2.2	文摘评测	381
14.2.3	代表系统	383
14.3	信息抽取	386

14.3.1 概述	386
14.3.2 信息抽取技术的发展及其研究现状	386
14.3.3 信息抽取系统基本构成与关键技术	388
第15章 口语信息处理与人机对话系统	390
15.1 汉语口语现象分析	390
15.1.1 概述	390
15.1.2 口语语言现象分析	391
15.1.3 冗余现象分析	393
15.1.4 重复现象分析	394
15.2 口语句子情感信息分析	395
15.2.1 情感词汇分类	395
15.2.2 口语句子情感信息分析	396
15.3 面向中间表示的口语解析方法	398
15.3.1 概述	398
15.3.2 中间表示格式	399
15.3.3 基于规则和 HMM 的统计解析方法	400
15.3.4 基于语义决策树的口语解析方法	405
15.4 基于中间表示的口语生成方法	410
15.4.1 基本思路	410
15.4.2 微观规划器	411
15.4.3 表层生成器	412
15.5 人机对话系统	413
15.5.1 系统组成	413
15.5.2 相关研究	414
附录 A: 项目作业	417
附录 B: 英汉术语对照表	419
参考文献	434