自然语言理解

内容回顾

No.95, Zhongguancun East Road
Beijing 100080, China

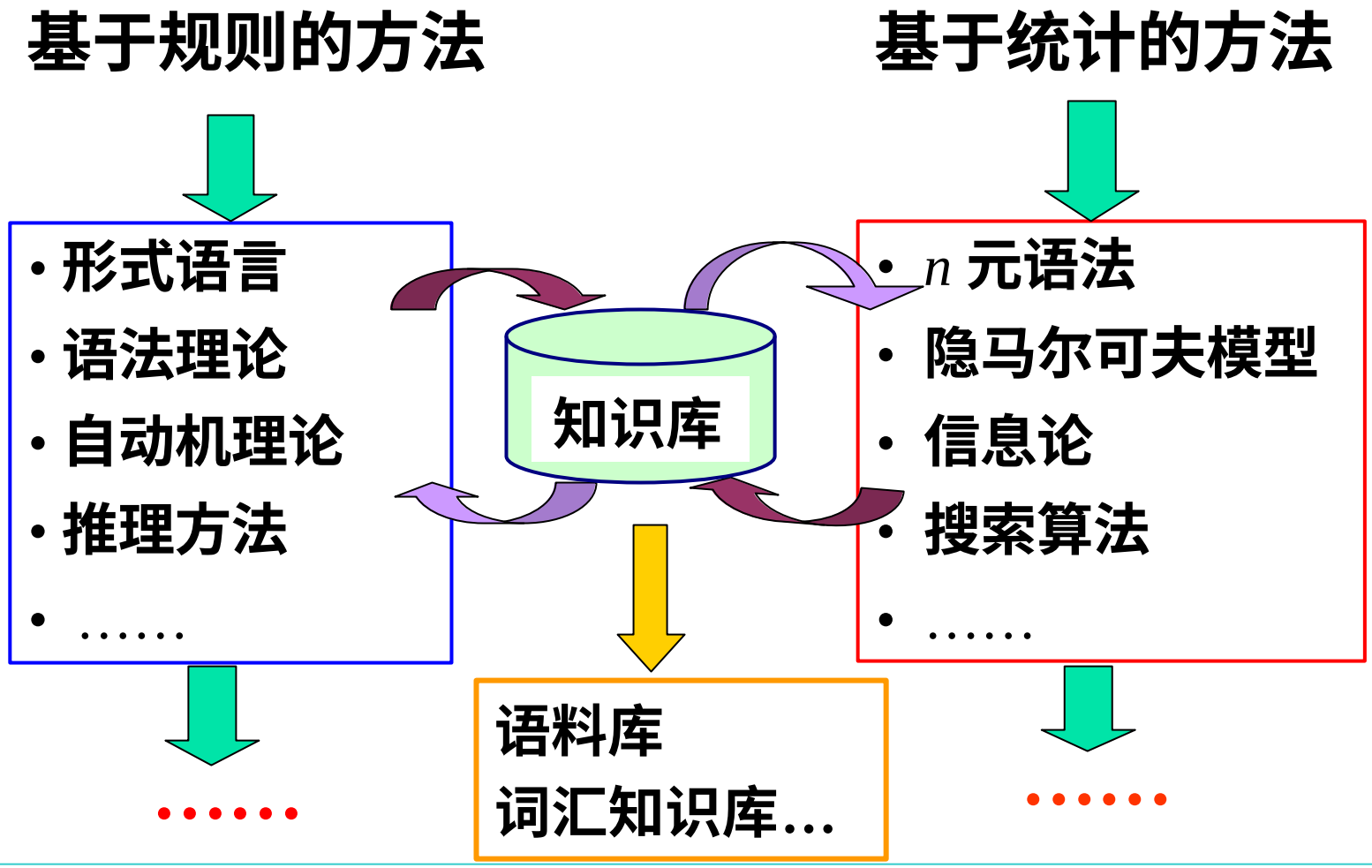
http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

1、目标

了解自然语言处理的理论和方法，实现相应的自然语言处理系统。

- 机器翻译
- 信息检索
- 文本分类
- 人机对话
- ……

2、系统实现方法


基于规则的方法
基于统计的方法
• 形式语言
• 语法理论
• 自动机理论
• 推理方法
• ……
知识库
• n 元语法
• 隐马尔可夫模型
• 信息论
• 搜索算法
• ……
语料库
词汇知识库…

