

YWCL2010

# 第五届全国青年计算语言学研讨会论文集

主办：中国中文信息学会

承办：华中师范大学

2010年10月11日-13日

## 前 言

全国青年计算语言学研讨会(YWCL)是由中国中文信息学会发起的、以青年学者和学生为主体的学术会议。会议每两年举办一次,其目的在于加强计算语言学研究领域青年学者和学生之间的学术交流与合作,促进国内计算语言学的研究和应用,提高计算语言学人才培养的水平。与其他学术会议相比,全国青年计算语言学研讨会的最重要的特点就是以青年学者和学生为主体,会议的各项活动真正做到指导委员会统筹下的青年学者与学生筹划、组织和参与。

全国青年计算语言学研讨会前称“全国学生计算语言学研讨会(SWCL)”。第一届全国学生计算语言学研讨会(SWCL2002)于2002年8月在北京大学计算语言学研究所召开,已成功举办过四届。2009年7月在“第十届全国计算语言学学术会议”上,根据与会代表的意见,从2010年起,将“全国学生计算语言学研讨会”易名为“全国青年计算语言学研讨会”,但保持研讨会的宗旨不变。本届研讨会,即第五届全国青年计算语言学研讨会(YWCL2010),是易名后的第一届研讨会,将于2010年10月在武汉华中师范大学召开。

本届研讨会共征集到论文122篇,经研讨会程序委员会严肃认真地评审,最终录用口头报告论文76篇,录用率为62.3%。程序委员会还从录用论文中精选出了一批论文推荐至《中文信息学报》发表。

本届研讨会论文集的内容非常丰富,收录的论文主要涵盖了以下四个主题:

- 1、词法、句法、语义和篇章分析
- 2、智能检索
- 3、语言资源建设及相关技术
- 4、机器翻译技术、系统及评测方法

此外还有少量论文关注于与自然语言处理相关的其他技术和应用。在这些论文中,有近三分之二关注于“词法、句法、语义和篇章分析”与“智能检索”两个主题,这一方面反映了寻求跨越语义鸿沟之道仍然是青年学者与学生的梦想所在,另一方面也反映了青年学者与学生一直在为将计算语言学研究的新成果进行转化并应用于其他领域的研究而付出不懈的努力,这些都显示了国内计算语言学研究的良好势头。

在此,让我们感谢全体作者和与会代表对研讨会的热情参与;感谢中国中文信息学会和本届会议指导委员会诸位老师的悉心指导;感谢全体程序委员会委员的辛勤劳动;感谢研讨会组织委员会的出色工作;感谢赞助单位的慷慨解囊;感谢华中师范大学为研讨会顺利召开所提供的大力支持。

预祝研讨会取得圆满成功!

