

中文语言资源建设与评测 ——回顾与展望

刘群

2008.11.24

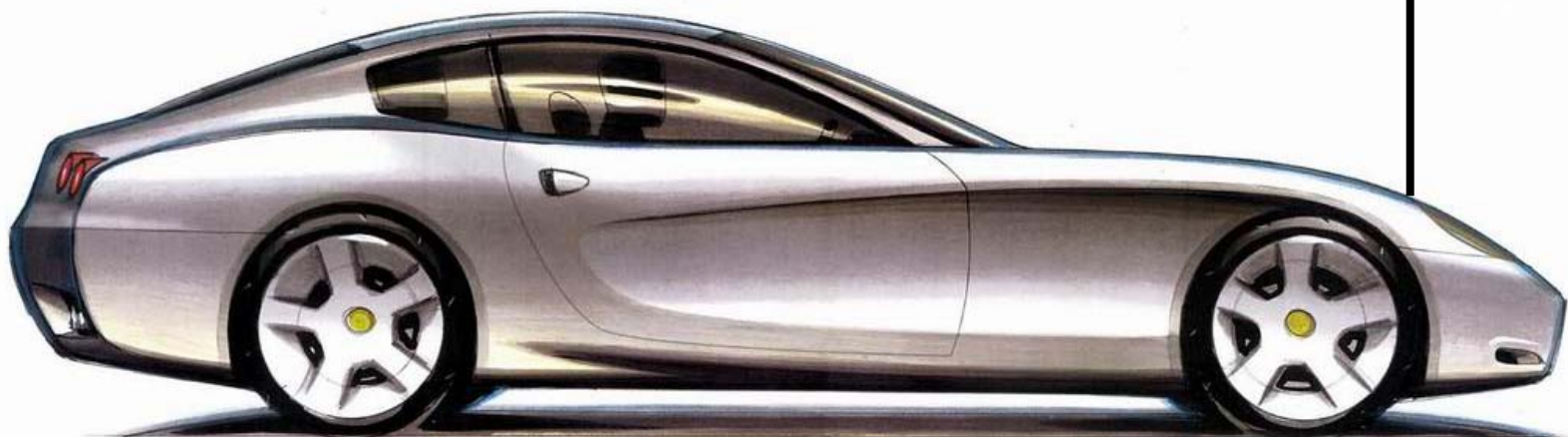
中文信息学会成立27周年学术年会

提纲

- 引言——统计自然语言处理的两大驱动力
- 回顾
- 展望

统计自然语言处理的两大驱动

统计自然语言处理



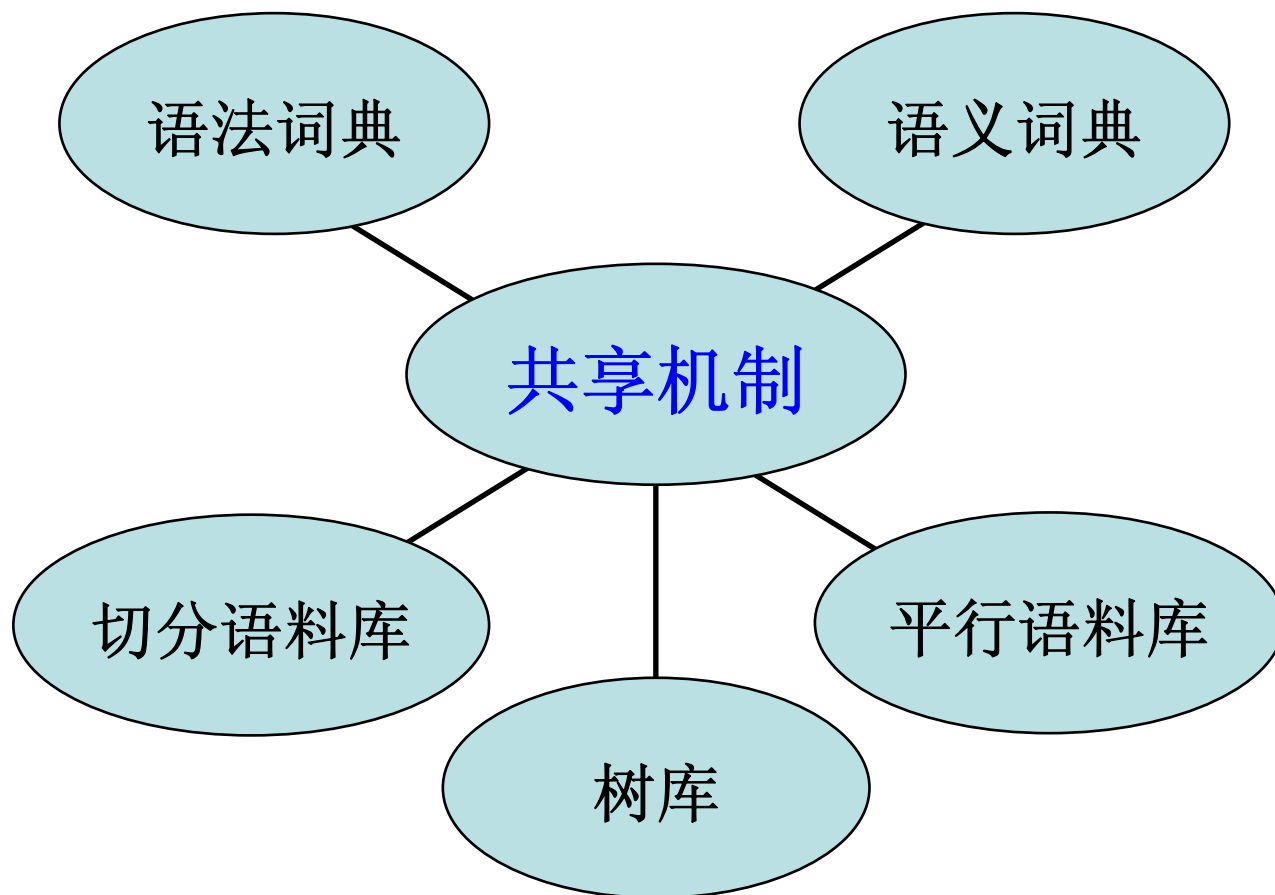
大规模共享
语言资源

公开周期性
技术评测与交流

提纲

- 引言
- 回顾
 - 语言资源建设
 - 中文信息处理评测
- 展望

中文语言资源建设成果



词典

- 北京大学语法信息词典
- 北京大学中文概念辞书 (CCD)
- 知网 (Hownet)
- 同义词词林
-等等

语料库

- 《人民日报》切分标注语料库
- 清华大学汉语句法树库
- 教育部汉语监测语料库和汉语平衡语料库
- 北京大学、中科院计算所和自动化所、哈工大、厦门大学等单位的汉英双语语料库
- 北大天网和搜狗实验室的信息检索语料库
- 香港城市大学的汉语共时语料库
- 台湾中研院的现代汉语平衡语料库
- 等等

资源共享平台

资源提
供单位

ChineseLDC
中文语言
资源联盟

资源使
用单位

支撑平台
中文信息学会

提纲

- 引言
- 回顾
 - 语言资源建设
 - 中文信息处理评测
- 展望

早期的863评测

- 我国最早的中文信息处理公开评测是863评测，开始于1990年代初
- 起步时间几乎与国外同类评测同步
- 极大地推动了我国的相关领域的研究工作
- 官方性质，比较侧重于比赛的成绩，而对通过评测进行学术交流重视不够，也对其自身发展造成一些不利影响

后期的863评测

2003-2005年间，863评测在停顿了五年之后恢复，连续举行了三届

- 形式上的变化

- 现场评测 → 网络评测

- 无训练语料不受限 → 有训练语料的受限评测

- 没有研讨会交流 → 举行专门的评测研讨会

- 项目增加，最多时8个大项10多个小项

- 参评系统最多时达100多个

与国际评测接轨，推动了国内研究的进展

中文信息学会评测

近年来，国内的自然语言处理技术评测发生了很大的变化。国家**863**计划不再直接对评测工作进行支持，但我国相关领域的评测工作并没有停顿下来。一些主要的研究机构开始以各种形式自发地开展学术性的评测工作，而这些评测工作近年来越来越多地纳入了中文信息学会的名义之下。

