

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。

版权所有, 翻印必究。

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

<http://www.tup.com.cn>

社总机: 010-62770175

地 址: 北京清华大学学研大厦

邮 编: 100084

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

I 大会邀请报告

- TREC2002 介绍及清华大学实验研究..... 张敏 马亮 马少平 陈群秀 1
大规模内容计算..... 白硕 程学旗 郭莉 王斌 余智华 刘群 13
对自动分词的反思..... 黄昌宁 高剑峰 李沐 26
哈工大自然语言处理研究进展..... 李生 39

II 词法、句法和语义

- 基于 DCC 的流行语动态跟踪与辅助发现研究..... 张普 47
流通度——字词使用情况测定的新方法..... 郑泽之 王强军 张普 54
农业病虫害词汇获取方法初探..... 郑家恒 杜永萍 宋礼鹏 61
基于 Bootstrapping 的领域词汇自动获取..... 陈文亮 朱靖波 姚天顺 张宇新 67
一种自适应概率语言模型的训练方法及其应用于中文分词
..... 徐志明 揭春雨 Jonathan Webster 73
使用互信息辅助在篇章范围内识别命名实体..... 郭志立 79
Co-Training 的机器学习方法在中文机构名识别中的应用
..... 吴雪军 朱靖波 王会珍 叶娜 张宇新 85
汉语机构名的构成模式..... 雷静 91
蒙古文人名自动识别研究..... 那顺乌日图 雪艳 淑琴 敖日格乐 97
典型参数平滑算法在词性标注中的性能评价..... 朱莉 孟遥 赵铁军 103
汉语组块的定义和获取..... 李素建 刘群 110
汉语部分分析研究..... 周强 116
时间短语的分析与识别..... 刘智颖 122
现代汉语常用动词带宾语能力调查..... 邢红兵 129
浅析“体词”的“动词”兼类现象..... 韦向峰 135
体词性并列结构的结构平行..... 吴云芳 141
面向计算机的二重复句层次划分研究..... 李晋霞 刘云 147
汉语句法分析建模中基于模型质量的特征选择方法..... 孟遥 赵铁军 杨沐昀 李生 154
现代藏语的句法组块与形式标记..... 江荻 160
现代藏语判定动词句主宾语的自动识别方法..... 黄行 江荻 167
规则和边界统计相结合的英语基本名词短语识别..... 梁颖红 赵铁军 翟舒 173
花园幽径句的某些形式特性..... 冯志伟 179

俄语句法结构的模式化描述及操作原理.....	傅兴尚	186
面向真实文本的汉语词义排歧模型研究.....	杨尔弘 李盛	193
定语类型和槽关系类型的对应及其对名词语义分析的作用.....	张卫国 梁社会	199
属性分析说略.....	陈小荷	206
隐马尔可夫模型和贝叶斯模型词义消歧对比研究.....	于江伟 刘挺 卢志茂 李生	214
基于统计的汉语词汇间语义相似度计算.....	关毅 王晓龙	221
利用语义特征生成搭配.....	赵晨光 蔡东风	228
基于《知网》的中文语块抽取器.....	董强 郝长伶 董振东	234
介连兼类词“以”的句法语义区别特征及消歧策略.....	方向红 宋春阳	240
基于情景理论分析 VA-语句.....	毛家菊 高峰 陈秋林 陆汝占	246
基于类义抽象的汉语复合词义的求解模式探索.....	宋春阳 陆汝占 方向红	252
谓词带定式的配价研究.....	王治敏 李勉东	258
基于语义依存关系的句子理解模型.....	李涓子 王作英	264
句处理中排歧问题补议.....	陆俭明 王黎	271
文本生成与理解的语言学模型——伊戈尔·梅里丘克《意思(=)文本》模型评介	易绵竹 南振兴 李绍哲 薛恩奎	278
基于格关系和配价的藏语动词再分类研究.....	陈玉忠 李保利 俞士汶	284
现代汉语复杂句蜕块研究.....	唐兴全	291
变异句蜕块的构成分析.....	孙雄勇	298
几种汉语移位现象的 HNC 研究.....	雒自清 郝惠宁 温锁林 张克亮	304
论旨网格的描写和 HNC 句类表示的比较分析.....	李千驹 唐兴全 林杏光	311
III 计算语言学资源建设及相关技术		
标注语料机器校对的研究与实践.....	曲维光 陈小荷	318
现代汉语语料的句子级语义标注.....	苗传江 刘智颖	325
双语语料库段落重组对齐方法研究.....	李维刚 刘挺 王震 李生	332
大规模非限定领域汉英双语语料库建设及句子对齐方法研究.....	刘非凡 赵军 徐波	339
蒙古语语料库建设现状分析和完善策略.....	华沙宝 巴达玛敖德斯尔	346
现代汉语语义词典(SKCC)的新进展.....	王惠 俞士汶 詹卫东	351
现代汉语述语形容词机器词典的研究与实现.....	尹一钺 陈群秀	357
基于知网的相关概念场的构建.....	董强 董振东	364
知网知识库描述语言.....	郝长伶 董强	371
汉语粘合式名词短语语义结构信息数据库.....	胡凤国 傅爱平	378
《中国大百科全书》人物传记知识提取加工规范.....	颜伟 王洁 尚英 宋柔	385
《信息处理用现代汉语分词词表》规范.....	孙茂松 王洪君 董秀芳	391

