

## 1. 陈小荷：先秦文献的信息处理

“先秦汉语词汇统计与知识检索”项目，准备对 25 种最重要的先秦传世文献进行词语切分、词性标注、个别常用词（包括古今字和通假字）的词义标注，建立先秦书面汉语（以下简称先秦汉语）的词汇知识库以及先秦文献的历史知识库并研制相应的检索系统。研究现代汉语的计算机处理技术在先秦汉语处理中的应用方法。

### (1) 词汇研究

- 先秦汉语有多少个不同的词，分别表示哪些概念？词语切分，做词表，同形词区分。在词的认定上可不可以找到某种起辅助作用的统计标准？
- 先秦汉语中单字词和多字词的使用频率是怎样的，组合型歧义消解。
- 将先秦汉语跟现代汉语的词汇统计比较。
- 从概念表达的角度来观察古今词汇对应关系。
- 从概念表达的角度来观察古今词汇替换现象。
- 从概念表达的角度来观察古今词义变化。全面标注古今字、通假字，以确定同一个字在不同概念上的频率。

### (2) 基于内容的计算

- 根据先秦文献建造历史事件、人名、地名三个知识库。
- 对事件、人物和地点进行相关性度量，人名、地名的“同名异指”和“异名同指”排歧。
- 进行事件、人名和地名等历史知识的交叉检索，为大众服务。

## 2. 袁毓林：文本蕴涵的类型和识别机制——从认知假设到计算模型

### (1) 信息检索碰到的难题

- a. Who is Ariel Sharon?
- b. Israel's Prime Minister, Ariel Sharon, visited Prague. —*T*
- c. Ariel Sharon is Israel's Prime Minister. —*H*
- 对于给定的问句 a 来说，文本 b 蕴涵了句子性假设 c；因此，b 可以作为 a 的答句。

### (2) 文本蕴涵的定义

在自然语言中，相同的意义可以用不同的形式来表达，或者从不同的文本中推断出来。可以把这种不同文本之间的同义和蕴涵关系系统称为文本蕴涵（textual entailment）。

### (3) 词汇层面的蕴涵的推理机制

- a. Yahoo bought Overture.
- b. Yahoo own Overture.
- 可以从中演绎下面这种蕴涵型式：  
$$X \leftarrow \text{subj buy obj} \rightarrow Y \Rightarrow X \leftarrow \text{subj own obj} \rightarrow Y$$

并把 Yahoo 和 Overture 作为支撑点。

#### (4) 词汇-句法层面的蕴涵的推理机制

- 在词汇-句法层面上，可以假定：文本  $T$  和假设  $H$  都是由一组可通过依存关系分析而得到的句法依存关系来表达的。
- 于是，如果  $H$  中的关系可以被  $T$  中的关系所覆盖，那么定义  $T$  和  $H$  之间具有蕴涵关系。

#### (5) 基于依存树匹配关系的识别机制

- 工作原理：寻找文本和假设的依存关系分析树之间的匹配关系。
- “文本-假设”对之间是否具有蕴涵关系，是根据它们之间的相似性来决定的。而这种相似性被定义为假设中能够跟文本匹配上的节点的比例。
- 经过试验，发现这种相似性的阈值是 50%。也就是说，当假设中能够跟文本匹配上的节点的比例达到或超过 50%时，可以说它们之间具有蕴涵关系；当这种比例小于 50%时，可以说它们之间不具有蕴涵关系。

#### (6) 利用谓词-论元结构的识别机制

- 工作原理：用简单的一般性的启发式和知识贫乏的方法来识别同义互释，用 NP 同指互参、NP 语块切分、RASP 和 Link 两个分析器来给“文本-假设”对中的每一个句子产生 (PAS)。
- 然后，用 WordNet 词汇链和一些专门的启发式规则来建立这些 PAS 中相应成分的语义相似性；
- 最后，为这些相应的 PAS 的结构相似性和相应词汇成分的相似性设定阈值，用以判断“文本-假设”对之间的蕴涵关系是否成立。

### 3. 宋柔：中文信息处理对汉语基础研究的需求

(1) 汉语标点句间句法关系的几何性质。回答问题：汉语标点句常常缺失成分，而代词又很少用，汉语的话语凭借什么机制来理解和生成？基础研究，希望将来能为应用研究作贡献。

(2) 汉字字形的形式化描述和计算，用于汉字教学中的错字、古籍中变形字等汉字的保真输入和分析。可以使计算机的标准字符集大大缩小，而带有怪字的文本可以输入、显示、打印、存储、编辑、传输、交流。

(3) 交搭模型替代传统的隐马模型做词性标注。利用自然语言前后文紧密联系所形成的相互约束，能提高准确性和效率。

### 4. 盛玉麒

#### 一、中文信息处理是语言教学研究的 new 领域

1. 人机系统不同于人际系统
2. 新问题和新的挑战
3. 新手段和新机遇

#### 二、汉语教学的新领域

1. 二语习得与电脑理解汉语
2. 电脑中文化与中文电脑化
3. 中文数字化与中文标准化
4. 资源开发与共享平台建设

### 三、汉语研究的新课题

1. 数据库
2. 知识库
3. 规则库
4. 方法库

### 四、基于语料库的汉语知识挖掘

1. 方兴未艾的语料库语言学
2. 资源性的认识与开发
3. 知识挖掘与语料库加工
4. 共建共享避免低水平重复劳动

### 5. 一些想法

#### (1) 自然语言处理的要求

- 面向**真实**文本和环境
- 正视**现实**，调查现实，尊重现实，承认现实，从语言现实出发进行研究，提出符合现实状态的语言理论和知识。
- 解决**实用**中的问题

#### (2) 计算机学科与语言学科必须深入交叉融合

- **计算机科学不可能充当自然语言处理的救世主。**计算机工作者不要以为语言学知识没有用，以为计算机用统计方法能解决全部问题。离开语言知识，数学方法和计算机技术不可能奏效。也不要以为现成的语言学理论和知识加上统计方法就够了。
- **语言学不可能充当自然语言处理的救世主。**语言学工作者不要以为现有的理论体系是天经地义的，也不要以为建立在现有理论体系下的一些语言知识可以解决很多自然语言处理的应用问题。目前的语言理论不能准确反映汉语现实，不能形式化、算法化，难以在自然语言处理中发挥效用。
- **计算机工作者要从事语言研究**（具有形式化、算法化的训练，对需求有深入理解）
- **语言学工作者要从事自然语言处理应用系统的研制**（具有语言辨析、语言现象归纳的训练，有活的语料库）
- **计算语言学工作者应当做原创性的研究**，这是职业责任。不能仅仅依靠原有的理论和别人的方法做语言实验室的实验员或语言资源库的标注工。

天：原创理论

空：发现规律/发明方法                      资源加工，方法使用

地：语言现实/实用需求

