1. 统计自然语言处理;
2. 序言; 序一; 序二; 第二版前言; 第一版前言;
3. Chapter 1: 绪论;
4. Node 1.1, 基本概念;
5. List 1.1.1, 语言学和语音学;
6. List 1.1.2, 自然语言处理;
7. List 1.1.3, 关于理解的标准;
8. 图灵测试; (2020-10-2)
9. Node 1.2, 自然语言处理研究的内容和面临的困难;
10. List 1.2.1, 自然语言处理研究的内容;
11. 机器翻译(machine translation), 自动文摘(automatic abstracting), 信息检索(information retrieval), 文档分类(document categorization), 问答系统(question-answering system); 信息过滤(information filtering); 信息抽取(information extraction); 文本挖掘(text mining); 舆情分析; 隐喻计算; 文字编辑和自动校对; 作文自动评分; 光读字符识别; 语音识别(speech recognition); 文语转换; 说话人识别; (2020-10-2)
12. List 1.2.2, 自然语言处理的几个层次;
13. 形态学(morphology); 语法学(syntax); 语义学(semantics); 语用学(pragmatics);
14. List 1.2.3, 自然语言处理面临的困难;
15. 歧义消解(disambiguation); (2020-10-2)
16. Node 1.3, 自然语言处理的基本方法及其发展;
17. List 1.3.1, 自然语言处理的基本方法;
18. 理性主义方法主张建立符号处理系统, 由人工整理和编写初始的语言知识表示体系(通常为规则), 构造相应的推理程序; 经验主义方法主张通过建立特定的数学模型来学习复杂的、广泛的语言结构，然后利用统计学、模式识别和机器学习等方法来训练模型的参数，以扩大语言使用的规模; 经过筛选、加工和标注等处理的大批量语料构成的数据库叫做语料库(corpus base); (2020-10-2)
19. List 1.3.2, 自然语言处理的发展;
20. 噪声信道模型; 隐马尔可夫模型; ①对于句法分析，基于单一标记的短语结构规则是不充分的；②短语结构规则在真实文本中的分布呈现严重的扭曲; (2020-10-2)
21. Node 1.4, 自然语言处理的研究现状;
22. 已开发完成一批颇具影响的语言资源库; 尚未建立起一套完整、系统的理论框架体系; (2020-10-2)
23. Node 1.5, 本书的内容安排; (2019-7-22)
24. Chapter 2: 预备知识;
25. Node 2.1, 概率论基本概念;
26. List 2.1.1, 概率;
27. List 2.1.2, 最大似然估计;
28. 通常用相对频率作为概率的估计值, 这种估计概率值的方法称为最大似然估计(maximum likelihood estimation); (2020-10-2)
29. List 2.1.3, 条件概率;
30. List 2.1.4, 贝叶斯法则;
31. List 2.1.5, 随机变量;
32. List 2.1.6, 二项式分布;
33. List 2.1.7, 联合概率分布和条件概率分布;
34. List 2.1.8, 贝叶斯决策理论;
35. List 2.1.9, 期望和方差;
36. Node 2.2, 信息论基本概念;
37. List 2.2.1, 熵;
38. 使熵值最大的概率分布最真实地反映了事件的分布情况，因为熵定义了随机变量的不确定性，当熵最大时，随机变量最不确定，最难准确地预测其行为; (2020-10-2)
39. List 2.2.2, 联合熵和条件熵;
40. 联合熵实际上就是描述一对随机变量平均所需要的信息量; 熵的连锁规则; (2020-10-2)
41. List 2.2.3, 互信息;
42. 反映的是在知道了Y的值以后X的不确定性的减少量, 可以理解为Y的值透露了多少关于X的信息量; 互信息体现了两变量之间的依赖程度; (2020-10-2)
43. List 2.2.4, 相对熵;
44. 相对熵是衡量相同事件空间里两个概率分布相对差距的测度; (2020-10-2)