第五届全国青年计算语言学研讨会程序委员会  
2010年7月

## 第五届全国青年计算语言学研讨会（YWCL-2010）组织情况

日期：2010年10月11日—13日

地点：武汉华中师范大学

发起单位：中国中文信息学会

承办单位：武汉华中师范大学

### 指导委员会

主席：

李宇明（中国教育部语言文字信息管理司司长、教授）

副主席：

孙茂松（清华大学计算机科学与技术系主任、教授，计算语言学专委会主任委员）

何婷婷（华中师范大学计算机科学系教授）

委员：（按拼音排序）

陈群秀（清华大学计算机科学与技术系教授）

程学旗（中科院计算所软件室研究员）

黄河燕（中科院计算机语言信息工程研究中心研究员）

黄居仁（香港理工大学文学院院长、教授）

黄萱菁（复旦大学计算机科学与工程系教授）

李茹（山西大学计算机与信息技术学院教授）

刘群（中科院计算所多语言交互技术评测实验室研究员）

刘绍明（日本富士施乐有限公司主任研究员）

刘挺（哈尔滨工业大学计算机学院信息检索实验室教授）

施水才（TRS信息技术有限公司总裁）

史晓东（厦门大学计算机与信息工程学院计算机系教授）

孙乐（中科院软件所中文中心研究员）

王海峰（百度搜索研发部高级科学家）

王小捷（北京邮电大学智能科学技术中心教授）

荀恩东（北京语言大学语言信息处理研究所教授）

于浩（富士通研究开发有限公司信息技术部部长）

杨尔弘（北京语言大学应用语言学研究所教授）

俞士汶（北京大学计算语言学研究所教授）

张敏（清华大学计算机科学与技术系副教授）

赵军（中科院自动化所研究员）

## 大会主席：

张 勇（武汉大学 博士生）

## 程序委员会

主席：

李 鹏（清华大学计算机科学与技术系 博士生）

副主席：

贾玉祥（北京大学计算语言学研究所 博士生）

涂新辉（华中师范大学 博士生）

委员：（按拼音排序）

丁泽亚（中国科学院声学研究所 博士生）

丁卓冶（复旦大学计算机学院 博士生）

黄哲煌（厦门大学 博士生）

江会星（北京邮电大学智能科学与技术中心 博士生）

姜尚仆（清华大学计算机科学与技术系 硕士生）

康生巧（沈阳航空工业学院 硕士生）

李双红（山西大学计算机与信息技术学院 硕士生）

李正华（哈尔滨工业大学信息检索研究中心 博士生）

梁社会（南京师范大学文学院 博士生）

齐振宇（中科院自动化所模式识别国家重点实验室 博士生）

仇 伟（上海交通大学计算机系 硕士生）

肖 桐（东北大学自然语言处理实验室 博士生）

熊 皓（中国科学院计算技术研究所 硕士生）

张 育（苏州大学 硕士生）

周明海（鲁东大学中文信息处理研究所 硕士生）

朱小飞（中国科学院计算技术研究所 博士生）

邹红建（北京语言大学应用语言学研究所 博士生）

## 组织委员会

主席：

胡 珀（武汉大学 博士生）

副主席：

张 勇（武汉大学 博士生）

委员：

陈劲光（华中师范大学 博士生）

舒江波（华中师范大学 博士生）  
李 芳（华中师范大学 博士生）  
宋 乐（华中师范大学 硕士生）  
杨 柳（华中师范大学 硕士生）  
段秀婷（华中师范大学 硕士生）  
张红春（华中师范大学 硕士生）  
江腾飞（华中师范大学 硕士生）  
万 剑（华中师范大学 硕士生）  
娄振霞（华中师范大学 硕士生）  
董婧灵（华中师范大学 硕士生）  
周琨峰（华中师范大学 硕士生）

赞助单位：  
富士施乐有限公司  
富士通研究开发有限公司  
教育部语言文字信息管理司  
中国中文信息学会  
TRS 信息技术有限公司  
百度公司

# 目 录

## I 词法、句法、语义和篇章分析

基于网络百科全书的中文关联语义知识获取	杨柳 何婷婷 涂新辉	1
特定主题概念关联的挖掘及其表示式的实现	丁泽亚 缪建明 张全	8
基于统计的词素切分算法	董兴华 杨雅婷 陈丽娟 周喜 吐尔洪·吾司曼	15
基于 PMI-IR 算法的 Blog 情感分类研究	段秀婷 何婷婷 宋乐	22
基于词典的名词性隐喻识别	贾玉祥 俞士汶	29
基于树核函数的中文语义角色标注研究	王步康 王红玲 袁晓虹 周国栋	36
基于错误驱动的现代汉语方位词用法规则的自动更新	吴云鹏 咎红英	43
基于北大网库的语义角色分类	杨敏 常宝宝	50
基于概率潜在语义分析的词汇情感倾向判别	宋晓雷 王素格 李红霞	57
朝鲜语对格的语义角色分析	李琳 毕玉德 陈洁	64
上下文边界可变的贝叶斯分类器词义消歧方法	吴崇斌 张全	71
基于例句语料库的现代汉语方位词用法自动识别研究	买志玉 赵丹 咎红英 张坤丽	77
句法特征在动词词义排歧中的应用	王宏显 周强	82
基于 TCRF 的核心框架元素标注	王智强 刘海静 李双红 李茹	89
基于规则的现代汉语连词用法自动识别研究	周丽娟 张坤丽 袁应成 咎红英	96
汉语句法成分中心词自动识别方法的研究	任晓娜 王莹莹 周俏丽 蔡东风	103
事件预期属性的标注	邹红建 杨尔弘	110
基于 MC-Value 的非句蜕广义对象语义块的边界识别	臧翰芬	117
汉语的计量特征在语言风格对比及作家判定中的应用——以韩寒《三重门》与郭敬明《梦里花落知多少》为例	陈芯莹 李雯雯 王燕 王璐 阚明刚	124
基于电影对白的现代汉语普通话语音历时对比分析	王燕 刘俊 阚明刚 侯敏 邹煜	131
汉语语篇修辞结构标注实验	邱武松	138
基于概率和句法分析的中文句子修剪	陈劲光 何婷婷 李芳 桂卓民	145
汉、蒙、藏、维分词与词性标注技术发展现状研究	通拉嘎	152
从迭句中辨识出三类花园幽径句	池哲洁 池毓焕 张全	159
现代韩国语“控制”类动词下位语义分类研究	陈洁 毕玉德 李琳	166

