孙茂松 陈群秀 主编

语言计算与基于内容的文本处理

清华大学出版社
北京

内容简介

本书是全国第七届计算语言学联合学术会议JSCL-2003（2003年8月9日～11日，哈尔滨）的论文集。书中选录的90篇论文是从全国各地（含港、台地区）和日本学者中征集到的130篇论文中精选出来的，另有4篇论文是大会邀请报告，也收录于本书中。本书内容包括下列6类：（1）大会邀请报告；（2）词法、句法和语义；（3）计算语言学资源建设及相关技术；（4）机器翻译技术、系统及评测方法；（5）信息检索（Web智能检索、跨语言检索、文本分类、文本过滤、问答系统等）；（6）其他（包括自动查错纠错、汉字输入法、汉盲转换、语音识别、汉字识别）。

本书充分展示了国内计算语言学研究与应用的最新进展，也展示了21世纪初中文信息处理和计算语言学研究的前沿和动向，对中文信息处理的基础研究和产品开发具有重要的参考价值。本书可供计算机、语言学等专业的科研人员、工程技术人员、大学教师和研究生选读。

书　　名：语言计算与基于内容的文本处理
作　　者：孙茂松　陈群秀　主编
出 版 者：清华大学出版社　　地　　址：北京清华大学学研大厦
　　http://www.tup.com.cn　　邮　　编：100084
　　社 总 机：010-62770175　　客户服务：010-62776969
组稿编辑：薛　慧
文稿编辑：赵彤伟
封面设计：常学影
印 刷 者：北京市清华园胶印厂
发 行 者：新华书店总店北京发行所
开　　本：787×1092　1/16　印张：40.25　字数：1005千字
版　　次：2003年7月第1版　2003年7月第1次印刷
书　　号：ISBN 7-302-06916-6/TP・5120
印　　数：1~700
定　　价：100.00元

全国第七届计算语言学联合学术会议组织情况

日　　期：　2003 年 8 月 9 日～11 日
地　　点：　哈尔滨工业大学西苑宾馆

发起单位：　中国中文信息学会
　　中国计算机学会
　　中国人工智能学会
　　北京市语言学会

赞助单位：　东芝（中国）研究开发中心
　　IBM 中国研究中心
　　微软亚洲研究院
　　TRS 信息技术有限公司
　　富士通研究开发中心有限公司
　　哈尔滨工业大学慧通计算机技术有限公司
　　中国中文信息学会

组织主办单位：　哈尔滨工业大学计算机科学与技术学院
　　清华大学智能技术与系统国家重点实验室

大 会 主 席：　李　生
大会副主席：　刘开瑛

程序委员会主席：孙茂松
程序委员会副主席：俞士汶　张　普
程序委员会委员：（以姓氏拼音字母次序为序）
　　蔡莲红　曹右琦　常宝宝　陈群秀　程学旗　董振东　冯志伟　傅爱平
　　黄河燕　宋　柔　王晓龙　苑春法　赵铁军　张　全　赵　军　周　明

组织委员会主席：赵培文
组织委员会副主席：陈群秀　曹右琦
组织委员会委员：唐降龙　韩纪庆　刘　挺　李海峰　于　浩　关　毅　杨沐昀

前 言

全国第七届计算语言学联合学术会议(JSCL-2003)由中国中文信息学会、中国计算机学会、中国人工智能学会和北京市语言学会联合发起，于2003年8月9日～11日在哈尔滨工业大学举行。

本次会议共征集到论文130篇，经大会程序委员会严肃认真的评审，最终录用90篇。4篇大会邀请报告也如愿收入到论文集中。本次会议论文大体上可分为六类：

(1) 大会邀请报告，4篇；

(2) 词法、句法和语义，42篇；

(3) 计算语言学资源建设及相关技术，12篇；

(4) 机器翻译技术、系统及评测方法，12篇；

(5) 信息检索（Web智能检索、跨语言检索、文本分类、文本过滤、问答系统等），18篇；

(6) 其他（包括自动查错纠错、汉字输入法、汉盲转换、语音识别、汉字识别），6篇。

这些类别表面上似乎都是"旧面孔"，但实际上，论文所涉及的内容无论是从深度还是从广度上，较JSCL-2001的论文均取得了长足的进步。对本论文集，我们尝试着作如下圈点：