45. List 2.2.5, 交叉熵;
46. 交叉熵的概念就是用来衡量估计模型与真实概率分布之间差异情况的; 模型的交叉熵越小, 模型的表现越好; (2020-10-2)
47. List 2.2.6, 困惑度;
48. 语言模型设计的任务就是寻找困惑度最小的模型，使其最接近真实语言的情况; (2020-10-2)
49. List 2.2.7, 噪声信道模型;
50. 一种自然语言的句子可以视为已编码的符号序列，但需要进行解码，使观察到的输出序列更接近于输入; (2020-10-2)
51. Node 2.3, 支持向量机;
52. 在高维特征空间使用线性函数假设空间的学习系统，在分类方面具有良好的性能; (2020-10-2)
53. List 2.3.1, 线性分类;
54. List 2.3.2, 线性不可分;
55. 对于非线性问题，可以把样本x映射到某个高维特征空间，在高维特征空间中使用线性学习器; (2020-10-2)
56. List 2.3.3, 构造核函数;
57. Chapter 3: 形式语言和自动机;
58. Node 3.1, 基本概念;
59. List 3.1.1, 图;
60. 无向图; 有向图; 连通图; 回路; (2020-10-2)
61. List 3.1.2, 树;
62. List 3.1.3, 字符串;
63. 字符串; 字符串连接; 符号串集合的乘积; 闭包运算; (2019-7-21)
64. Node 3.2, 形式语言;
65. List 3.2.1, 概述;
66. 穷举法; 文法; 自动机法; (2020-10-3)
67. List 3.2.2, 形式语法的定义;
68. 形式语法; 推导; 句子; (2020-10-3)
69. List 3.2.3, 形式语法的类型;
70. 正则语法; 上下文无关文法; 上下文有关文法; 无约束文法; (2020-10-3)
71. List 3.2.4, CFG识别句子的派生树表示;
72. 二义性文法; (2019-7-23)
73. Node 3.3, 自动机理论;
74. List 3.3.1, 有限自动机;
75. 确定性有限自动机; DFA接受的语言; 不确定的有限自动机; NFA接受的语言; (2020-10-3)
76. List 3.3.2, 正则文法和自动机的关系;
77. List 3.3.3, 上下文无关文法和下推自动机;
78. List 3.3.4, 图灵机;
79. 线性界限自动机; (2019-7-23)
80. Node 3.4, 自动机在自然语言处理中的应用;
81. List 3.4.1, 单词拼写检查;
82. List 3.4.2, 单词形态分析;
83. heavy: heavier, heaviest; (2020-10-3)
84. List 3.4.3, 词性消歧;
85. 词性标注的方法很多，有限状态转换机方法是其中的一种; (2020-10-3)
86. Chapter 4: 语料库和语言知识库;
87. Node 4.1, 语料库技术;
88. List 4.1.1, 概述;
89. 语料库语言学研究的基础是大规模真实语料; (2020-10-3)
90. List 4.1.2, 语料库语言学的发展;
91. List 4.1.3, 语料库的类型;
92. 平衡语料库和平行语料库; 通用语料库和专用语料库; 共时语料库和历时语料库; 生语料和标注语料库; (2020-10-7)
93. List 4.1.4, 汉语语料库建设中的问题;
94. 语料库建设的规范问题; 产权保护和国家语料库建设问题; (2020-10-7)
95. List 4.1.5, 典型语料库介绍;
96. LDC中文树库; 命题库, 名词化树库和语篇库; 布拉格依存树库; BTEC口语语料; 现代汉语口语语料库; 台湾中研院语料库; (2019-7-23)
97. Node 4.2, 语言知识库;
98. List 4.2.1, WordNet;
99. List 4.2.2, FrameNet;
100. List 4.2.3, EDR;
101. List 4.2.4, 北京大学综合型语言知识库;
102. List 4.2.5, 知网;
103. List 4.2.6, 概念层次网络; (2019-7-24)
104. Node 4.3, 语言知识库和本体论;
105. Chapter 5, 语言模型;