3、例子

理解句子：我们有一家三星级酒店。

问题："理解"的标准是什么？

- **就句子本身描述的内容让系统回答问题：**
  - **我们有什么？**
  - **谁有酒店？**
  - **有几家酒店？**
  - **几星级酒店？等等**
- **用特定的语义表示形式描述：**
  - **框架表示；**
  - **其它语义表示形式**

IF 表示

IF (Interchange Format) **表示：**

- **说话人标志**(Speaker),**表示说话人的身份**
- **语句意图**(SpeechAct),**代表说话者的意图**
- **概念**(Concept),**表示句子的主题**
- **参数**(Argument),**表示句子的具体内容**

IF: a:give-information+existence+accommodation

(experiencer=we,
accommodation-spec=(quantity=1,(accommodation-class=three_star)))

问题与方法

实现过程中的问题：

（1）汉语分词与词性标注问题：

- 最大分词方法
- 统计方法
- 有限自动机方法

（2）短语结构、句法结构分析

- 规则方法
- 统计方法

问题与方法

（3）语义的分析和归并问题：

- 规则方法
- 统计方法
- 有限自动机方法

（4）知识库和基础语料：

- 词典
- 语义概念库
- 规则库
- 标注语料

基本思路

➢ 基于概念语块的统计解析方法

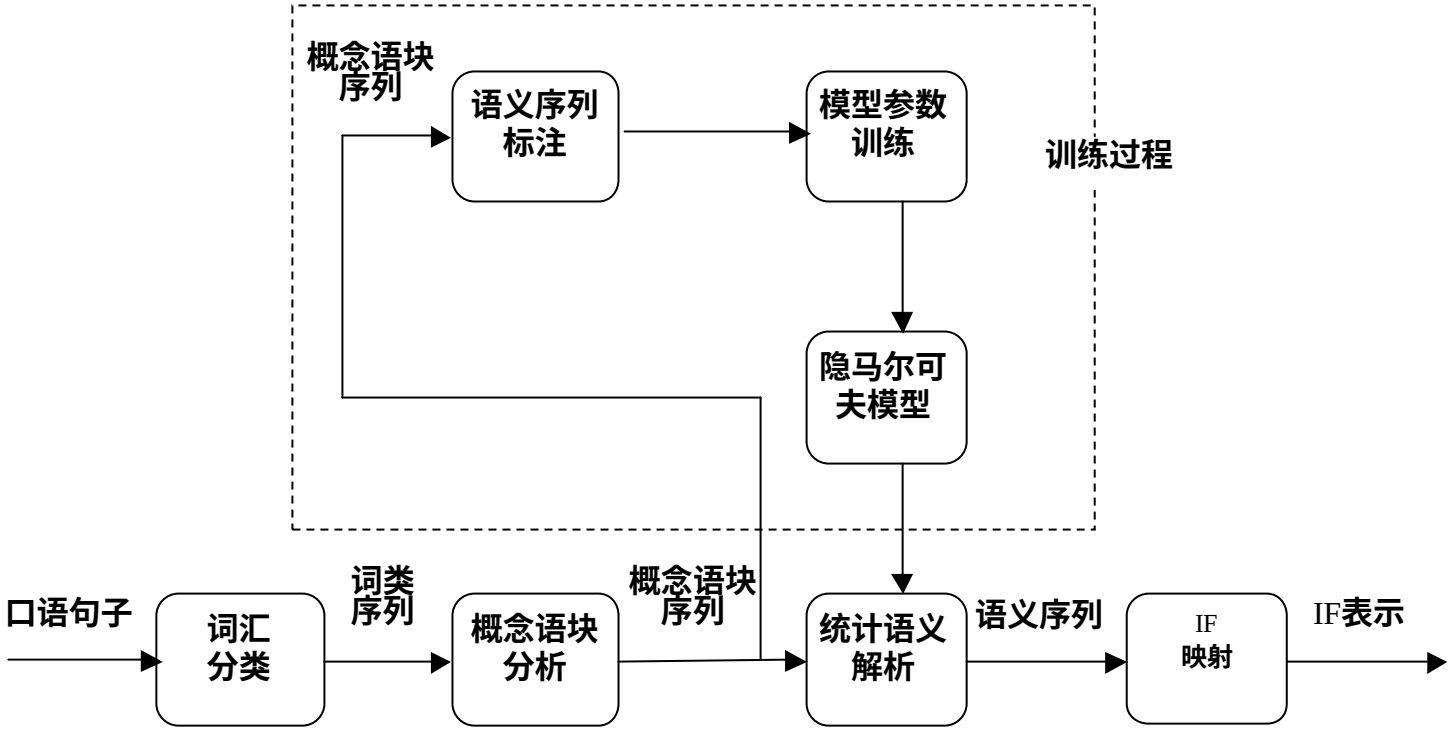
- ✓ 概念语块的定义和分析
- ✓ 概念语块到IF片段的转换
- ✓ 通过HMM,把概念语块序列解析IF表示
- ✓ 统计口语解析模型HMM的改进