- 大会邀请报告极具启发性。《TREC2002介绍及清华大学实验研究》和《大规模内容计算》介绍了我国学者在文本信息检索领域的最新成果。在国际上著名的TREC-11文本信息检索比赛中，来自清华大学智能技术与系统国家重点实验室及中国科学院计算技术研究所的两个研究小组共取得了Web Track、Novelty Track和Adaptive Filtering Track中6项指标（寻找关键资源、查找网页入口、抽取相关信息、抽取新信息、T11SU、T11F）第一的优异成绩，令人倍感振奋。《对自动分词的反思》对我们在自动分词乃至中文信息处理整个领域的思维模式和技术路线进行了重新审视，高屋建瓴。《哈工大自然语言处理研究进展》使我们深切体会到自然语言处理研究在带动研究型大学计算机学科建设中所占据的位置。

- "词法、句法和语义"和"计算语言学资源建设及相关技术"两大类别占全部录用论文的60%，超过"半壁江山"，表明学者日益重视语言计算"基本功"的锤炼，更加着眼于长远而并不盲目追求"短平快"。这无疑是十分可喜的现象。而其中一个值得注意的动态是，针对语义研究的论文比重较以往大幅攀升，实际上反映了今后几年中文计算的一个重要的研究趋势。

- "机器翻译技术、系统及评测方法"类别的论文占全部录用论文的13.3%，其内容显示这个领域的研究目前处于比较平稳的发展阶段。

- "信息检索（Web智能检索、跨语言检索、文本分类、文本过滤、问答系统等）"类别的论文占全部录用论文的20%。显然，此类研究的上升势头正猛。在应用上与Web相结合，无疑当属天作之合。

- "其他（包括自动查错纠错、汉字输入法、汉盲转换、语音识别、汉字识别）"类别的论文仅占全部录用论文的6.7%。

展望今后几年，我们认为计算语言学的主流研究将围绕"语言计算"与"基于内容的文本处

理”这两大主题展开。前者属“老生常谈”，但它是基础，基础不牢，后者的构建势必摇摇晃晃，不可能有什么真正的前途；后者说法上似乎有点儿新意（注意，这里的“基于内容”，大致等同于“基于语义”，与图像、视像检索中所谓“基于内容”实则“基于物理属性”的提法存在根本不同）。我们应该刻意记取，这是我们研究的目的。没有它，前者将失去“动力之源”。具体而言，关于汉语的研究前沿很可能包括：

- 应用驱动（主要以 Web 为背景）的浅层汉语语言分析技术（主要指分词和局部句法分析），尤其是在天然存在一定分析错误率的条件下；
- 汉语语义计算研究；
- 大规模汉语语义资源的整合与建设；
- 词法、句法、语义一体化的汉语分析模型的研究；
- 语义 Web 及其语言支撑平台（例如，通用 ontology 及面向具体应用领域的各种专用 ontologies 的建设。相关标准的制定）；
- 大规模双语对齐语料库的建设；
- 自然语言处理技术、OCR 技术、语音识别等技术与基于内容的图像、视像处理技术的融合；
- 大规模语言计算资源共享平台与机制的建设。

我们要积极应对这些新的挑战，立足于国内现实需求，紧密依托国际学术舞台，通过坚持不懈的奋斗，力争把研究做细、做深入；把应用做实、做大。

最后，让我们感谢大会邀请报告讲者和全体作者对大会的热情支持；感谢程序委员会的辛勤劳动（全部评审过程是在“非典型肺炎”肆虐的非常时期完成的）；感谢大会组织委员会的出色工作；感谢清华大学出版社在出版方面的积极配合；感谢赞助单位的慷慨解囊；感谢中国中文信息学会计算语言学专业委员会为会议牵头所做的贡献。没有这些共同的努力，就不可能产生这本精彩纷呈的论文集，也不可能成就国内计算语言学者再度欢聚一堂的“满汉全席”。

编 者

2003 年 6 月 24 日

（当日，世界卫生组织宣布解除对北京旅游警告并将北京从 SARS 疫区名单中删除）

写于清华园

目　录

III 计算语言学资源建设及相关技术

VI 其他