

# 《自然语言标注》

## 图书基本信息

书名：《自然语言标注》

13位ISBN编号：9787564142812

出版时间：2013-6-1

作者：[美]普斯特若夫斯基 (James Pustejovsky), [美]斯塔布斯 (Amber Stubbs)

页数：324

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《自然语言标注》

## 内容概要

《自然语言标注:用于机器学习(影印版)》可以手把手地指导你一种经验证的标注开发周期——把元语添加到你的训练语料库中来帮助机器学习算法更有效工作的过程。你无需任何编程或者语言学方面的经验就可以上手。《自然语言标注:用于机器学习(影印版)》通过每一步中的详细示例,你将学到“标注开发过程”是如何帮助你建模、标注、训练、测试、评估和修正你的训练语料库。你也将了解到一个实际标注项目的完整演示。

# 《自然语言标注》

## 作者简介

作者：（美国）普斯特若夫斯基（James Pustejovsky）（美国）斯塔布斯（Amber Stubbs）是Brandeis大学的教授，他在该大学的计算机科学系讲解和研究人工智能及计算语言学。刚刚获得了Brandeis大学标注方法论的博士学位。她现在是SUNYAlbany大学的博士后

## 书籍目录

Preface

1. The Basics

The Importance of Language Annotation

The Layers of Linguistic Description

What Is Natural Language Processing ?

A Brief History of Corpus Linguistics

What Is a Corpus ?

Early Use of Corpora

Corpora Today

Kinds of Annotation

Language Data and Machine Learning

Classification

Clustering

Structured Pattern Induction

The Annotation Development Cycle

Model the Phenomenon

Annotate with the Specification

Train and Test the Algorithms over the Corpus

Evaluate the Results

Revise the Model and Algorithms

Summary

2. Defining Your Goal and Dataset

Defining Your Goal

The Statement of Purpose

Refining Your Goal : Informativity Versus Correctness

Background Research

Language Resources

Organizations and Conferences

NLP Challenges

Assembling Your Dataset

The Ideal Corpus : Representative and Balanced

Collecting Data from the Internet

Eliciting Data from People

The Size of Your Corpus

Existing Corpora

Distributions Within Corpora

Summary

3. Corpus Analytics

Basic Probability for Corpus Analytics

Joint Probability Distributions

Bayes Rule

Counting Occurrences

Zipf's Law

N—grams

Language Models

Summary

4. Building Your Model and Specification

Some Example Models and Specs

Film Genre Classification

Adding Named Entities

Semantic Roles

Adopting ( or Not Adopting ) Existing Models

Creating Your Own Model and Specification : Generality Versus Specificity

Using Existing Models and Specifications

Using Models Without Specifications

Different Kinds of Standards

ISO Standards

Community—Driven Standards

Other Standards Affecting Annotation

Summary

5. Applying and Adopting Annotation Standards

Metadata Annotation : Document Classification

Unique Labels : Movie Reviews

Multiple Labels : Film Genres

Text Extent Annotation : Named Entities

Inline Annotation

Stand—off Annotation by Tokens

Stand—off Annotation by Character Location

Linked Extent Annotation : Semantic Roles

ISO Standards and You

Summary

6. Annotation and Adjudication

The Infrastructure of an Annotation Project

Specification Versus Guidelines

Be Prepared to Revise

Preparing Your Data for Annotation

Metadata

Preprocessed Data

Splitting Up the Files for Annotation

Writing the Annotation Guidelines

Example 1 : Single Labels——Movie Reviews

Example 2 : Multiple Labels——Film Genres

Example 3 : Extent Annotations——Named Entities

Example 4 : Link Tags——Semantic Roles

Annotators

Choosing an Annotation Environment

Evaluating the Annotations

Cohen's Kappa ( K )

Fleiss's Kappa ( K )

Interpreting Kappa Coefficients

Calculating K in Other Contexts

Creating the Gold Standard ( Adjudication )

Summary

7. Training : Machine Learning

What Is Learning ?

Defining Our Learning Task

- Classifier Algorithms
- Decision Tree Learning
- Gender Identification
- Naive Bayes Learning
- Maximum Entropy Classifiers
- Other Classifiers to Know About
- Sequence Induction Algorithms
- Clustering and Unsupervised Learning
- Semi—Supervised Learning
- Matching Annotation to Algorithms
- Summary
- 8. Testing and Evaluation
- 9. Revising and Reporting
- 10. Annotation : TimeML
- 11. Automatic Annotation : Generating TimeML
- A. List of Available Corpora and Specifications
- B. List of Software Resources
- C. MAE UserGuide.
- D. MAI UserGuide
- E. Bibliography
- Index

# 《自然语言标注》

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:[www.tushu000.com](http://www.tushu000.com)