语言监测中词语构造能力的分析及其应用——曾小兵 邱丽娜 张普 张志平 杨尔弘 173

## II 语言资源建设及相关技术

- “非常”、“特别”还是“相当”——基于语料库的用法计量研究——阚明刚 王燕 王华英 180
- 基于动态流通语料库的连词考察——李艳娇 杨尔弘 187
- 基于标注语料库的现代汉语状元槽序研究——周明海 亢世勇 194
- 面向汉韩机器翻译的隐喻研究及隐喻知识库构建设想——徐超 201
- 维吾尔语口语语音语料库的设计与研究——杨雅婷 马博 王磊 吐尔洪·吾司曼 李晓 208
- 俄语军事缩略语知识库的构建——徐进 215
- 基于句对质量和覆盖度的统计机器翻译训练语料选取——姚树杰 肖桐 朱靖波 221
- 基于句子级的领域倾向词表构建——张小琴 蒋秀凤 228
- 基于流形排序的领域词抽取方法——宋涛 李素建 234
- 从语义关系的复杂性看语义词典建设——严灿勋 刘慧敏 241

## III 机器翻译技术、系统及评测方法

- 汉语对应英语定语从句结构的一种自动翻译方法——王雷 常宝宝 俞士汶 248
- 移动终端机器翻译设备的解码定点化方法——李响 徐金安 刘群 吕雅娟 姜文斌 255
- 利用依存限制抽取长距离调序规则——涂兆鹏 刘群 林守勋 261
- 基于规则的名词短语预调序——牟小峰 荀恩东 268
- 基于最大熵短语重排序模型的特征抽取算法改进——孙萌 姚建民 吕雅娟 刘群 姜文斌 275
- 一种改进词语对齐的新方法——罗维 吉宗诚 吕雅娟 刘群 282
- 模糊匹配在树到串翻译模型中的应用——熊皓 刘洋 刘群 289
- 题录信息的机器翻译方法——李贤华 于淼 吕雅娟 296
- 汉藏短语抽取——诺明花 张立强 刘汇丹 吴健 丁治明 303

## IV 智能检索(信息检索、信息抽取、文本挖掘、情感分析与热点问题

## 发现、话题跟踪、文本分类、文本过滤、自动文摘、问答系统等)

中文博客标签调查分析及标签推荐模型的研究	宋洪鑫 李蕾 刘冬雪	310
一种适用于语言模型的检索词扩展方法	张斌 周延泉	317
汉语情感问题类型分类研究	葛正荣 李婷玉 姚天昉	324
基于信息结构的突发事件文本事件信息自动抽取策略研究	曾青青 杨尔弘 朱丹青	331
面向传媒语言语料库的关键词自动抽取研究	吴继媛 孙淳 侯敏	338
基于百科知识库的主题扩展研究	闻彬 何婷婷	345
越南语文献中字母缩略语自动提取研究	张海云 张超静 毕玉德	351
面向查询的多模式自动摘要研究	李芳 何婷婷	358
基于流行排序的查询推荐方法	朱小飞 郭嘉丰 程学旗 杜攀	365
少数民族汉语考试作文自动评分的特征提取研究	蔡黎 彭星源 柯登峰 赵军	372
一种基于认知情景框架的文本分类方法	李月伦 李湘 常宝宝 袁毓林	379
基于语句相似度的网页标题抽取方法	李国华 咎红英	386
LDA 主题驱动的中文多文档自动文摘方法	张明慧 王红玲 周国栋	393
唐诗文本自动分类的算法研究	匡海波 陈小荷	399
基于主题情感句的汉语评论文倾向性分析	杨江 侯敏 王宁	406
基于 LDA 的关键词抽取方法	翁伟 王厚峰	413
IR4QA 系统中基于维基百科的查询扩展	周斌 刘茂福 陈建勋	418
面向音乐领域的文本检索与挖掘系统	付瑞吉 秦兵 刘挺	424
音乐领域典型事件抽取方法研究	丁效 宋凡 秦兵 刘挺	431
依存信息在蛋白质关系抽取中的作用	刘兵 徐华 钱龙华 周国栋	438
基于维基百科类别的文本特征表示	王锦 王会珍 张俐	445
维基百科人物属性自动获取方法研究	孟新萍 王会珍 张俐	452
基于汽车领域的情感问答系统设计及实现	栾家阳 张文波 姚天昉	459
隐喻化新词的考察	李惠 冯敏萱	466

## V 其他

基于音系理论的变调自动处理模型	贺俊杰	473
-----------------	-----	-----



基于法律文本的藏语句子边界识别	赵维纳 刘汇丹 于新 吴健 张普	480
基于 Uniscribe 和 OpenType 的蒙古文字处理软件 MWord 的设计与实现	斯·劳格劳 华沙宝 萨如拉	487
基于声频特征的维吾尔语语音端点检测方法	杨雅婷 马博 王磊 吐尔洪·吾司曼 李晓	494
基于 XML 的语言技术平台	李正华 车万翔 刘挺	501
Win32 平台下女书拼音输入法的设计与实现	王鹏 孙茂松	508
基于日志分析的中文输入法用户行为研究	许丹青 刘奕群 岑荣伟 马少平 茹立云 杨磊	515