课程内容回顾

No.95, Zhongguancun East Road
Beijing 100080, China
http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

1 任务与目标

了解自然语言处理的理论和方法，设计并实现自然语言处理系统。

- 机器翻译
- 信息检索
- 文本分类
- 内容提取
- 人机对话
- …….

2 系统实现方法

基于规则的方法

- 形式语言
- 语法理论
- 推理方法
- ……

……

自动机方法

知识库

语料库
词汇知识库…

基于统计的方法

- *n* 元语法
- 隐马尔可夫模型
- 信息论
- 搜索算法
- ……

……

3 例子-1

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

3 例子-1

- 理解结果表示形式：
  - 框架表示
  - 其它语义表示形式

3 例子-1

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

a: give-information+existence+accommodation

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

3 例子-1

◆ 实现思路：

分析每个词的含义→ 词之间的相互关系→短语的识别→短语的含义→句子的含义

◆ 可采用的方法：

- 有限状态自动机方法
- 规则分析方法
- 统计方法

3 例子-1

◆方法确定：

➢ 采用规则和统计方法相结合的方法

给定句子S, 寻找中间表示 IF，那么，

$$\hat{I}F = \arg\max_{\hat{I}F}(IF \mid S)$$

3 例子-1

- 实现过程中的问题：
  - 汉语自动分词
  - 词义解释与消歧
  - 语块（短语）识别
  - 语义概念组合
  - 语义提取

3 例子-1

◆每个问题的处理方法：

- 汉语自动分词
  - 最大分词方法（FMM, BMM,…）
  - 最短路径方法
  - 基于HMM统计方法

… …

3 例子-1

- 词义解释与消歧

  —建立一部词典：格式＋内容

  #词条，词性，词义标记

  - 词条收集
    - 现有词典
    - 语料收集、提取
  - 词性划分与标记符号的确定
  - 语义划分
  - 标记符号

  词义区分：

  —规则方法

  —统计方法

3 例子-1

- 语块（短语）识别

[我们]$_{NP}$ [有]$_{VP}$ [一 家 三星级 酒店]$_{NP}$ 。 $_{PU}$

- 规则方法:
  - NP → NN
  - VP → VV
  - MP → NUM L
  - ……
- 统计方法：
  - HMM方法
  - 条件随机场
  - ……
- 自动机

3 例子-1

- 语义概念组合与语义提取
  - 规则方法
  - 统计方法
  - 自动机方法
  - 混合方法

3 例子-1

◆ 基于概念语块的统计解析方法

- 基本思想

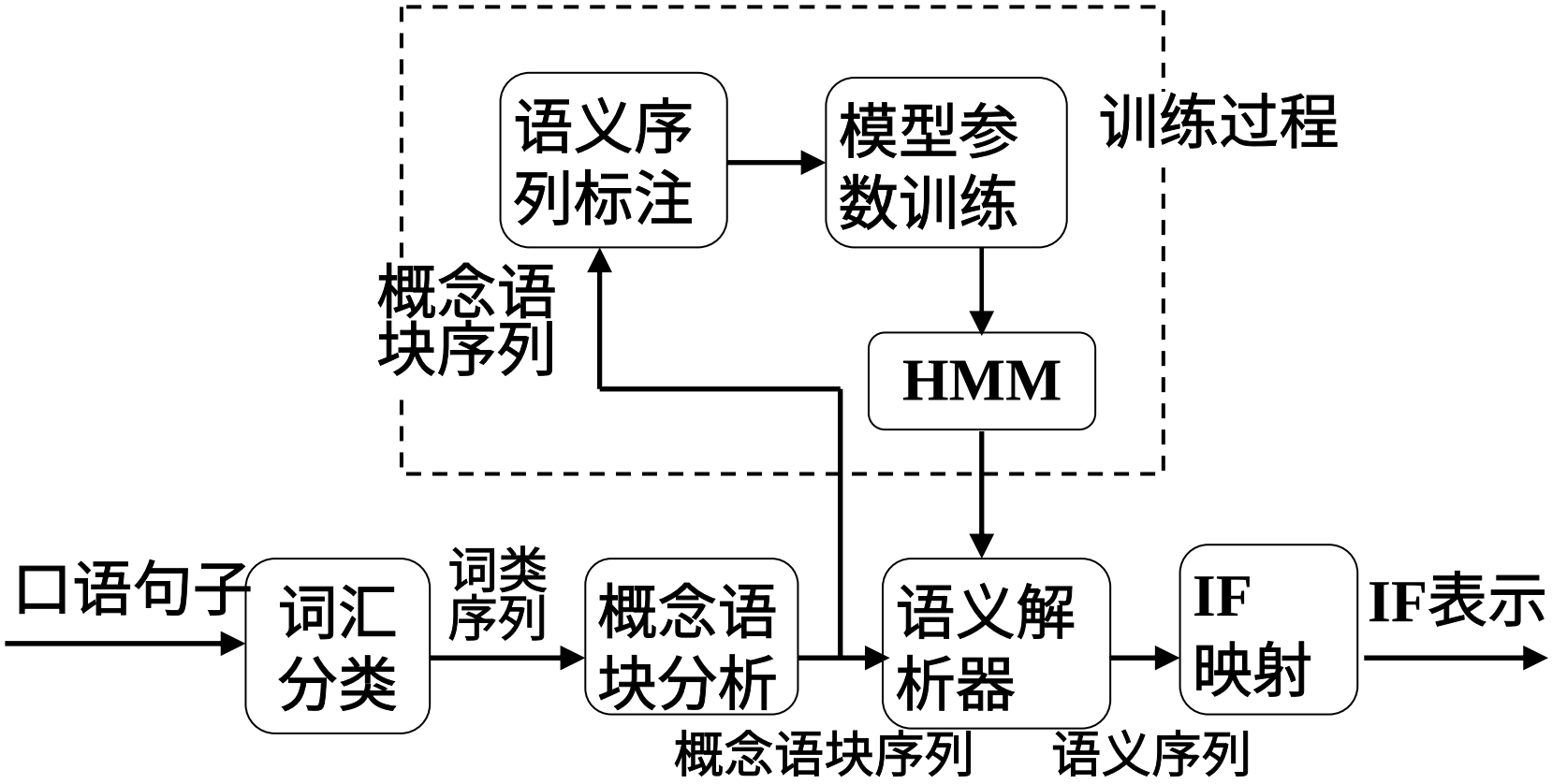
  利用规则对句子进行概念语块分析，并且得到概念语块的内部层次结构，在此基础上利用HMM对句子进行解析。

