



# 中文信息处理热点问题讨论 总结

---

2008-11-24



# 两个话题

---

- 汉语分析中句法和语义的关系问题
- 自然语言处理与信息检索的关系问题

# 汉语分析中句法和语义的关系 问题



---



# 问题的提出

---

- 汉语缺乏形式上的标记，带有很强的意合色彩，因此有学者认为汉语可以省去句法分析阶段直接进行语义分析。
- 在汉语分析中，句法和语义究竟是怎样的关系，语义分析一定要以句法分析为基础吗？语义能够引导句法分析的过程吗？句法和语义能够相互融合与协同，使汉语分析技术取得新的突破吗？



# 发言者

---

- 董振东，HOWNET
- 宋柔，书面发言，北京语言大学
- 姬东鸿，武汉大学
- 靳光瑾，教育部
- 张全，中科院声学所
- 袁毓林，北大
- 李茹，山西大学
- 周强，清华大学



# 董振东

---

- 中国人对于词性是不敏感的，这与西方语言的使用者是很不同的。中文的词性标记体系是不成体系的。
- 中国人对于意义及其组合规律是很敏感的，甚至是不学自通的。
- 汉语的语法特性是：以意义为基础、以语序以及虚词的运用为主要语法手段、以音律为辅助语法手段的。



# 董振东

---

## 做什么？

- 做以意义为基本单元的中文短语的结构和意义组合的自动识别；
- 做中文的结构歧义和词语的多义歧义的“定点清除”；
- 继续开发基于知网的意义计算的工具包，包括机译系统；



# 宋柔

---

- 汉语的句法结构没有明确的定义。句法结构的标注不但计算机难以自动实现，人工完成也难以规范。
- 就计算机处理来说，其实无所谓句法和语义，关键是形式化、算法化、能算化。
  - 形式化就是要用封闭有限的符号串表示各个要素及要素间的关系
  - 算法化就是要制定出一个无歧义的、可终止的、覆盖各种可能情况的操作流程
  - 能算化就是算法复杂度不高，在可控范围内
- 现在许多人在谈语义和语用。关键问题是能否设计出一个理论体系，满足上述三点要求，而且计算结果符合汉语实际。



# 姬东鸿

---

- 汉语缺乏句法标记，但不缺乏语义标记；
- 直接总结语义规则很难，但可从资源中学习规则：
  - **传统做法：** 较小语言单位组合成较大语言单位的规律
  - 较小语言单位的语义结构组合成较大语言单位的语义结构的规律



# 张全

---

- 问题：将汉语句法和语义分析结合起来
- 解决方案：建立语言感知的处理模式（HNC）
  - 形成概念基元体系和句类体系，或者称语义基元空间和句类空间。
  - 通过自然语言符号体系映射成为HNC概念关联符号体系的方式将自然语言中的语义内容显式的表示出来，完成自然语言语句的理解，为各种语言信息处理提供基础。
  - 目前HNC已经形成了面向语句理解处理的句类分析处理技术，这一技术基于知识和规则，已经应用于实际的信息处理系统中。



# 周强

## 基本语言假设：

- 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
- 语言分析的目标：
  - 找到有效的方法自动识别出这些块？
  - 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



# 靳光瑾

- 句法结构上的关系映射了语义关系，在汉语的分析中句法、语义能够相互融合与协同
- 以动词“剪”为例说明了句法语义分析如何融合



# 李茹

## 坚持构建汉语框架网

- 继续构建了300个框架，研究汉语多义词和高频词语框架表示方式
- 展开汉语框架语义角色基础上的句义理解计算模型研究；
  - 提出了一种用层叠条件随机场模型进行汉语框架元素自动标注方法；
  - 分析框架和框架之间的概念关系及框架到框架的推理机制；
- 进行网络文本语料库框架语义深加工，以《中国分类主题词表》为纲，构建领域本体库，研发基于句义理解的信息检索实验系统，期望通过框架语义角色的分析使汉语分析技术取得新的突破。