106. Node 5.1, n元语法;
107. 可以将历史w1w2…wi-1按照某个法则映射到等价类E(w1w2…wi-1), 而等价类的数目远远小于不同历史的数目; 将两个历史wi-n＋2…wi-1wi和vkn＋2…vk-1vk映射到同一个等价类, 当且仅当这两个历史最近的n-1（1≤n≤l）个词相同; n元文法; 二阶文法是一阶马尔可夫链; (2020-10-7)
108. Node 5.2, 语言模型性能评价;
109. 交叉熵和困惑度越小越好，这是我们评估一个语言模型的基本准则; (2020-10-7)
110. Node 5.3, 数据平滑;
111. List 5.3.1, 问题的提出;
112. 平滑(smoothing)技术就是用来解决这类零概率问题的; “平滑”处理的基本思想是“劫富济贫”，即提高低概率（如零概率），降低高概率，尽量使概率分布趋于均匀; (2020-10-7)
113. List 5.3.2, 加法平滑方法;
114. List 5.3.3, 古德图灵估计法;
115. List 5.3.4, Katz平滑方法;
116. List 5.3.5, Jelinek-Mercer平滑方法;
117. List 5.3.6, Witten-Bell平滑方法;
118. List 5.3.7, 绝对减值法;
119. List 5.3.8, Kneser-Ney平滑方法;
120. List 5.3.6, 算法总结; (2019-7-24)
121. Node 5.4, 其他平滑方法;
122. List 5.4.1, Church-Gale平滑方法;
123. List 5.4.2, 贝叶斯平滑方法;
124. List 5.4.3, 修正的Kneser-Ney平滑方法; (2019-7-24)
125. Node 5.5, 平滑方法的比较; (2019-7-24)
126. Node 5.6, 语言模型自适应方法;
127. List 5.6.1, 基于缓存的语言模型;
128. List 5.6.2, 基于混合方法的语言模型;
129. List 5.6.3, 基于最大熵的语言模型; (2019-7-24)
130. Chapter 6, 概率图模型;
131. Node 6.1, 概述; (2019-7-24)
132. Node 6.2, 贝叶斯网络; (2019-7-24)
133. Node 6.3, 马尔可夫模型; (2019-7-25)
134. 一个马尔可夫链的状态序列的概率可以通过计算形成该状态序列的所有状态之间转移弧上的概率乘积而得出;
135. Node 6.4, 隐马尔可夫模型;
136. List 6.4.1, 求解观察序列的概率;
137. 前向变量; 后向变量; backward procedure; 维特比算法; 维特比变量; Viterbi algorithm; HMM的参数估计; forward-backward algorithm; (2019-7-25)
138. 6.5, 层次化的隐马尔可夫模型; (2019-7-25)
139. 6.6, 马尔可夫网络; (2019-7-25)
140. 6.7, 最大熵模型; 最大熵原理; 最大熵模型的参数训练; (2019-7-25)
141. 6.8, 最大熵马尔可夫模型; (2019-7-25)
142. 6.9, 条件随机场; (2019-7-25)
143. Chapter 7, 自动分词、 命名实体识别与词性标注;
144. 7.1, 汉语自动分词中的基本问题; 汉语分词规范问题; 歧义切分问题; 交集型切分歧义; 多义组合型切分歧义; 未登录词问题; (2019-7-25)
145. 7.2, 汉语分词方法; N-最短路径方法; 基于词的n元语法模型的分词方法; 由字构词的汉语分词方法; 基于词感知机算法的汉语分词方法; 基于字的生成式模型和区分式模型相结合的汉语分词方法; 其他分词方法; 分词方法比较; (2019-7-27)
146. 7.3, 命名实体识别; 方法概述; 基于CRF的命名实体识别方法; 基于多特征的命名实体识别方法; 模型描述; 词形模型; 词性模型; 混合模型; 词形和词性上下文模型; 实体模型; 人名实体模型; 地名和机构名实体模型; 单字地名实体模型; 简称机构实体模型; 专家知识; 人名识别的专家知识; 地名识别的专家知识; 机构名识别的专家知识; 模型训练; 测试结果; (2019-7-27)
