

面向机器学习的自然语言标注



[面向机器学习的自然语言标注_下载链接1](#)

著者:[美] 普斯特若夫斯基 (James Pustejovsky) ,

出版者:机械工业出版社

出版时间:2017-2-1

装帧:平装

isbn:9787111555155

自然语言理解是人工智能的一个重要分支，主要研究如何利用计算机来理解和生成自然语言。本书重点介绍了自然语言理解所涉及的各个方面，包括语法分析、语义分析、概

念分析、语料库语言学、词汇语义驱动、中间语言、WordNet、词汇树邻接文法、链接文法、基于语段的机器翻译方法、内识别与文本过滤、机器翻译的评测等，既有对基础知识的介绍，又有对新研究进展的综述，同时还结合了作者（James Pustejovsky，生成词库理论的创始人）多年的研究成果。本书内容全面、详略得当，结合实例讲解，使读者更易理解。

作者介绍:

James Pustejovsky教授是美国布兰代斯（Brandeis University）大学计算机科学系和Volen国家综合系统中心教授。先后在美国麻省理工学院和马萨诸塞大学获得学士学位和博士学位。Pustejovsky教授主要从事自然语言的理论和计算研究。研究领域包括：计算语言学、词汇语义学、知识表征、话语语义学、时间推理和抽取等。已经出版多部专著。

目录: 前言1

第1章基础知识7

1.1语言标注的重要性7

1.1.1语言学描述的层次8

1.1.2什么是自然语言处理9

1.2语料库语言学简史10

1.2.1什么是语料库13

1.2.2语料库的早期应用15

1.2.3当今的语料库17

1.2.4标注类型18

1.3语言数据和机器学习24

1.3.1分类25

1.3.2聚类25

1.3.3结构化模式归纳26

1.4标注开发循环26

1.4.1现象建模27

1.4.2按照规格说明进行标注30

1.4.3在语料库上训练和测试算法31

1.4.4对结果进行评价32

1.4.5修改模型和算法33

总结34

第2章确定目标与选择数据36

2.1定义目标36

2.1.1目标陈述37

2.1.2提炼目标：信息量与正确性38

2.2背景研究43

2.2.1语言资源44

2.2.2机构与会议44

2.2.3自然语言处理竞赛45

2.3整合数据集46

2.3.1理想的语料库：代表性与平衡性47

2.3.2从因特网上收集数据47

2.3.3从人群中获取数据48

2.4语料库的规模49

2.4.1现有语料库50

2.4.2语料库内部的分布51

总结53

第3章语料库分析54

3.1语料库分析中的基本概率知识	55
3.1.1联合概率分布	56
3.1.2贝叶斯定理	58
3.2计算出出现次数	58
3.2.1齐普夫定律 (Zip's Law)	61
3.2.2n元语法	62
3.3语言模型	63
总结	65
第4章建立模型与规格说明	66
4.1模型和规格说明示例	66
4.1.1电影题材分类	69
4.1.2添加命名实体	70
4.1.3语义角色	71
4.2采用 (或不采用) 现有模型	73
4.2.1创建模型和规格说明: 一般性与特殊性	74
4.2.2使用现有模型和规格说明	76
4.2.3使用没有规格说明的模型	78
4.3各种标准	78
4.3.1ISO标准	78
4.3.2社区驱动型标准	81
4.3.3影响标注的其他标准	81
总结	82
第5章选择并应用标注标准	84
5.1元数据标注: 文档分类	85
5.1.1单标签标注: 电影评论	85
5.1.2多标签标注: 电影题材	87
5.2文本范围标注: 命名实体	90
5.2.1内嵌式标注	90
5.2.2基于词例的分离式标注	92
5.2.3基于字符位置的分离式标注	95
5.3链接范围标注: 语义角色	96
5.4ISO标准和你	97
总结	97
第6章标注与审核	99
6.1标注项目的基本结构	99
6.2标注规格说明与标注指南	101
6.3准备修改	102
6.4准备用于标注的数据	103
6.4.1元数据	103
6.4.2数据预处理	104
6.4.3为标注工作分割文件	104
6.5撰写标注指南	105
6.5.1例1: 单标签标注——电影评论	106
6.5.2例2: 多标签标注——电影题材	108
6.5.3例3: 范围标注——命名实体	111
6.5.4例4: 链接范围标注——语义角色	112
6.6标注人员	114
6.7选择标注环境	116
6.8评价标注结果	117
6.8.1Cohen的Kappa (κ) 算法	118
6.8.2Fleiss的Kappa (κ) 算法	119
6.8.3解释Kappa系数	122
6.8.4在其他上下文中计算 κ 值	123
6.9创建黄金标准 (审核)	125

总结	126
第7章训练：机器学习	129
7.1何谓学习	130
7.2定义学习任务	132
7.3分类算法	133
7.3.1决策树学习	135
7.3.2朴素贝叶斯学习	140
7.3.3最大熵分类器	145
7.3.4其他需要了解的分类器	147
7.4序列归纳算法	148
7.5聚类和无监督学习	150
7.6半监督学习	150
7.7匹配标注与算法	153
总结	154
第8章测试与评价	156
8.1测试算法	157
8.2评价算法	157
8.2.1混淆矩阵	157
8.2.2计算评价得分	159
8.2.3解释评价得分	163
8.3可能影响算法评价的问题	164
8.3.1数据集太小	164
8.3.2算法过于适合开发数据	166
8.3.3标注中的信息过多	166
8.4最后测试得分	167
总结	167
.....	
第9章修改与报告	169
第10章标注：TimeML	179
第11章自动标注：生成TimeML	199
第12章后记：标注的未来发展趋势	217
附录A可利用的语料库与标注规格说明列表	227
附录B软件资源列表	249
附录CMAE用户指南	269
附录DMAI用户指南	276
附录E参考文献	282
• • • • •	(收起)

[面向机器学习的自然语言标注_下载链接1](#)

标签

自然语言处理

机器学习

nlp

计算机

人工智能

Linguistics

NLP

想读的书

评论

@ memect

简直是AI生成的书，仿佛说了很多，却一点用都没有。附录的数据集还可以。

粗疏

数据标注实际上是在定义问题，这才是难点，挺有意思

了解了一些标注的方法和数据集，算是开阔眼界的书籍。

较全面的NLP机器标注。

对语料库构建方法和标注过程均有较详细的介绍，并贯穿以例子帮助读者理解其中概念，可以为初涉自然语言处理领域研究者提供参考。另，Pustejovsky的学生Sauri的博士论文与这本书也有些关系。

这本书主要是供开拓眼界用的，汇集了许多资源列表，讲解了很多背景知识，中英文术语对照做的挺好。对各种标注方法的优劣做了比较，印象较深的是内嵌式标注和分离式标注的优劣。

[面向机器学习的自然语言标注_下载链接1](#)

书评

[面向机器学习的自然语言标注_下载链接1](#)