

# 块驱动的汉语句法语义分析

周强

清华大学信息技术研究院  
语音和语言技术中心

CIPS-2008 研讨会，北京，2008-11-24

# 研究动机

- 基本假设：句子理解是块识别和块内、块间词汇关联性分析交互作用的过程
- 我们设想的句子内容理解过程：
  - 确定句子的基本信息单元
    - 实体、动作、性状、属性、数量...
  - 把握句子的事件描述单元，形成事件骨架树（块间关系）
    - 不同层次的事件描述：小句、从句
    - 中心成分分析：主、谓、宾 ...
  - 明确骨架成分中各个基本单元之间的内在联系（块内关系）
    - 实体-属性、实体-数量、动作-形状 ...
  - 形成句子内容的完整理解（基本单元 → 事件骨架树）
    - ‘骨架’ + ‘血肉’
- 这里的基本信息单元和事件描述单元就表现为不同层次的块

# 句子理解实例

- 句子：现在，友谊汽车服务公司的1000多辆出租车都备有一把伞。
  - 基本单元：友谊汽车服务公司、1000多辆、出租车、一把、伞
  - 事件骨架：出租车-备有-伞
  - 单元关系：友谊汽车服务公司 – 出租车、1000多辆 - 出租车、一把 - 伞
  - 完整理解：
    - 备有 (拥有事件)
      - 拥有者=出租车(归属=友谊汽车服务公司, 数量= 1000多辆),
      - 拥有物=伞(数量=一把),
      - 时间=现在

# 研究切入点

- 问题：
  - 能否找到有效的方法自动识别出这些块？
  - 能否有效计算这些块内和块间的词汇关联性？
- 解决方案：
  - 建立分层次汉语块描述体系
  - 构建大规模的块标注库
  - 开发有效的块分析器
  - 构建词汇概念网络
    - 词汇关联度计算和词汇关联性分析
  - 开发块驱动的汉语句法语义一体化分析器

# 汉语块描述体系设计

- 基本认识：块是句法语义信息的结合体
  - 内部的词语关联性是句法语义联系的桥梁
- 分层次的块描述体系
  - 基本块：形成句子的基本信息单元
    - 比词更大的信息单元：宏观调控、手机银行、参观图书馆
    - 可能是人脑中组织信息的基本单元？
    - 词库中静态的字/词进入动态组块成句的重要桥梁
  - 功能块：形成句子的事件描述单元
    - 自顶向下的中心成分分析：主-状-谓-宾-补-定-中心语
    - 形成事件骨架树的基本单元
      - 小句层面：主-状-谓-宾-补
      - 从句层面：主-状-谓-宾-补-定-中心语

# 不同层次的汉语块标注库

- 以清华汉语树库TCT为基础资源
  - 自动提取不同层次的块标注库
- 基本块标注库
  - 多个实义词按照特定句法关系（定中、状中、并列、述宾、述补）聚合形成的信息单元
  - 句法成分：名词块、数量块、时间块、空间块、动词块、形容词块、副词块
  - 平均块长度：1.4个词
- 功能块标注库
  - 占据特定句式的不同功能位置的信息单元块
  - 功能标记：主语、谓语、宾语、状语、补语、定语、中心语
  - 平均块长度：2.3个词
- 基本块 → 功能块：90%可以简单对应
  - 1:1 --- 76%，N:1 --- 14% (一层规则对应)

# 块分析器开发进展

- 基本数据集：TCT的所有新闻类文本
  - 总规模：约20万词，80%训练，20%测试
- 基本块分析器
  - CRF模型：局部语境词语+词类+组合特征
  - 开放测试F-measure达到90%左右
- 功能块分析器
  - CRF模型：局部语境词语+词类+组合特征
  - 开放测试F-measure达到85%左右
- 分析难点
  - ‘+v’基本块：巴基斯坦流亡总统，淮河排污整治行动
  - 复杂从句边界：定语从句，主宾语从句
- 进一步性能提升需要引入更多的词汇关联性知识
  - 最重要：‘V-N’，‘N-N’

# 词汇关联性描述

- 词汇关联性 (Lexical Relatedness)
  - 实义词汇之间可能形成的各种概念联系
- 主要关联性类别
  - 语义关联性：同义/反义、属性-值-主体、事件-角色 ...
  - 句法关联性：定中、状中、述宾、述补、主谓 ...
  - 主题关联性：医生-护士-病人-医院（医疗主题、小桥-流水-河岸（空间主题）
- 关键问题：能否形成一个可计算、可扩展的词汇关联性快速计算模型？

# 词汇概念网络开发

- 基本假设：
  - 各种关系融合形成的图结构关系网络可以提供全局性的词汇关联性计算知识
- 图结构的词汇关联关系融合体
  - 三个节点层次：词语、义项、语义类
  - 关系边：{<N<sub>1</sub>, N<sub>2</sub>, relation, probability>}
- 构建思路：
  - 融合现有语言资源，形成基础词汇概念网络
  - 利用块分析技术，不断学习补充新的可靠句法关联和主题关联对

# 词汇概念网络：目前状态

- 数据来源：
  - 提取知网、同义词词林的语义关联对；
  - 提取TCT和现代汉语辞海的句法关联对；
- 总规模：
  - 词语节点-90752，义项节点-108362，语义类节点-2560，关系边总数约100万
- 正在进行相关的词汇关联度计算研究，为句法结构排歧、词语义项排歧和语义角色标注提供有力支持

谢谢关注!

Q & A