基于概念语块的统计解析方法

- 基本思想
  - 利用规则对句子进行概念语块分析，并且得到概念语块的内部层次结构,在此基础上利用HMM对句子进行解析
- 特点
  - 在于对句子进行深层次语义分析的同时，保持了统计方法较高的鲁棒性

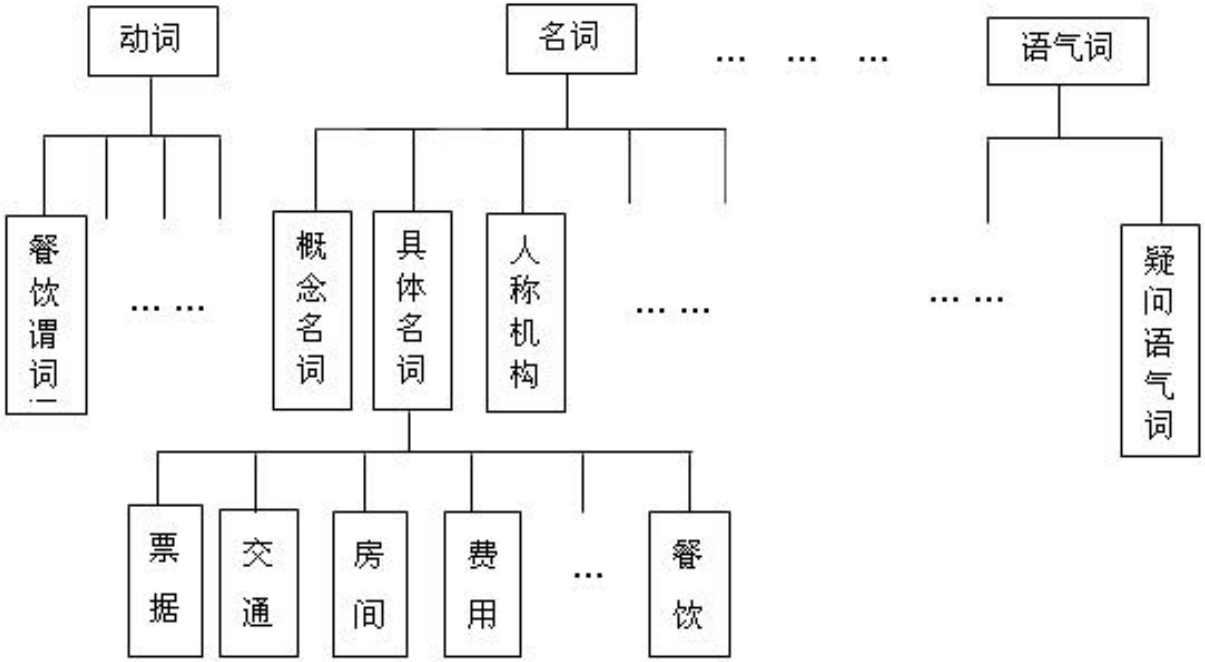
基于概念语块的统计解析方法

- 基本结构
  - 训练和解析两部分



基于概念语块的统计解析方法

- 词汇分类
  - 根据词汇的语法特征和语义功能，在词性标注的基础上对词汇进一步分类
  - 最底层的小类一共有324个



基于概念语块的统计解析方法

- **对于一些表示具体意义的实词, 给出了这些词汇对应的IF值**
  - **比如“星期五”对应的IF值为“friday”,单人间对应的IF值为single等**

词类和词类包含的词汇

词类	词类包含的词汇
N_C_LANGUAGE	**汉语**:name-chinese; **英语**:name-english; **阿拉伯语**:name-arabic; **西班牙语**:name-catalan
N_O_CITY	**北京**:name-beijing; **香港**:name-hongkong; **深圳**:name-shenzhen; **上海**:name-shanghai
N_C_WEEK	**星期一**: Monday; **星期二**:Tuesday; **星期三**:Wednesday; **星期四**:Thursday
N_C_NAME	**姓名 名字 全名 大名**
V_INCLUDE	**包括 带有 加上 加**

