SWCL2006

The 3rd Students' Workshop on Computational Linguistics

Sponsor:

The Chinese Information Processing Society of China

Organizer:

ShenYang Institute of Aeronautical Engineering
Human-Computer Intelligence Research Center

前　言

学生计算语言学研讨会，是由中国中文信息学会发起的、以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学生之间的学术交流和合作，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。与同类会议相比，学生计算语言学研讨会的一大特点是以学生为主体，会议的各项活动真正做到指导委员会统筹下的学生筹划、学生组织和学生参与。

继第一届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，第二届全国学生计算语言学研讨会（SWCL2004）于 2004 年 8 月在北京语言大学信息科学学院成功召开之后，第三届学生计算语言学研讨会（SWCL2006）由沈阳航空工业学院人机智能研究中心承办，并将于 2006 年 8 月 16 日—18 日在沈阳航空工业学院举行。

本次会议共征集到论文 106 篇，程序委员会深信每篇稿件都饱含着作者的艰辛努力，因此，程序委员会为每篇论文都组织了至少两位委员进行评审。通过委员评审和主席复核两个阶段，最终确定了论文录用结果。会议最终录用口头报告论文 68 篇，张贴论文 21 篇，本届会议论文的数量和录用论文的质量都超过了上届会议，这在反映出国内计算语言学研究水平不断提高的同时，也见证了学生计算语言学研讨会的成长。本届会议口头报告论文的主题大致分为以下六类：

[1] 词法、句法与语义分析，17 篇；

[2] 知识表示与机器学习，17 篇；

[3] 信息检索、抽取与过滤，12 篇；

[4] 语料库语言学，12 篇；

[5] 文本自动分类，5 篇；

[6] 机器翻译，5 篇。

计算语言学是一个交叉学科，从此次会议论文的涉及内容来看，无论是在理论方面、还是应用方面，计算语言学的许多研究领域都有同学们参与研究的踪迹，有些研究还嵌入到了实际的产品当中。在此基础上，我们希望此次研讨会能够对计算语言学研究领域的学生交流和研究进展做出应有的贡献。

最后，让我们感谢全体作者和与会代表对会议的热情参与；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；感谢全体程序委员的辛勤劳动；感谢会议组织委员会的出色工作；感谢赞助单位的慷慨解囊；感谢沈阳航空工业学院人机智能研究中心为会议顺利召开所提供的大力支持。

预祝会议取得圆满成功！

第三届全国学生计算语言学研讨会程序委员会

2006 年 8 月

SWCL-2006 组织机构

发起单位：中国中文信息学会
承办单位：沈阳航空工业学院人机智能研究中心

指导委员会：
主　任：　李宇明 教授（中国教育部语言文字信息管理司司长）
副主任：　孙茂松 教授（中国中文信息学会计算语言学专业委员会主任委员）
　　蔡东风 博士（沈阳航空工业学院教授）
　　任福继 博士（日本德岛大学工学部教授）
委　员：（按拼音排序）
　　陈群秀 （清华大学计算机系教授）
　　程学旗 （中科院计算所软件室研究员）
　　何婷婷 （华中师范大学计算机科学系教授）
　　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 （中研院语言学研究所研究员，北京大学计算语言学研究所兼职教授）
　　黄萱菁 （复旦大学计算机科学与工程系教授）
　　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 （富士施乐有限公司研究员）
　　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副研究员）
　　王小捷 （北京邮电大学智能科学技术研究中心教授）
　　苟恩东 （北京语言大学副教授）
　　俞士汶 （北京大学信息科学技术学院计算语言学研究所教授）
　　张　敏 （清华大学计算机系讲师）
　　赵　军 （中科院自动化所副研究员）
　　赵铁军 （哈尔滨工业大学计算机学院语言技术研究中心教授）
　　朱靖波 （沈阳东北大学信息学院计算机软件所副教授）

