中文信息处理战略研讨会

中文信息处理的技术突破口

刘群

江西婺源

2012-4-14

中文信息处理已经取得的突破

* 显示排版：CCDOS、UCDOS、方正、华光
* 汉字输入：联想、五笔、微软、搜狗、亚伟
* 语音合成：讯飞
* 语音识别：讯飞、自动化所、声学所
* 文字识别：汉王
* 中文检索：百度、搜狗、TRS
* 词语切分：ICTCLAS……
* 句法分析：？

与其他语言处理的差距

* 句法分析
* 语义角色标注
* 机器翻译
* 问答

下一个突破口在哪里？

* 句法？（技术层面）
* 语义？（技术层面）
* 机器翻译？（应用层面）
* 自动问答？（应用层面）

* 核心：某种语义！（绕过句法、直达某种语义）

* 机器翻译和自动问答的成熟都将依赖于语义问题的解决！

语义表示应该是怎样的？

* 词义
  * Ontology? （wordnet）
  * Hownet?
* 结构义
  * 依存? （句法到语义的过渡？）
  * 语义依存? （刘挺）
  * FrameNet?
  * 语义角色?
  * 逻辑?
  * 话题层次（宋柔）
  * 情境?

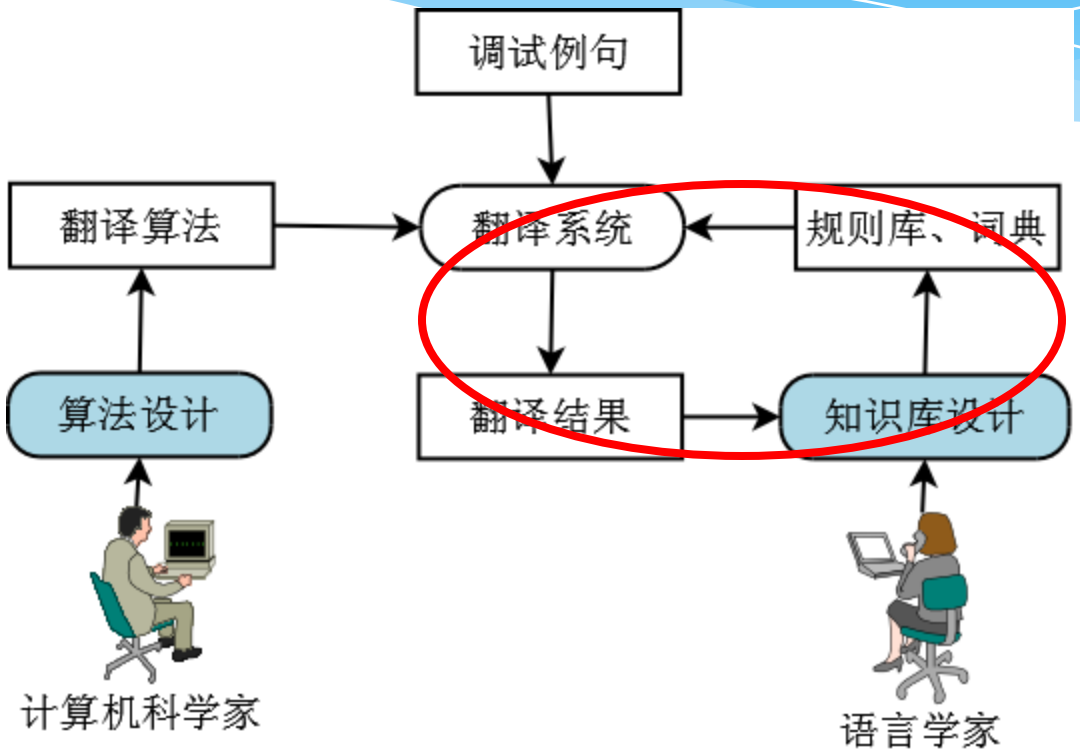
语义资源建设

* 资源为王?
* 是不是有足够的资源就能解决汉语语义问题?
* 语义资源建设的投入应该投向何方?
* 众包方式如何应用?