147. 7.4, 维吾尔语人名识别方法研究; 维语人名构成特点; 维吾尔语人名类型; 基于CRF的维语人名识别方法; (2019-7-27)
148. 7.5, 词性标注; 概述; 基于统计模型的词性标注方法; 基于规则的词性标注方法; 统计方法和规则方法相结合的词性标注方法; 词性标注中的生词处理方法; (2019-7-27)
149. 7.6, 词性标注的一致性检查与自动校对; 词性标注的一致性检查方法; 词性标注自动校对方法; (2019-7-27)
150. 7.7, 关于技术评测; (2019-7-30)
151. Chapter 8, 句法分析;
152. 8.1, 句法结构分析概述; 基本概念; 语法形式化; 基本方法; 常用的短语结构分析器; (2019-7-30)
153. 8.2, 基于PCFG的基本分析方法; 位置不变性; 上下文无关性; 祖先无关性; 面向PCFG的内向外向算法; 内向算法; 外向算法; 选择句子的最佳结构; PCFG的概率参数估计; 分析实例; (2019-7-30)
154. 8.3, 词汇化的短语结构分析器; (2019-7-30)
155. 8.4, 非词汇化句法分析器; (2019-7-30)
156. 8.5, 其他相关研究; PCFG方法的改进; 数据驱动的分析方法; (2019-7-30)
157. 8.6, 短语结构分析器性能评价; 评价指标; 标记正确率; 标记召回率; 交叉括号数; 短语结构分析器性能比较; (2019-7-30)
158. 8.7, 层次化汉语长句结构分析; 标点符号在句法分析中的作用; 层次化汉语长句结构分析的思路; 汉语标点符号的分类; 句法规则提取方法; HP分析算法; 长句分割; 第一级分析; 子树合并; 第二级分析; (2019-8-6)
159. 8.8, 浅层句法分析; 概述; 基本名词短语的定义; 基于SVM的base NP识别方法; 基于window的base NP识别方法; 训练阶段; 测试阶段; 基于CRF的base NP识别方法; (2019-8-6)
160. 8.9, 依存语法理论简介; (2019-8-6)
161. 8.10, 依存语法分析; 概述; 生成式依存分析方法; 二元亲和词汇模型; 选择偏好模型; 递归生成模型; 判别式依存分析方法; 确定性依存分析方法; Yamada分析方法; Nivre分析方法; 其他相关研究; 基于序列标注的分层式依存分析方法; (2019-8-6)
162. 8.11, 依存分析器性能评价; 评价指标; 无标记依存正确率; 带标记依存正确率; 依存正确率; 根正确率; 完全匹配率; 依存分析性能比较; 英语依存分析性能比较; 汉语依存分析性能比较; (2019-8-6)
163. 8.12, 短语结构与依存结构之间的关系; (2019-8-12)
164. Chapter 9, 语义分析;
165. 9.1, 词义消歧概述; (2019-8-12)
166. 9.2, 有监督的词义消歧方法; 基于互信息的消歧方法; 基于贝叶斯分类器的消歧方法; 基于最大熵的词义消歧方法; (2019-8-13)
167. 9.3, 基于词典的词义消歧方法; 基于词典语义定义的消歧方法; 基于义类词典的消歧方法; 基于双语词典的消歧方法; Yarowsky算法及其相关研究; (2019-8-13)
168. 9.4, 无监督的词义消歧方法; (2019-8-13)
169. 9.5, 词义消歧系统评价; (2019-8-13)
170. 9.6, 语义角色标注概述; (2019-8-13)
171. 9.7, 语义角色标注基本方法; 自动语义角色标注的基本流程; 基于短语结构树的语义角色标注方法; 基于依存关系树的语义角色标注方法; 基于语块的语义角色标注方法; 语义角色标注的融合方法; (2019-8-13)
172. 9.8, 语义角色标注的领域适应性问题; (2019-8-13)
173. 9.9, 双语联合语义角色标注方法; 基本思路; 系统实现; 源语言部分; 目标语言部分; 论元对齐概率模型; 实验; 实验数据; 参数选取; 实验结果; (2019-8-20)
174. Chapter 10, 篇章分析;
175. 10.1, 基本概念; (2019-8-20)
176. 10.2, 基本理论; 言语行为理论; 声明类; 指令类; 许诺类; 表达类; 宣告类; 中心理论; 修辞结构理论; 脉络理论; 篇章表示理论; (2019-8-20)