大会主席：赵小兵 （北京语言大学 博士生）

程序委员会
主　席：　刘奕群（清华大学计算机系 博士生）
副主席：　车万翔（哈尔滨工业大学计算机学院信息检索实验室 博士生）
　　朱　虹（北京大学信息科学技术学院计算语言学研究所 博士生）
委　员：（按拼音排序）
　　陈毅东（厦门大学计算机与信息工程学院计算机系 博士生）
　　费仲超（复旦大学计算机科学与工程系 博士生）
　　郭瑞杰（中科院计算所软件室 博士生）
　　何中军（中科院计算所多语言交互技术评测实验室 博士生）
　　黄　瑾（中科院计算所多语言交互技术评测实验室 硕士生）
　　金　澎（北京大学信息科学技术学院计算语言学研究所 博士生）

人员或用户进行词条的甄选，得到一个小规模词表。然后，利用这个词表进行自动分词，在未切分的汉字串中，抽取出更多的词条，由人工进行判定。不断地重复这个人机交互的过程，最终完成对文本的分词。

使用统计方法进行自动抽词，必然面对两个问题，即探测未登录词和适应不同的分词标准。从技术上考虑，这两个问题可以看作两个过程，一是如何自动地提取候选字串并进行自动筛选，保证词语的精确率和召回率；二是如何通过人机交互来确定词表，让词语符合用户的分词标准。本文提出了改进的后缀数组抽词算法，对抽取的候选词语采用互信息（MI）进行过滤，得到了性能较好的自动抽词模块。同时，提供较好的人机交互界面，便于用户增删词语。

3.2 改进的后缀数组自动抽词算法

提取候选字串时，最大的问题是会产生大量垃圾。如，假设在一个文本中，字串“萨达姆”、“萨达”和“达姆”的频次都为 10 次。很明显，“萨达”和“达姆”是需要过滤的字串。针对这一问题，目前主要有两种做法：一种是基于哈希表，计算文本中所有的 n 元字串的频次，然后使用频次相减法来过滤（金翔宇 等 2001）。另一种是利用排序的后缀数组，直接把数组序列中前缀相同的字串提取出来，这样可以排除掉“萨达”。同样地，建立排序的前缀数组来排除“达姆”（Luo Zhiyong et al. 2004）。这两种方法，在时间和空间上开销过大，难以满足实用系统的需要，也无法提取频次为 1 的词语。为了解决计算效率问题，我们提出了改进算法，只需建立一个排序的后缀数组，就可以完成排除子串的过程。首先，利用排序的后缀数组，可以排除“萨达”。然后，进一步利用上文的后缀信息来提取候选串。如果上文有相同的字符，则不算作候选字串。由于“达姆”所在的后缀数组，其上文必定为“萨”，可以被排除。通过计算它们上文相同的长度，即上文最长公共后缀（LLCS，Longest Left Common Suffix）的长度，就可以跳过这些被长串完全覆盖的子串。在不增加空间开销的前提下，把算法的时间复杂度由原来的$O(N^2)$降到了$O(N*\lg N)$。由此，利用邻串的 LCP、LLCS 值，可以从文本中自动获得大量的 n 元字串。图 2 是从 1998 年 1 月人民日报语料中提取出来的一部分以“中国”开头的后缀数组。

了解了中国，从而向往中国，想去中国看看，但直到去世也未圆这个梦。
闭界的共同努力，增进各国人民对中国真实情况的了解，促进相互间友好
望老教授考察教师住房和市场指出中国知识分子勤奋敬业爱国精神是民族
清和奉美林都有一个强烈的感受，中国知识分子太可贵了，他们勤奋、敬业
染力自有扣人心弦处。邓在军，在中国知道她的人并不少。她开创了中央电视台
通过国家检收。该工程是辽宁省和中国石化总公司联合兴建的大型石油化工项目
探和原油生产任务。１９９７年，中国石油天然气总公司为寻找新的可采资源，
为国民经济发展作出更大的贡献。中国石油天然气总公司去年共发现１０个亿吨
略格局已基本形成。在今天召开的中国石油天然气总公司工作会议上，周永康总
）国务院总理李鹏今天下午在接见中国石油天然气总公司工作会议代表时强调，
起来。侯祥麟今年８５岁，现在是中国石油天然气总公司高级顾问。侯老向
届工业矿山度开物管理政策”与“中国矿山复垦技术指南”两项软课题，对我国
布，南非总统曼德拉已任命南非驻中国研究中心主任盛克瑞为首任驻华大使。
时发展有相适应的部分。但是，在中国确定中长期目标时，需要重视在现行经济