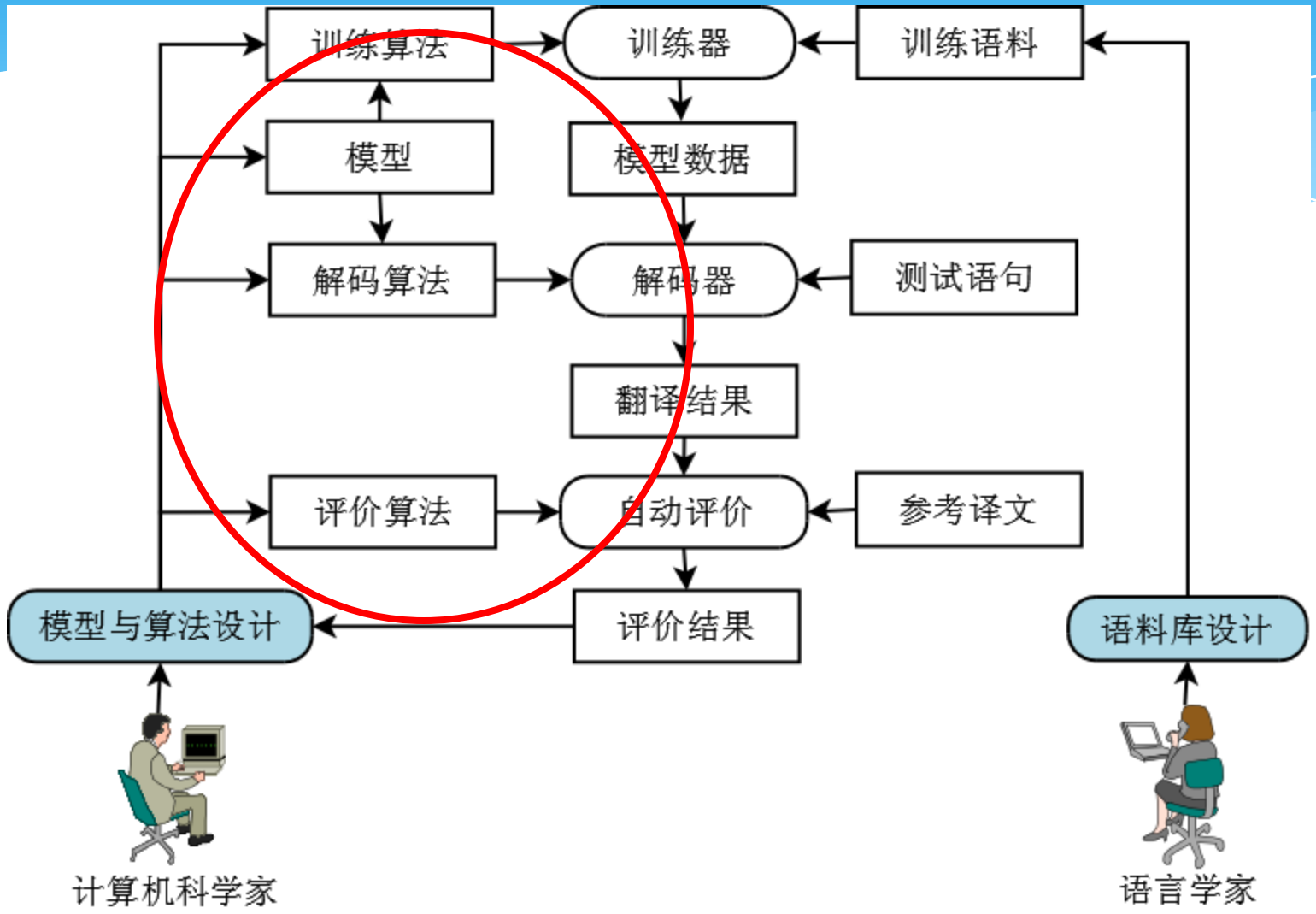
语义分析算法

* 训练
  * 有监督
  * 无监督
  * 半监督
* 分析

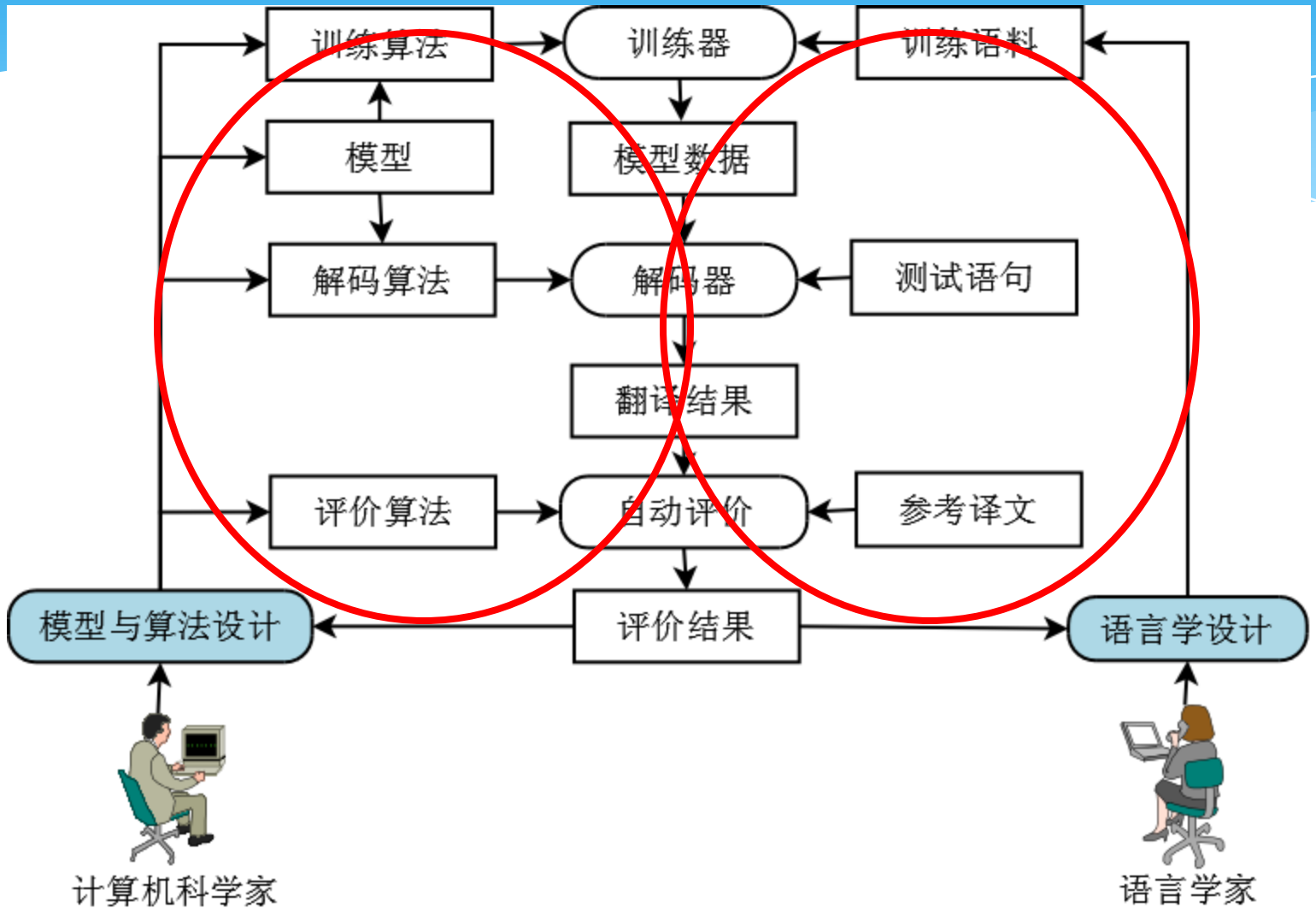
基于规则的机器翻译研究范式



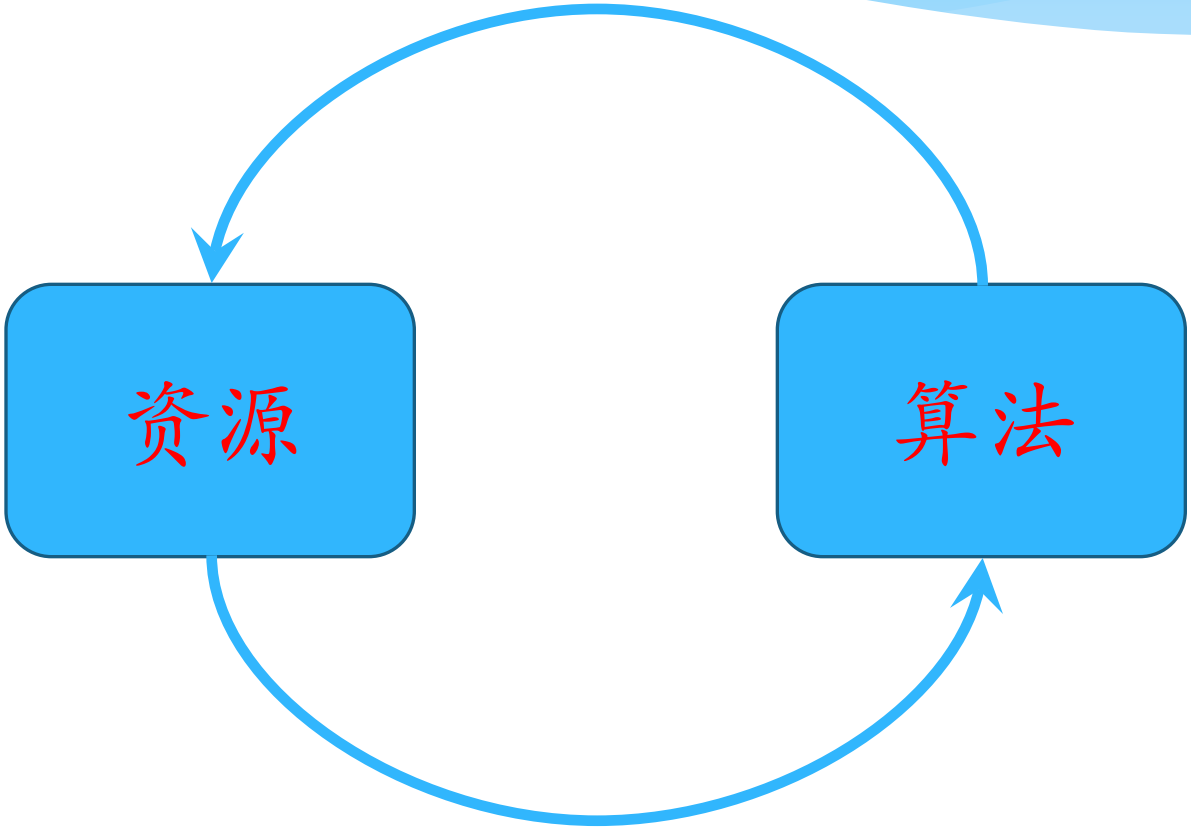
统计机器翻译研究范式


训练算法
训练器
训练语料
模型
模型数据
解码算法
解码器
测试语句
翻译结果
评价算法
自动评价
参考译文
模型与算法设计
评价结果
语料库设计
计算机科学家
语言学家

规则与统计结合的新研究范式



统计为主还是资源为王？



建议

* 资源建设：各种语言理论百花齐放
  * Hownet
  * Wordnet（北大CCD）
  * Framenet（刘开瑛）
  * 配价词典（北大）
  * 语义角色标注（SRL）
  * 语义依存（刘挺）
  * 话题层次（宋柔）

建议

* 评测引导
  * 客观任务：不依赖于理论
  * 理论pk

建议

* 语义资源建设与评测需要进行顶层设计

- 组织召集资源与评测系列会议

谢谢！