基于概念语块的统计解析方法

- 规则的概念语块分析方法
  - 概念语块的定义
    - 句子中不依赖于其它词汇而能表示限定领域内的某种概念的最长部分
      - “我想预订一个单人间中”的“我”、“想”、“预订”、“一个单人间”
      - “我想定一个人住的”中的“一个人住的”
    - 表示同一个概念的语块被划分为一类
      - 所有能够表示“某个具体的人”的概念语块,都归属于PERSON类概念语块,比如“你”、“你们”、“他们”、“我的朋友”等
    - 270种概念语块

基于概念语块的统计解析方法

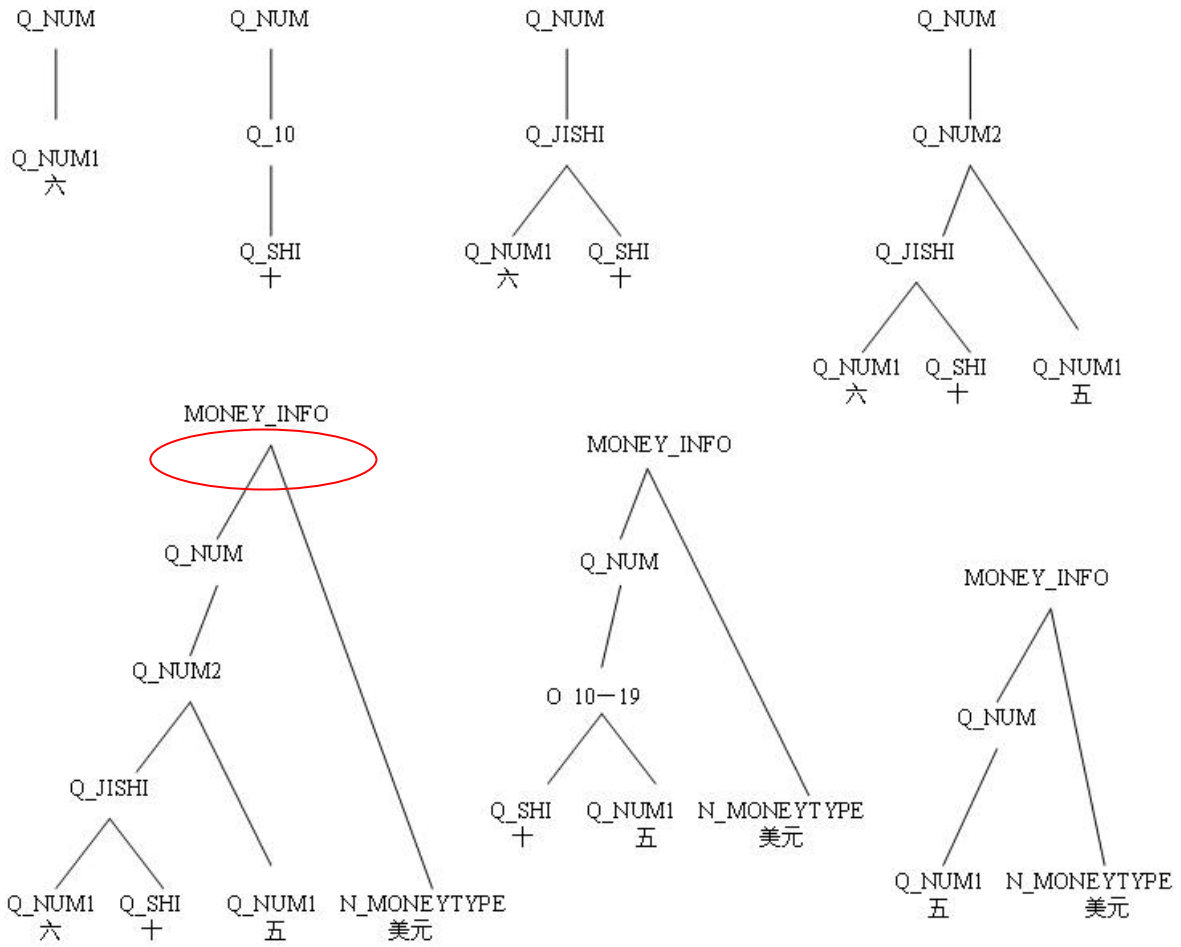
- **分析器为线图分析**(chart-parsing)
- **文法为标准的**PCFG
  - **规则描述的内容是词类或者概念语块之间如何组成新的概念语块**