# 袁毓林

---

- 提出文本蕴涵计算的词汇解决方案
- 研究目的：为问答系统提供识别同义表达形式的理论、方法和语言知识资源
- 技术路线：
  - 立足于词汇层面，以动词为核心，分别语义情境类型，来控制问题的规模；
  - 充分利用已经建设成的谓词-论元结构知识库；
  - 通过浅层的句法-语义分析，把句子之间的蕴涵关系简化和落实到句子中谓词之间的蕴涵关系；暂时不考虑论元之间的蕴涵关系。



## 小结

---

- 语义研究在汉语信息处理中占有重要地位
- 应该加强汉语语言学研究 and 计算语言学研究两方面的研究



# 自然语言处理与信息检索的 关系问题

---



# 问题的提出

---

- 随着搜索引擎的成功，自然语言处理也焕发了新的生机，但在信息检索界不少学者对自然语言处理技术对信息检索的贡献表示质疑，认为自然语言处理只能在问答这样精准化的检索问题上发挥作用，而在大规模网页的检索、分类、过滤等课题上由于自然语言处理技术精度不高、速度较慢，因此往往帮不上忙。
- 那么，是否自然语言处理真的只是点缀？从长远看，自然语言处理能够对信息检索起到更为重要的作用吗？



# 发言者

---

- 马少平，清华
- 洪涛，百度
- 王斌，中科院计算所
- 孙乐，中科院软件所
- 李涓子，清华
- 孙斌，北大
- 赵军，中科院自动化所



# 马少平

---

- 长期以来，研究者们不断尝试着在狭义的信息检索任务（文档检索）中使用自然语言处理，然而结果并不能令人满意。
  - 复杂性较低的基本自然语言处理技术（包括去除停止词、分词、取词根等），计算消耗小，简单易行，虽然那对信息检索的帮助很小，仍然是在信息检索实验平台中推荐使用的；
  - 复杂性高的高级自然语言处理技术，包括句法分析、短语识别、命名实体识别、概念抽取、指代消解和词义消歧等，计算消耗大，精度不高，对信息检索基本没有帮助。
- 原因
  - 自然语言处理技术的精度不够高，存在错误，即便有一些积极影响也会被消极影响所掩盖；
  - 不使用自然语言处理的方法（例如使用统计方法），之所以能够大幅度地提高检索效果的原因是它其实已经蕴含了语言学的知识，并且它解决的问题是相对容易的那些，剩下留给自然语言处理的都是难得多的。



# 马少平

---

- 自然语言处理在广义的信息检索任务（例如问答系统、信息抽取）中已经发挥了很大的作用。这些将是未来自然语言处理应用的发展方向之一。
  - 为了使NLP技术更好地应用在信息检索任务中，要针对信息检索任务优化NLP技术
  - 自然语言处理与信息检索有效融合而成的统一模型，也将是研究者们关注的重点。在基于语言模型的IR中加入自然语言处理获得的成功可以看作是这一方向取得的成果。



# 洪涛

---

- 从产业的角度看，NLP一定对IR有大的帮助
- 百度招聘了很多有NLP背景的学生就说明了一点
- 索引的层级（词汇和短语的多层次标引）、广告分析等都需要用到NLP技术
- NLP和IR相互促进和融合