IV 机器翻译技术、系统及评测方法

- 基于 Link Grammar 的英蒙机器翻译系统 敖其尔 王斯日古楞 吉日木图 399
- 基于翻译记忆库与基于规则的汉维-维汉机器辅助翻译系统方法与框架研究
..... 吐尔根·依布拉音 艾尔肯·伊米尔 阿布力米提·阿不都热依木 405
- 机器翻译中汉语词节点的识别 王厚峰 412
- 基于汉英机器翻译的名词回指分析——句组研究之二 侯敏 孙建军 418
- 基于锚词对的英汉双语语段对齐模型 吴尉林 屈刚 陆汝占 425
- 一种汉英翻译模板提取方法 杨二宝 吕学强 朱靖波 姚天顺 431
- 基于信息熵的候选实例模式检索算法 张孝飞 陈肇雄 黄河燕 俞旻 437
- 汉语和英语逗号的对比分析及其翻译处理 张全 444
- 面向机器翻译的日语形态素解析方法 隋福民 黄德根 451
- 基于 CFC (正确性信心指数) 的学习型信赖机器翻译系统 李应潭 458
- 机器翻译测评结果的一致性 曹冬林 李堂秋 史晓东 蔡经球 464
- 汉语分词在机器翻译评价中的影响 徐冰 姚建民 杨沐昀 赵铁军 470

V 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)

- 面向 TDT 的主题相似性计算模型 朱靖波 陈文亮 姚天顺 476
- 基于标引技术的特定领域 XML 文本自动生成 刘桐菊 于浩 赵铁军 482
- 主题 Web 信息采集的研究与设计
..... 李盛韬 吴丽辉 于满泉 潘文锋 余智华 王斌 程学旗 488
- Web 关键资源发现中的链接分析技术 刘悦 王斌 杨志峰 张鑫 495
- 基于浅层分析的网页相关度研究 替红英 苏玉梅 孙斌 俞士汶 501
- 基于网页上下文分析的图片检索 刘金松 于浩 西野文人 507
- 面向英汉的跨语言信息检索关键技术研究 张玥杰 郭依昆 吴立德 513
- 面向双语句对检索的汉语句子相似度计算 车万翔 刘挺 秦兵 李生 520
- 弱指导的统计隐含语义分析及其在跨语言信息检索中的应用 金千里 赵军 徐波 527
- 一种快速的多模式串匹配算法及其在实时汉语文本分类系统中的应用
..... 张鑫 程学旗 谭建龙 王映 534
- 基于大规模真实文本的平衡语料分析与文本分类方法 陈克利 宗成庆 王霞 540
- 基于 Winnow 算法的文本过滤 赵林 夏迎炬 黄莹菁 吴立德 546
- TREC 自适应信息过滤中的目标优化技术研究 许洪波 王斌 程学旗 白硕 553
- 话题检测与跟踪技术的发展与研究 骆卫华 刘群 程学旗 560
- 基于最大熵模型的 QA 系统置信度评分算法 游斓 周雅倩 黄莹菁 吴立德 567
- 人机口语对话系统中否定结构的处理 郭荣 高峰 毛家菊 陆汝占 575
- 基于查询语义的数据库中文界面研究 张凯 吴丽辉 李盛韬 程学旗 581
- 基于动态知识库的问答系统研究 王树西 刘群 白硕 王斌 程学旗 姜吉发 587

VI 其他

- 模式匹配和句型成分分析相结合的语法错误自动检查 龚小谨 罗振声 骆卫华 593
- 中文自动查错与人机交互纠错系统的研究与实现——简介语料中文自动校对系统
..... 吴岩 蔺荪 600
- 基于多知识分析的汉盲转换算法..... 黄河燕 陈肇雄 黄静 607
- 为何汉字形码输入法难以走出“难”的困境？——谈谈一些技术上的欠妥观点
..... 张小衡 614
- 基于多路差别子空间的语速变化语音的识别 吕成国 韩纪庆 王承发 621
- 基于知识模型的手写中国地址识别系统..... 王春恒 堀田悦伸 諏访美佐子 直井聪 627