177. 10.3, 篇章衔接性研究; 基于指代消解的衔接性相关研究; 基于词汇衔接的衔接性相关研究; (2019-8-20)
178. 10.4, 篇章连贯性研究; 基于信息性的连贯性相关研究; 详述关系; 并列关系; 对比关系; 基于意图性的连贯性相关研究; (2019-8-20)
179. 10.5, 篇章标注语料库; (2019-8-20)
180. 10.6, 关于汉语篇章分析; (2019-8-20)
181. Chapter 11, 统计机器翻译;
182. 11.1, 机器翻译概述; 机器翻译的发展; 机器翻译方法; memory-based machine translation; example-based machine translation; statistical machine translation; neural network machine translation; 机器翻译研究现状; (2019-8-20)
183. 11.2, 基于噪声信道模型的统计机器翻译原理; (2019-8-20)
184. 11.3, IBM的5个翻译模型; 模型1; 模型2; 模型3; 模型4; 模型5; (2019-8-21)
185. 11.4, 基于HMM的词对位模型; (2019-8-21)
186. 11.5, 基于短语的翻译模型; 模型演变; 短语对抽取方法; 基于GIZA++的短语对抽取方法; (2019-8-21)
187. 11.6, 基于柱搜索的解码算法; 翻译选项; 核心算法; 重组假设; 柱搜索; (2019-8-21)
188. 11.7, 基于最大熵的翻译框架; 模型介绍; 对位模型与最大近似; 对位模板; 特征函数; 参数训练; (2019-8-21)
189. 11.8, 基于层次短语的翻译模型; 概述; 模型描述; 参数训练; 解码方法; (2019-8-21)
190. 11.9, 树翻译模型; 树到树的翻译模型; 树到串的翻译模型; 串到树的翻译模型; (2019-8-22)
191. 11.10, 树模型的相关改进; 源语言句法增强的串到树翻译模型; 基于无监督树结构的翻译模型; (2019-8-22)
192. 11.11, 句法模型解码算法; (2019-8-22)
193. 11.12, 基于谓词论元结构转换的翻译模型; (2019-8-22)
194. 11.13, 各种翻译模型的分析; (2019-8-22)
195. 11.14, 集外词翻译; 数字和时间表示的识别与翻译; 命名实体翻译; 概述; 基于结构的组织机构名翻译; 命名实体识别与对齐的联合方法; 普通集外词的翻译; (2019-8-22)
196. 11.15, 统计翻译系统实现; 语料准备和预处理; 工具准备; 模块构建; 系统调试; (2019-8-22)
197. 11.16, 系统融合; 句子级系统融合; 最小贝叶斯风险解码; 通用线性模型; 短语级系统融合; 词汇级系统融合; 构建混淆网络; 特征和特征权重的优化; 构建混淆网络的词对齐方法; 基于编辑距离的词对齐方法; 基于语料库单语句对的词对齐方法; 基于语言学知识的单语句对的词对齐方法; (2019-8-26)
198. 11.17, 译文质量评估方法; 概述; 技术指标; 主观评测指标; 自动评测指标; BLEU方法; NIST方法; mWER方法; mPER方法; GTM方法; METEOR方法; 相关评测; IWSLT口语翻译评测; NIST机器翻译评测; WMT机器翻译评测; CWMT机器翻译评测; 有关自动评测方法的评测; (2019-8-26)
199. Chapter 12, 语音翻译;
200. 12.1, 语音翻译的基本原理和特点; 语音翻译的基本原理; 语音翻译的特点; (2019-8-26)
201. 12.2, 语音翻译的研究现状; (2019-8-26)
202. 12.3, C-STAR, A-STAR和U-STAR; C-STAR概况; A-STAR和U-STAR; (2019-8-26)
203. 12.4, 系统与项目介绍; 短语翻译器Phraselator; Verbmobil口语翻译系统; ATR-MATRIX口语翻译系统; JANUS口语翻译系统; CASIA口语翻译系统; NESPOLE!语音翻译研究项目; (2019-8-30)