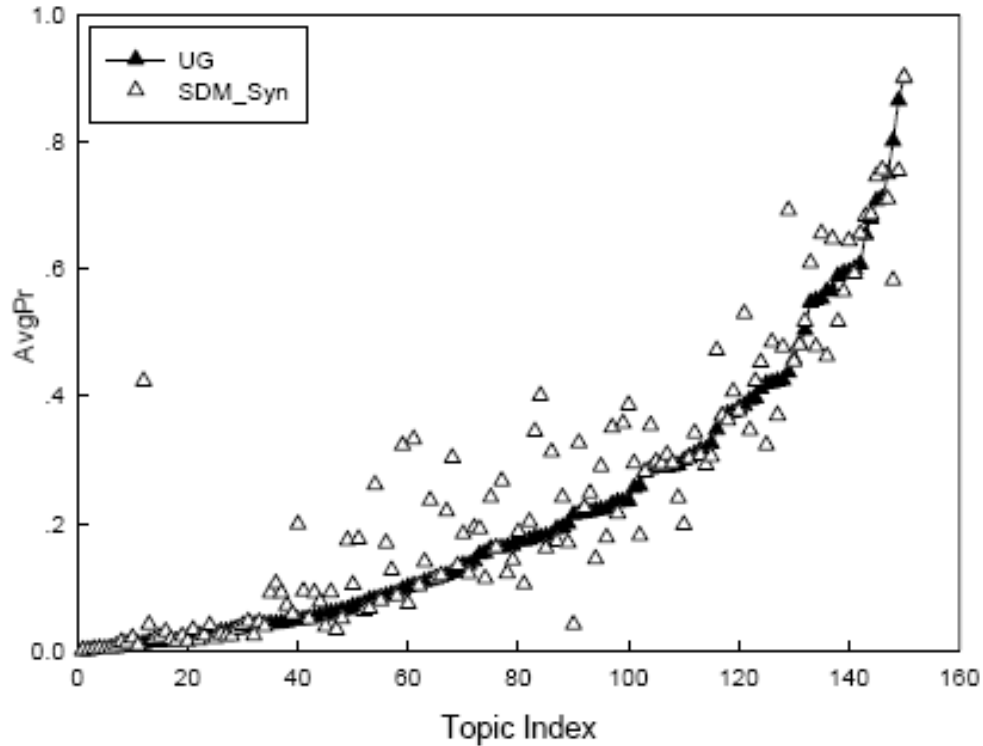
# 王斌：NLP和IR密不可分

---

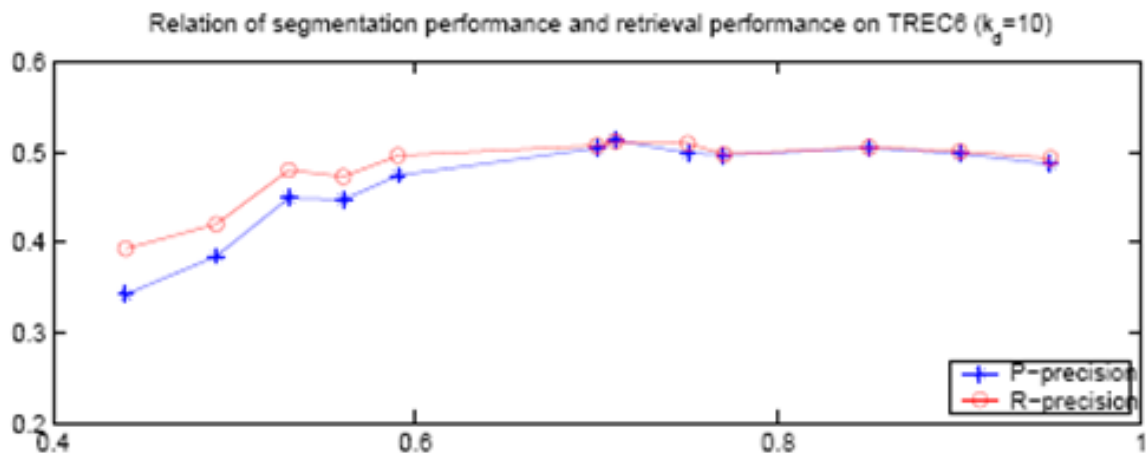
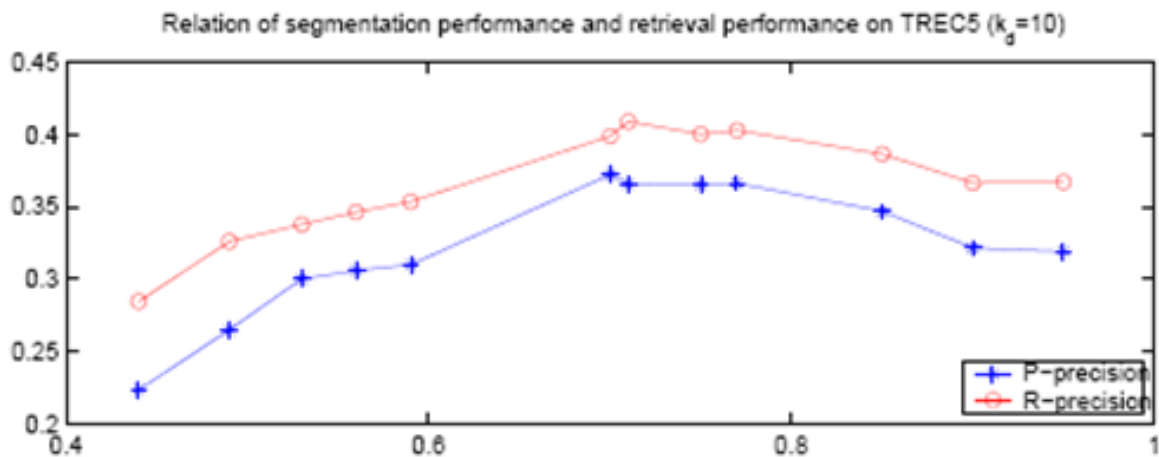
- NLP中的文本表示和文本处理是IR中的重要组成部分
- 一些(新的)应用是NLP和IR的融合产物
  - Question Answering (QA)
  - Information Extraction (IE)
  - Cross-language IR (CLIR)
  - Sentiment Analysis (SA)
  - Text Summarization (TS) –query specific TS
- 两个领域的技术互相引用
  - Language Modeling—from NLP to IR
  - Maximum Entropy —from NLP to IR
  - Vector Space Model—from IR to NLP
  - Kernel based similarity—from IR to NLP
  - .....

