

文章编号: 1003-0077(2014)01-0125-02

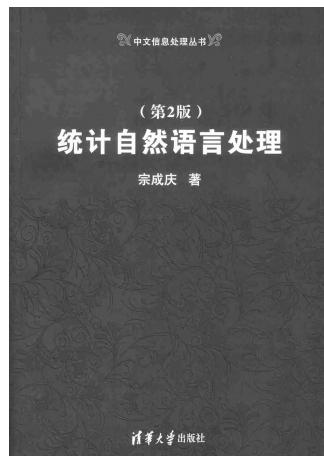
## 内容丰富多彩, 阐述深入浅出 ——评《统计自然语言处理》(第 2 版)

俞士汶

(计算语言学教育部重点实验室(北京大学); 北京大学 计算语言学研究所)

宗成庆博士著《统计自然语言处理》一书自 2008 年问世以来, 已在计算语言学与自然语言处理学界产生了广泛影响, 被很多大学、研究所指定为硕士生、博士生的必读参考书。该书第 1 版很快售罄。参照读者反馈的意见, 作者对该书进行了增删、修改和磨砺, 于 2013 年 8 月推出了第 2 版。在清华大学出版社组织出版的《中文信息处理丛书》中, 这种情况是不多见的。

《统计自然语言处理》(第 2 版)全书共 16 章, 洋洋洒洒 87.5 万余字, 全面介绍了统计自然语言处理的基本概念、理论和方法, 既有词法分析、句法分析、语义分析和篇章分析等核心技术, 也有机器翻译、文本分类、信息抽取、自动文摘、情感计算以及口语信息处理与人机对话等应用系统, 而且对形式语言与自动机、语言模型和概率图模型、语料库与语言知识库这些自然语言处理赖以实现的理论模型和数据资源给予了详细的介绍, 并以概率论、信息论和机器学习等基础知识作为铺垫, 可谓一应俱全, 丰富多彩。综观全书, 脉络清晰, 条理井然, 内容全而不繁, 阐述深入浅出, 不失为一部上乘之作。



除了内容丰富之外, 本书还有以下特点: (1) 从第 7 章至第 9 章, 在介绍自然语言处理的核心方法

时, 从基本概念或提出问题开始, 到基本方法和各种方法的对比和改进, 按照开展研究工作的基本思路逐步展开, 并且利用实验数据对各种方法或模型进行客观的比较, 给读者留下思考的空间, 让读者做出自己的判断。如第 7 章关于词语切分方法的比较, 第 8 章关于短语结构分析器和依存分析器的性能比较。(2) 在介绍每一部分内容时, 都引用了相关的代表性论文, 包括一些在计算语言学领域顶级国际会议上获奖的优秀论文, 如统计机器翻译中基于最大熵的翻译模型、基于层次短语的翻译模型等, 有利于读者了解相关领域的主流方法和代表性成果。(3) 对本领域使用的专业术语都给出了规范的英文注释, 为读者阅读英文文献和撰写英文论文提供参照和帮助。(4) 对于很多开源工具, 如支持向量机(SVM)、条件随机场(CRF)、最大熵(ME)、隐马尔可夫模型(HMM)等, 以及作者的课题组所实现的工具软件, 如汉语分词系统(Urheen)和句法分析器(Oboe)等, 都给出了明确的网址, 方便读者直接使用, 还可以进行对比实验。

之所以能写出这样一本好书, 是因为作者既有较深的学术造诣, 又有丰富的实践经验。宗成庆博士自 2004 年起就在中国科学院自动化研究所任研究员和博士生导师, 已有 10 年时光的高级学术活动的历练, 是自然语言处理学界的知名学者。他所取得的科研成果在国内外享有盛誉, 产生了较大的影响。例如, 他主持研发的多语言机器翻译系统多次在国际和全国机器翻译系统评测中取得第一名的优异成绩, 并在实际应用中取得了卓越的效果。宗成庆博士也在本书中介绍了自己近年来的一些研究成果, 如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、双语联合的语义角色标注方法、基于谓词论元结构转换的翻译模型等, 这些工作大都发表在高层次的国际学术会议或期刊上。

本书前言提及, 对于书中的重要内容, 作者都邀

请了同行专家或专门从事相关研究的博士研究生进行校对，并就某些问题反复进行讨论和核实。这种严谨的治学态度是值得褒扬的。

与《统计自然语言处理》第 1 版相比，第 2 版删除了一些相对陈旧的内容，如统计机器翻译中基于词的翻译方法等，增加了近年来的一些热点研究内容，如第 10 章篇章分析、第 11 章中一些新的翻译模型和第 15 章关于情感信息抽取的内容等，且在很多章节中都给出了具体的实例，而不只是介绍数学模

型，如基于词的 n 元语法模型的分词方法和基于概率上下文无关文法 (PCFG) 的句法分析方法等，有利于读者理解和实现相关算法。

综上所述，《统计自然语言处理》第 2 版的问世顺应了大数据时代自然语言处理研究和开发的需求，是自然语言处理领域的一件幸事，有利于中文信息处理事业的发展。有理由相信，本书的出版一定会受到广大读者的欢迎。