图 2 1998 年 1 月语料中“中国”的排序后缀数组示例

利用改进的后缀数组提取 n 元字串，只能解决长串包含短串的问题和系统的时空开销问题，并不能直接提取出真正的词，依然存在以下问题：

1） 只能提取频次为 2 以上的 n 元字串，导致频次为 1 的词无法提取。

2） 跳跃出现相同的字串，频次需要累加。如，上图中的多处出现的“中国”。

3） n 元字串大于真正的词，边界不好确定，这种情况数量庞大，需要筛选和过滤。如，“中国石油天然气总公司工作会议”。

4） n 元字串小于真正的词，边界不好确定，这种情况数量很少，如“哥伦比亚”、“无与伦比”造成的“伦比”。

针对这四个问题，我们提出了相应的解决方案：

1） 使用左右扩展法在后缀数组中提取频次为 1 的低频字串。方法是利用频次为 1 的二字串进行左右扩展。对汉字串 ABCDE，如“起诉萨达姆”中，假设“起诉、诉萨、萨达、达姆”的频次分别为 5、1、1、1。假设以“萨达”为出发点，即 CD 的频次为 1，向左扩展，如果 BC 的频次为 1，扩展为 BCD（诉萨达）；再往左扩展，如果 AB 的频次大于 1，则删除 B；以此类推不断向左扩展，确定左边界；以相同的方式可以确定右边界。同时屏蔽掉后缀数组中其他出发点的二字串。最后得到频次为 1 的“萨达姆”。需要说明的是，该方法也只能获取一

参加审稿人员名单

刘奕群（清华大学计算机系 博士生）

车万翔（哈尔滨工业大学计算机学院信息检索实验室 博士生）

朱　虹（北京大学信息科学技术学院计算语言学研究所 博士生）

陈毅东（厦门大学计算机与信息工程学院计算机系 博士生）

费仲超（复旦大学计算机科学与工程系 博士生）

郭瑞杰（中科院计算所软件室 博士生）

何中军（中科院计算所多语言交互技术评测实验室 博士生）

黄　瑾（中科院计算所多语言交互技术评测实验室 硕士生）

金　澎（北京大学信息科学技术学院计算语言学研究所 博士生）

陆　敏（中科院自动化所 硕士生）

苗雪雷（沈阳航空工业学院自然语言处理研究室 硕士生）

乔　维（清华大学计算机系 博士生）

谭红叶（哈尔滨工业大学计算机学院语言技术研究中心 博士生）

王会珍（东北大学信息学院计算机软件所 博士生）

王　洁（北京语言大学计算机系 博士生）

魏勇鹏（清华大学计算机系 硕士生）

杨志豪（大连理工大学计算机系 博士生）

袁彩霞（北京邮电大学 博士生）

张大鲲（中科院软件所中文中心 博士生）

赵小兵（北京语言大学 博士生）

支　流（北京大学信息科学技术学院计算语言学研究所 硕士生）

周　浪（北京中科院计算机语言信息工程研究中心 博士生 ）

瞿国忠（华中师范大学计算机科学系 硕士生）

王　强（哈尔滨工业大学智能技术与自然语言处理实验室 博士生）

SWCL-2006 日程安排（总）

8 月 16 日	8:00-8:30	注册	国际交流中心一楼	主席：赵小兵
	8:30-9:00	开幕式	国际交流中心学术报告厅	
	9:00-9:30	合影留念	国际交流中心楼前	
	9:30-10:10	特邀报告 1：任福继	国际交流中心学术报告厅	
	10:10-10:50	特邀报告 2：张桂平		
	10:50-11:10	休息		
	11:10-11:50	特邀报告 3：王小川	国际交流中心学术报告厅	
	12:00	午餐	航院宾馆餐厅	
	13:30-15:30	session1A, session1B	分会场 1 分会场 2	
	15:30-15:50	休息		
	15:50-17:30	session2A, session2B	分会场 1 分会场 2	
	17:30	宴会	航院宾馆餐厅	
	19:30	与企业界代表的座谈会	国际交流中心学术报告厅	
8 月 17 日	8:00-8:40	特邀报告 4：黄居仁	国际交流中心学术报告厅	主席：刘奕群
	8:40-9:20	特邀报告 5：程学旗		
	9:20-9:30	休息		
	9:30-11:30	poster 和 demo session	实验楼二楼大厅	
	11:30	午餐	航院宾馆餐厅	
	12:30-18:00	参观沈阳世界园艺博览会		
	19:00	晚餐	航院宾馆餐厅	
8 月 18 日	8:00-10:20	session3A, session3B	分会场 1 分会场 2	
	10:20-10:30	休息		
	10:30-12:30	session4A, session4B	分会场 1 分会场 2	
	12:30	午餐	航院宾馆餐厅	
	13:30-15:30	session5A, session5B	分会场 1 分会场 2	
	15:30-15:50	休息		
	15:50-17:50	session6A, session6B	分会场 1 分会场 2	
	17:50-18:20	闭幕式	国际交流中心学术报告厅	
	18:30	宴会	航院宾馆餐厅	