# 王斌：句法分析&IR

- 针对IR方法 句法分析IR语言检索效果

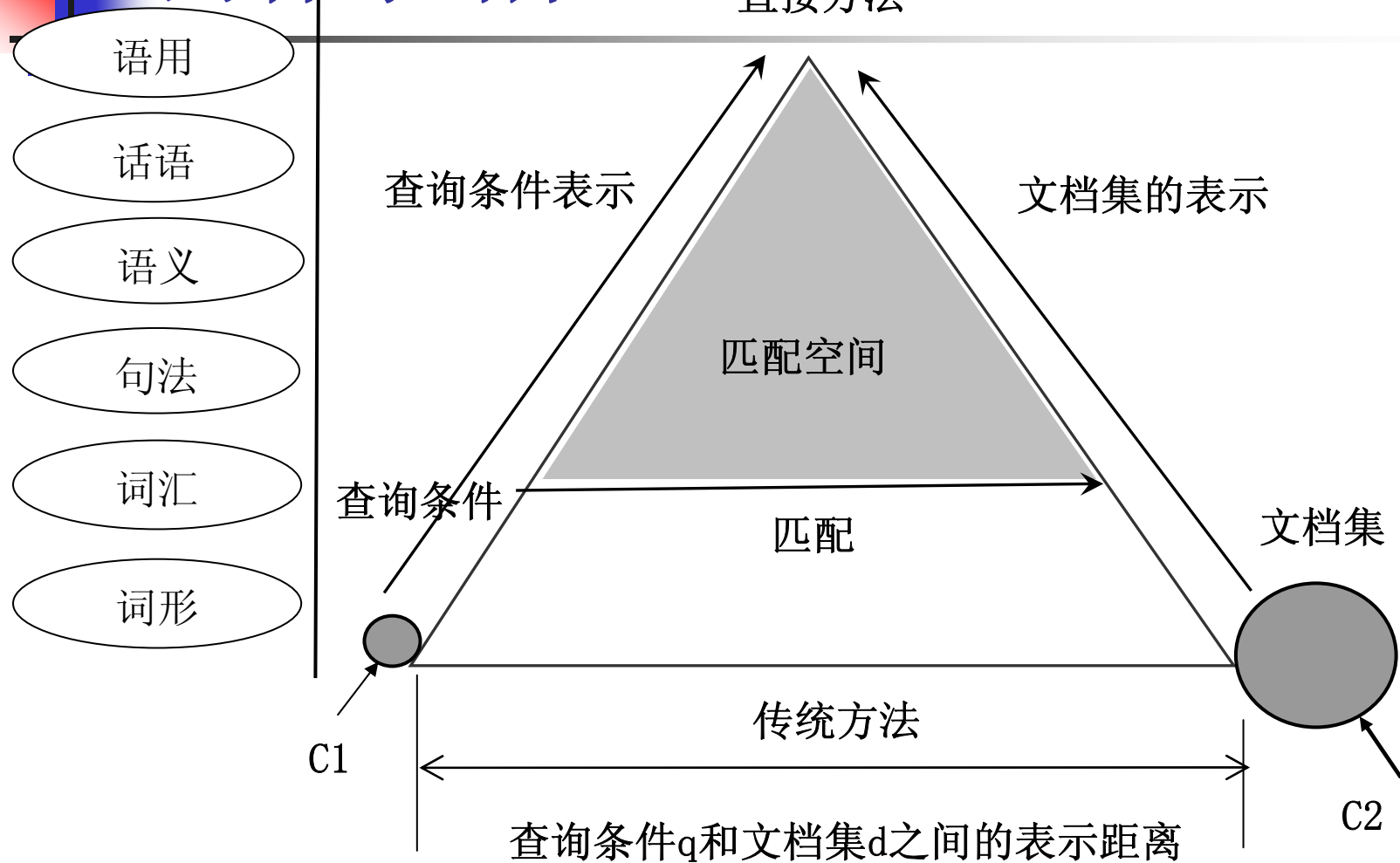


# 王斌：要研究面向IR的NLP



中文分词和检索性能的关系图

# 孙乐：NLP技术与IR相融合的体系结构



# 孙乐

文本表示是**NLP**与**IR**相融合的关键技术之一，通过在文本表示中引入**NLP**技术可以减小查询与文档集之间的匹配空间

- 目前文本表示方法中使用的要么是局部信息，要么是全局信息，这两者的结合有可能使一些深层次的**NLP**技术引入**IR**
- 随着计算机硬件性能的不断提高，新的**IR**模型将会出现，从文本表示来看，现有的**SVM IR**，**LM IR**模型将有可能被以**LDA Model**为代表的图模型所取代

# 李涓子：以POWERSET为例说明 NLP在语义搜索中的应用

The screenshot displays the PowerSet search engine interface. The search query is "Tim Berners-Lee". The results are categorized into "Wikipedia Articles" and "World Wide Web".

Annotations on the screenshot include:

- Text introduction**: A green speech bubble pointing to the introductory text of the "Tim Berners-Lee" article.
- Structured person information**: A green speech bubble pointing to the structured data fields such as "Date of Birth" and "Place of Birth".
- Extracted Facts about Tim**: A green speech bubble pointing to the "Facts from Wikipedia" section, which lists key facts like "invented: HTTP".