ROOM_INFO->Q_Q_QUAN ROOM_UNIT ROOM	*1*	***几间房***
ROOM_INFO->Q_NUM ROOM_UNIT	*1*	***两间***
ROOM_INFO->Q_Q_QUAN ROOM_UNIT	*1*	***几间***
ROOM_INFO->Q_NUM ROOM_UNIT ROOM_TYPE	*1*	***两个单人间***

- **分析结果是一系列候选节点，每个节点代表一个概念语块**
  - **节点的名称就是概念语块的类型，节点所覆盖的词汇即概念语块的内容**
  - **长度优先，概率优先**

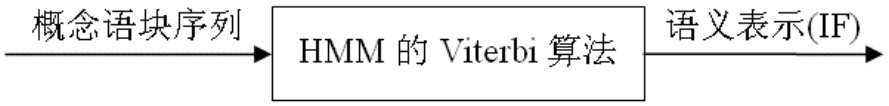
➢概念语块分析示例：

“价格是六十五美元”中的 “六十五美元”部分的解析结果



基于概念语块的统计解析方法

- 统计语义解析方法的核心－HMM
  - HMM包括4部分：状态S; 观察O; 状态转移概率矩阵A=a_{ij}; 从状态输出观察的概率分布矩阵B=$b_j(k)$; 初始状态概率分布向量π= π_i
  - 当利用HMM进行语义解析时
    - 模型的状态S相当于口语句子的语义(IF)
    - 模型的观察*O*相当于句子的概念语块序列
    - 解析的过程相等于：给定一个概念序列,如何选择最优的语义符号组合(IF)？→ Viterbi 搜索算法



➢HMM的Viterbi算法:

输入：观察序列$O_1O_2.....O_T$和 HMM 模型参数$\lambda=(A,B,\pi)$

输出：状态序列$S_1S_2.....S_T$

Step 1． 初始状态

$$\delta_1(i)=\pi_i b_i(o_1)，\qquad 1\le i\le N$$

$$\varphi_1(i)=0$$

Step 2． 递推计算

$$\delta_t(j)=\max_{1\le i<n}[\delta_{t-1}(j)a_{ij}]b_j(o_t)，\qquad 其中\ 2\le t\le T，\quad 1\le j\le N$$

$$\varphi_t(j)=\arg\max_{1\le i<n}[\delta_{t-1}(j)a_{ij}]b_j(o_t)，\qquad 其中\ 2\le t\le T，\quad 1\le j\le N$$

Step 3． 终结状态

$$P^*=\max_{1\le i\le n}[\delta_T(i)]$$

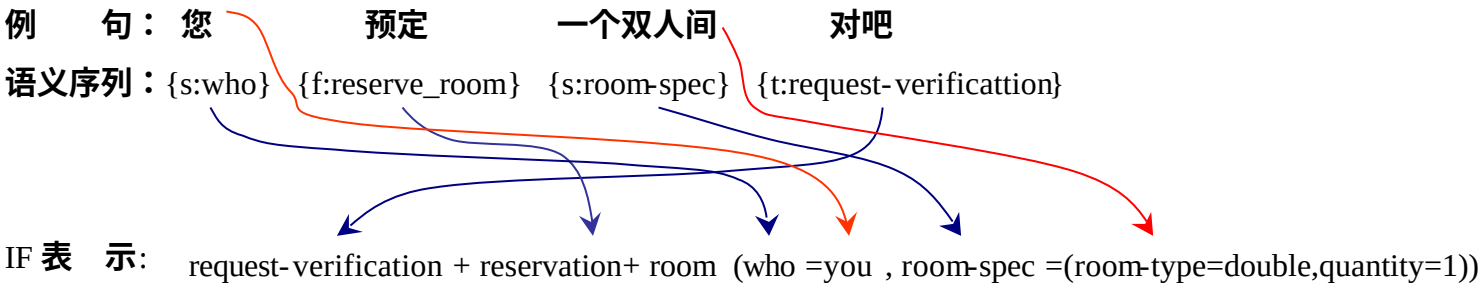
$$q_T^*=\arg\max_{1\le i\le n}[\delta_T(i)]$$

Step 4．路径回溯

$$q_t^*=\varphi_{t+1}(q_{t+1}^*)\qquad t=T-1,T-1,...1$$

基于概念语块的统计解析方法

- HMM 参数需要从语料中训练获得
- 统计模型得到的是一个线形化的IF，将其层次化



方法测试

- 正确率、召回率
- 与其它方法的比较
- 问题分析

Thanks

谢谢!