SWCL-2006 日程安排（分）

场次 session1A: 8 月 16 日 13:30-15:30　分会场 1
主持人：杨志豪　　主题：文本自动分类

13:30-13:50	关于文本分类中特征降维方式的研究
13:50-14:10	基于背景知识的文本自动分类
14:10-14:30	中文网页形式自动分类
14:30-14:50	一种基于主题的文本聚类方法
14:50-15:10	面向对外汉语报刊教学的文本难易度分类
15:10-15:30	Impact of the Size of Training Set on Text Categorization

场次 session1B: 8 月 16 日 13:30-15:30　分会场 2
主持人：张大鲲　　主题：机器翻译

13:30-13:50	日中机器翻译中汉语副词的数据处理
13:50-14:10	基于非连续短语的统计翻译模型
14:10-14:30	中国哈萨克阿拉伯文与哈萨克斯拉夫文文本转换
14:30-14:50	翻译规则优化中的分层优化方法
14:50-15:10	统计机器翻译中短语切分的新方法
15:10-15:30	影响统计翻译系统性能的因素分析

场次 session2A: 8 月 16 日 15:50-17:30　分会场 1
主持人：孙景广　　主题：语义分析

15:50-16:10	语料库语义成分标注的若干问题
16:10-16:30	中文语义角色标注的特征工程
16:30-16:50	中文褒贬义词语倾向性的分析
16:50-17:10	基于知网的中文问题自动分类
17:10-17:30	基于语义理解的文本倾向性识别机制

场次 session2B: 8 月 16 日 15:50-17:30　分会场 2
主持人：李姣　　主题：知识表示

15:50-16:10	基于标注语料库的组合歧义检测与消解
16:10-16:30	基于规则方法的汉语到语义网络语言的转换研究
16:30-16:50	一种基于 HNC 理论的领域知识表示研究
16:50-17:10	生物文献的本体建模及其在语义查询中的应用
17:10-17:30	词汇语义相似度计算中相关技术的分析

场次 session3A: 8 月 18 日 8：00－10：20 分会场 1

主持人：瞿国忠　　主题：信息检索

8:00-8:20	基于主题词对的文档重排方法
8:20-8:40	基于大规模日志分析的网络搜索引擎用户行为分析
8:40-9:00	语义理解下的自然语言处理及信息检索模型
9:00-9:20	一种利用链接分析的 Web 话题跟踪方法
9:20-9:40	基于网页框架和规则的网页噪音去除方法
9:40-10:00	Web 信息检索中相关词提示技术与评测
10:00-10:20	基于改进编辑距离和依存结构的句子相似度计算

场次 session3B: 8 月 18 日 8：00－10：20 分会场 2

主持人：钱小飞　　主题：词法与句法分析(1)

8:00-8:20	基于机器学习的分词不一致自动识别研究
8:20-8:40	蒙古文编码转换软件的设计与实现
8:40-9:00	面向大型叙事作品的指人成分识别
9:00-9:20	面向中文陌生文本的人机交互式分词方法
9:20-9:40	针对 SVM 中文分词特性的个性化后处理设计
9:40-10:00	在篇章中面向产品类的命名实体识别研究
10:00-10:20	维吾尔语的词性标注校对初探

场次 session4A: 8 月 18 日 10：30－12：30 分会场 1

主持人：肖镜辉　　主题：词法与句法分析(2)