- Wikipedia Documents**: A green speech bubble pointing to the list of search results under "Wikipedia Articles", including "Tim Berners-Lee", "World Wide Web", and "History of the World Wide Web".



# 李涓子：为什么语义搜索需要NLP

---

- 语义元数据的自动生成
- 网页信息的语义标注
- 实体的关联——Web of Entity

# 孙斌: NLP研究和Web IR应用将在水平上逐步重叠——

- 为什么自然语言处理的长期研究努力和结果没有在实际的信息检索应用中产生直接和显著的作用？特别是当前的大众化网络信息搜索应用似乎还没有为自然语言处理研究提供很大的用武之地
  - 首先是这些应用的技术层次还不高，利用简单的字符串匹配和统计、链接分析和数据库管理就可以满足大多数信息检索的要求（包括字面匹配、图像和音视频文件名称等）。
  - 其次是NLP技术的应用还不够“低”，一直着重于某些“小众化”的课题，没有把大众的日常信息需求当作重点研究内容。认为：未来大众信息检索的应用层次将逐步提高：结构化内容、问答式检索等-----从狭义的信息检索走向广义的信息检索
  - 另一方面，NLP研究界和产业界会把越来越多的日常生活需求当作严肃的研究和技术开发课题。
  - 当二者的重合达到某种临界程度，就会看到二者的相互促进，形成良性反馈的局面。



# 赵军：IR能为NLP做什么？

---

- NLP领域常用的统计方法对于小概率事件的预测无能为力
- 对于以上问题的处理，海量的、冗余的网络信息的有效挖掘可以提供新的手段
- 从以命名实体翻译为例，说明如何利用网络挖掘技术辅助翻译。



# 结论

---

- 目前，NLP对IR已经起到一定的帮助作用，但还不够明显
- 未来，NLP将通过文本表示、信息抽取、检索模型等对IR起到重要而全面的推动作用



谢谢

---

2008-11-25