204. 12.5, 口语翻译方法; 基于对话行为分析的口语翻译方法; 基于句子类型的口语翻译方法; (2019-8-30)
205. Chapter 13, 文本分类与情感分类;
206. 13.1, 文本分类概述; (2019-8-30)
207. 13.2, 文本表示; 向量空间模型VSM; 向量的相似性度量; (2019-8-30)
208. 13.3, 文本特征选择方法; 基于文档频率的特征提取法; 信息增益法; X^2统计量; 互信息法; (2019-8-30)
209. 13.4, 特征权重计算方法; 布尔权重; 绝对词频; 倒排文档频度; TF-IDF; TFC; ITC; 熵权重; TF-IWF; (2019-8-30)
210. 13.5, 分类器设计; 朴素贝叶斯分类器; 支持向量机的分类器; k-最近邻法; 基于神经网络的分类器; 线性最小平方拟合法; 决策树分类器; 模糊分类器; Rocchio分类器; 基于投票的分类方法; bagging算法; boosting算法; (2019-8-30)
211. 13.6, 文本分类性能评测; 评测指标; 正确率, 召回率, F测度值; 微平均和宏平均; 相关评测; (2019-8-31)
212. 13.7, 情感分类; 按机器学习方法分类; 按研究问题分类; (2019-8-31)
213. Chapter 14, 信息检索与问答系统;
214. 14.1, 信息检索概要; 背景概述; indexing; relevance or similarity; 基本方法和模型; 布尔模型; 向量空间模型; 点积法; 余弦法; Dice方法; Jaccard方法; 概率模型; 语言模型; 倒排索引; 文档排序; (2019-8-31)
215. 14.2, 隐含语义标引模型; 隐含语义标引模型; 概率隐含语义标引模型; 弱指导的统计隐含语义标引模型; (2019-9-2)
216. 14.3, 检索系统评测; 检索系统评测指标; 准确率, 召回率和F测度值; P@10; R-precision; 最差x%个查询主题的准确率; 信息检索评测活动; TREC评测; NTCIR评测; 863评测; (2019-9-2)
217. 14.4, 问答系统; 概述; 系统构成; 基本方法; QA系统评测; List问题; Definition问题; Context问题; Passage问题; Other问题; (2019-9-2)
218. Chapter 15, 自动文摘与信息抽取;
219. 15,1, 自动文摘技术摘要; (2019-9-2)
220. 15.2, 多文档摘要; 问题与方法; 文摘评测; (2019-9-2)
221. 15.3, 信息抽取; 概述; 传统的信息抽取技术; 开放式信息抽取; 开放式实体抽取; 实体消歧; 实体聚类消歧法; 实体链接消歧法; 开放式实体关系抽取; (2019-9-2)
222. 15.4, 情感信息抽取; (2019-9-2)
223. 15.5, 情感分析技术评测; COAE评测; CCF TCCI评测; TAC评测; 实体链接entity-linking; 槽填充slot-filling; 冷启动知识库扩展cold start knowledge base population; (2019-9-2)
224. Chapter 16, 口语信息处理与人机对话系统;
225. 16.1, 汉语口语现象分析; 概述; 口语语言现象分析; 词长分布; 对话语句的长度分布; 词类分布统计; 非规范语言现象的出现几率; 冗余现象分析; 重复现象分析; (2019-9-2)
226. 16.2, 口语句子情感信息分析; 情感词汇分类; 口语句子情感信息分析; (2019-9-4)
227. 16.3, 面向中间表示的口语解析方法; 概述; 中间表示格式; 基于规则和HMM的统计解析方法; 词汇分类; 词义语块分析; 统计解析过程; 组块解释方法; IF的生成; 基于语义决策树的口语解析方法; (2019-9-4)
228. 16.4, 基于MDP的对话行为识别; (2019-9-4)
229. 16.5, 基于中间表示的口语生成方法; 基本思路; 微观规划器; 表层生成器; (2019-9-4)
230. 16.6, 人机对话系统; 系统组成; 语音识别器; 语言解析器; 问题求解模块; 语言生成器; 对话管理模块; 语音合成器; 相关研究; (2019-9-4)
231. -