<<智能信息处理>>

图书基本信息

书名：<<智能信息处理>>

13位ISBN编号：9787030291356

10位ISBN编号：7030291352

出版时间：2010-10

出版时间：科学出版社

作者：郑家恒　等著

页数：318

前言

从20世纪90年代开始，国际自然语言处理领域发生了一些重大变化，重要特征之一就是转向对大规模真实文本的研究和处理。
以大规模真实文本为基础的语料库研究和知识自动获取受到高度重视。
显然，大规模真实文本的处理是计算语言学今后一个时期的战略目标，建设高质量的大规模语料库是中文信息处理领域的基础性工程。
基于语料库的语言研究是计算语言学的一个重要领域，语料库的建立为语言学的研究提供了丰富的语言现象，为计算语言学学者从加工的语料库中获取语言知识、建立语言模型、研究语言信息处理技术提供了翔实的语言信息数据。
作为研究资源的语料库的价值是通过对语料的加工来体现的，对语料库加工的层次越高，语料库的应用价值就越高。
希望本书的出版能促进语料库加工方法和技术的发展，为基于语料库的相关研究和应用提供支撑。
作者及其课题组从事语言信息处理的教学与研究已有二十多年。
近年来，作者有幸承担了若干国家863计划项目（中文文本自动切词和词性标注软件及其评测技术研究（863-306-03－09-4）、大规模中文文本语料库深加工质量检验技术研究（2001AAll4031））、国家自然科学基金项目（大规模中文文本语料库分词与词性标注一致性检验技术研究（60473139）、基于中文文本的计算机中介通信中欺骗检测研究（60775041））、省部级项目及横向合作项目等。
这些项目的研究成果为本书的编写提供了关键性支持。
多年来，刘开瑛、黄昌宁等诸位学术前辈都为作者的相关研究思路和方法提供了许多指导。
本书编写过程中，山西大学梁吉业、李德玉、李茹、王文剑、王素格等教授为作者提供了多方面的支持。
魏善德、任玉、魏莉、魏丽霞、樊勇、王振宇、刘博、张剑锋、何苑、温艳霞、毋菲等同学也为本书的出版做了许多文字校对方面的工作，谨在此一并表示深深的感谢。

<<智能信息处理>>

内容概要

本书以作者主持的国家项目、省部级项目及合作项目等为依托，以课题组近年来的研究成果为基础，重点介绍语料库深加工中的若干技术和方法，涉及分词、词性标注、句法分析、语义标注以及相关加工中的自动校对和一致性检验技术。
同时，对语料库加工质量的评价技术和语料库的相关应用做了详细介绍。
各章节的顺序展示了语料库加工中由浅人深的发展过程。
　本书可作为计算机、语言学等专业高年级本科生、研究生教材，也可作为自然语言处理和计算语言学研究人员的参考书。

<<智能信息处理>>

书籍目录

章节摘录

插图：关于语料库（corpus）的定义主要有以下几种：（1）McEnery和Wilson指出：“总体来说，多篇文本的集合就是语料库，但在现代语言学中使用语料库这个术语时，更倾向于包含更多的内涵，主要有采样（sampling）收集、有代表性（representativeness）、规模有限（finite size）、机器可读（machine-readable）、标准参考数据（a standard reference）等内涵特征。
”（2）语料库就是某种语言在实际运用中的大量实例集合，这些例子可以是书面文本，也可以是语音形式的文本。
（3）语料库是根据外部原则选择的电子形式的文本或文本片段的集合。
该集合能够代表一种语言，或一种语言的分支，或一种语言的变体，并可作为语言学研究使用的数据源[引。
这里外部原则（external criteria）是指通过文本的交流功能来选择文本的原则。
与外部原则相对的一个概念就是内部原则（internal criteria），具体指按照文本反映的语言细节来选择文本。
在上述的几种定义中，定义（1）使用最多，认为语料库不是简单收集的文本集合，而是通过采样收集，具有代表性，规模大小可以确定，是机器可读的标准数据。
但是Kilgarriff和Grefenstette提出了异议，认为McEnery和Wilson混淆了“什么是语料库”和“什么是好的、适合于某项语言研究的语料库”这两个问题，他们认为语料库就是文本的集合。
然而在具体使用中，有些研究者认为有许多文本的集合并不一定是语料库。
最具有争议的莫过于万维网（WWW）了。
WWW刚出现时，人们因为不了解搜索引擎，也不清楚对WWW如何采样，觉得WWW相当神秘。
因此，文献指出：“WWW不是语料库，因为其维度未知且不断变化，而且WWW最初也不是从语言学角度来设计的。
”

编辑推荐

《智能信息处理:汉语语料库加工技术及应用》：智能科学技术著作丛书

<<智能信息处理>>

版权说明