中文信息学会评测

- 目前这些在中文信息学会名义下的评测有：
 - 机器翻译评测
 - 汉语处理评测
 - 情感计算评测
 -希望会有更多.....
- 这些评测与以往的**863**评测相比
 - 更加注重参评单位的自主性
 - 更加强调评测本身的学术性
 - 评测之后的技术交流更加深入

提纲

- 引言
- 回顾
- 展望
 - 存在的问题
 - 期待

语言资源建设的问题

- 具有汉语特色的工作还不多，这类工作大多还停留在词典层面（如汉语语法信息词典、知网）
- 语料库建设的大部分工作还在模仿西方的做法，以语言学理论为指导的大规模语料库建设仍然比较少见，影响也不大
 - 国外的工作如PennTreeBank、PropBank、FrameNet等，都有很强的西方语言学背景，未必适合于汉语
 - 国内这方面工作开始起步，如清华周强的工作、宋柔的工作
- 语言资源建设与算法模型研究仍然处于割裂状态，语言学家和计算机科学家的交流太难、太少

技术评测方面的问题

- 技术研讨仍然不够深入，还很少有从国内评测中诞生的新思想和新方法
（如陈毅东的基于多目标的词语对齐）
- 从适合国情的应用中提炼评测方面还做得不够
 - 既要有很强的应用背景
 - 国家大型活动（奥运、世博会）
 - 国家安全（互联网安全中心、军方需求）
 - 具有普遍意义的企业需求
 - 又要有很高的学术价值
 - 不能沦为产品测试
 - 学术意义明确→可能开启新的研究领域（如信息抽取）

中文自然语言处理的科学问题

Open Question #1

- 汉语的句法分析

- 现象:

- 在类似的宾州树库上，句法分析性能比英文低**10%**
 - 在类似的语义标注语料库上，汉语语义标注性能比英文低更多

- 问题:

- 现有的基于树库的汉语句法分析方法是否适合于汉语？
 - 也许汉语的句法分析是否应该更多的利用语义篇章知识，而不仅仅是词性标记信息？
 - 如何将更多的语义篇章知识引入的汉语句法分析中？

中文自然语言处理的科学问题

Open Question #2

- 机器翻译

- 现象:

- 在国际NIST机器翻译评测中，在类似语料规模和测试环境下，阿英翻译的性能（BLEU值）比汉英翻译高15%左右
 - 基于句法的统计机器翻译模型在汉英翻译中取得的效果比阿英翻译更加显著
 - 与直观感觉似乎相反，统计机器翻译在英汉翻译上的效果似乎比汉英翻译更糟糕→汉语生成比理解更难？

- 问题:

- 涉及汉语的机器翻译似乎需要依赖更多的句法语义知识
 - 如何将更多的句法语义知识引入汉外和外汉机器翻译中？

提纲

- 引言
- 回顾
- 展望
 - 问题
 - 期待

期待

- 在汉语语言资源建设与技术评测的推动下，能够解决中文自然语言处理面临的主要科学问题，这有赖于：
 - 先进的语言学理论应用于大规模语言资源建设
 - 适合于先进语言学理论的算法模型
 - 语言学家和计算机科学家的密切合作
- 国内的自然语言处理技术评测：
 - 更多更深入的技术交流碰撞出更多创新的思想
 - 更多地面向我国国民经济建设和国家安全的需要，从中提炼出既有应用价值，又有学术意义的评测任务

本专题安排

- 16:10-16:20 刘 群：中文语言资源建设与评测的进展与展望
- 16:20-16:30 陶建华：ChineseLDC进展
- 16:30-16:40 何婷婷：国家语言资源监测语料库介绍
- 16:40-16:50 俞士汶：北京大学综合语言知识库
- 16:50-17:00 刘奕群：海量规模中文网络信息检索评测语料库的设计与实现
- 17:00-17:10 赵 军：信息检索和情感分析评测情况介绍
- 17:10-17:20 周 强：中文信息学会句法分析评测介绍
- 17:20-17:30 邹嘉彦：香港城大泛华语共时同题语料库介绍
- 17:30-17:40 易绵竹：基于本体语义学的语言资源观
- 17:40-18:00 **自由发言**

向希望发言但未能安排的同行致歉！欢迎大家踊跃发言！

谢谢！