10:30-10:50	词汇化概率句法分析与动词子语类框架获取的互动方法
10:50-11:10	面向句法分析的样本选择
11:10-11:30	粤拼序列自动切分算法的研究
11:30-11:50	语音识别后文本纠错处理
11:50-12:10	LTP：语言技术平台
12:10-12:30	基于生语料、最大匹配切分语料以及熟语料的中文词频估计方法

场次 session4B: 8 月 18 日 10：30－12：30 分会场 2

主持人：安娜　　主题：语料库语言学(1)

10:30-10:50	小学生语言偏误分析
10:50-11:10	语文词典标注词性的基本原则
11:10-11:30	语料库中的插入语标注研究
11:30-11:50	基于标注语料库的《新闻联播》语言特征统计分析
11:50-12:10	基于多语境的相关词自动提取
12:10-12:30	基于语料统计的以“不”开头双字分词不一致研究

场次 session5A: 8 月 18 日 13：30－15：30 分会场 1

主持人：袁彩霞　　主题：语料库语言学(2)

13:30-13:50	基于半监督最大熵模型的汉语词性标注
13:50-14:10	带标注语料库中切分变异的统计分析及思考
14:10-14:30	中文缩略语知识库建设
14:30-14:50	《英汉蒙电子词典》的设计与实现
14:50-15:10	《蒙古语语法信息词典字符分库》的建立及意义
15:10-15:30	基于语料库的数量名短语识别

场次 session5B: 8 月 18 日 13：30－15：30 分会场 2

主持人：王波　　主题：机器学习 (1)

13:30-13:50	条件随机域模型和实验分析
13:50-14:10	中文单词聚类的比较研究
14:10-14:30	现代汉语动态助词“了”的自动生成研究
14:30-14:50	汉语空间关系中射体识别问题的研究与分析
14:50-15:10	基于特征选择和语义扩展的词序列核函数研究
15:10-15:30	基于结构描述的汉字字形相似度计算

场次 session6A: 8 月 18 日 15：50－17：50 分会场 1

主持人：郎君　　主题：机器学习 (2)

15:50-16:10	中国人名性别自动识别
16:10-16:30	媒体用语中的语误分析
16:30-16:50	基于抽样的两阶段支持向量机训练算法
16:50-17:10	基于传媒语音语料库的不同语言样式统计分析
17:10-17:30	Dotplotting 文本分割技术的分析与改进
17:30-17:50	关键词密度分布法在偏重摘要中的应用研究

场次 session6B: 8 月 18 日 15：50－17：50 分会场 2

主持人：乔春庚　　主题：信息抽取与过滤

15:50-16:10	术语自动提取中的领域度计算方法研究
16:10-16:30	规则与统计相结合的案件名称识别
16:30-16:50	中文事件抽取中事件类别的自动识别
16:50-17:10	基于数据挖掘思想的网页正文抽取方法的研究
17:10-17:30	基于用户聚类的电子商务推荐系统
17:30-17:50	汉语 base NP 识别：错误驱动的组合分类器方法

Poster：8 月 17 日 9：30—11：30 实验楼二楼大厅

主持人：陈志玮

基于渡越矩阵的复句关系词自动标注初探
基于规则的复句中的关系词标注探讨
基于复句语料库的分词系统的研究
中国 EFL 学习者自动作文评分探索
文本篇章结构的自动标引
一种基于 HTML 位置信息的查询扩展技术
基于标注语料库的情景分析
外国人汉语虚词辅助学习系统研究
基于标注语料库以[S][P][O]为样本的句系研究
汉语依存树库的构建
基于语义统计的中文自动文摘研究
基于条件随机域的中文命名实体识别
“不是”的用法及自动处理研究
现代汉语“名+名+名”组合的统计分析
短语结构树到依存树的转换
基于问句相似度的中文 FAQ 问答系统研究
基于 PageRank 和锚文本的网页排序研究
基于条件随机域的生物医学命名实体识别
新编同义词词林语义分类体系
路径表达式在分词算法中的应用
构建“尹湛纳希辞典”的设想

目 录

I 词法、句法与语义分析

II 知识表示与机器学习

III 信息检索、抽取与过滤

IV 语料库语言学

V 文本自动分类

VI 机器翻译

VII 张贴论文