- 特点

  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。

3 例子-1

- 基本结构：一训练和解析两部分

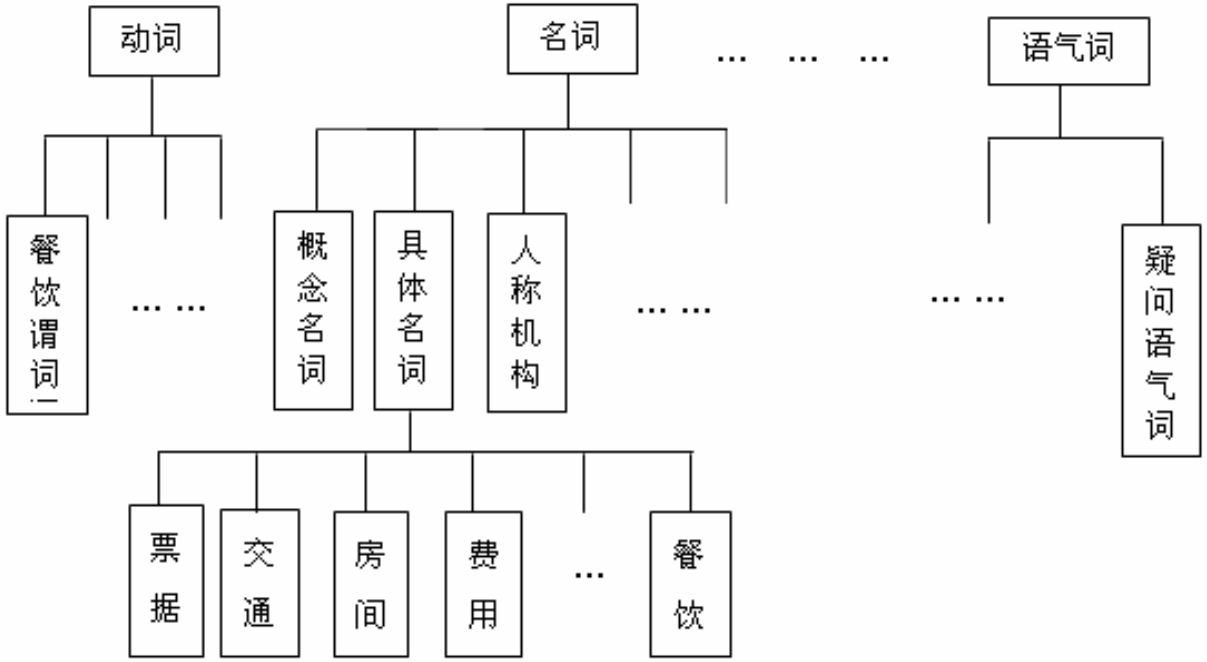


3 例子-1

- 主要工作
  - ✓ 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

3 例子-1

- 词汇语义分类
  - 根据词汇的语义功能，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。



3 例子-1

对于一些表示具体意义的实词, 给出这些词汇对应的IF表示的标记值，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	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

3 例子-1

- 概念语块的定义
  - 表示同一个概念的语块被划分为一类

    例如，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

3 例子-1

- 分析器为线图分析(chart-parsing)
- 文法 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

ACCOM_CLASS → *Q_NUM CLASS* *1* 五星级

ROOM_INFO → *Q_Q_QUAN ROOM_UNIT* *1* 几间

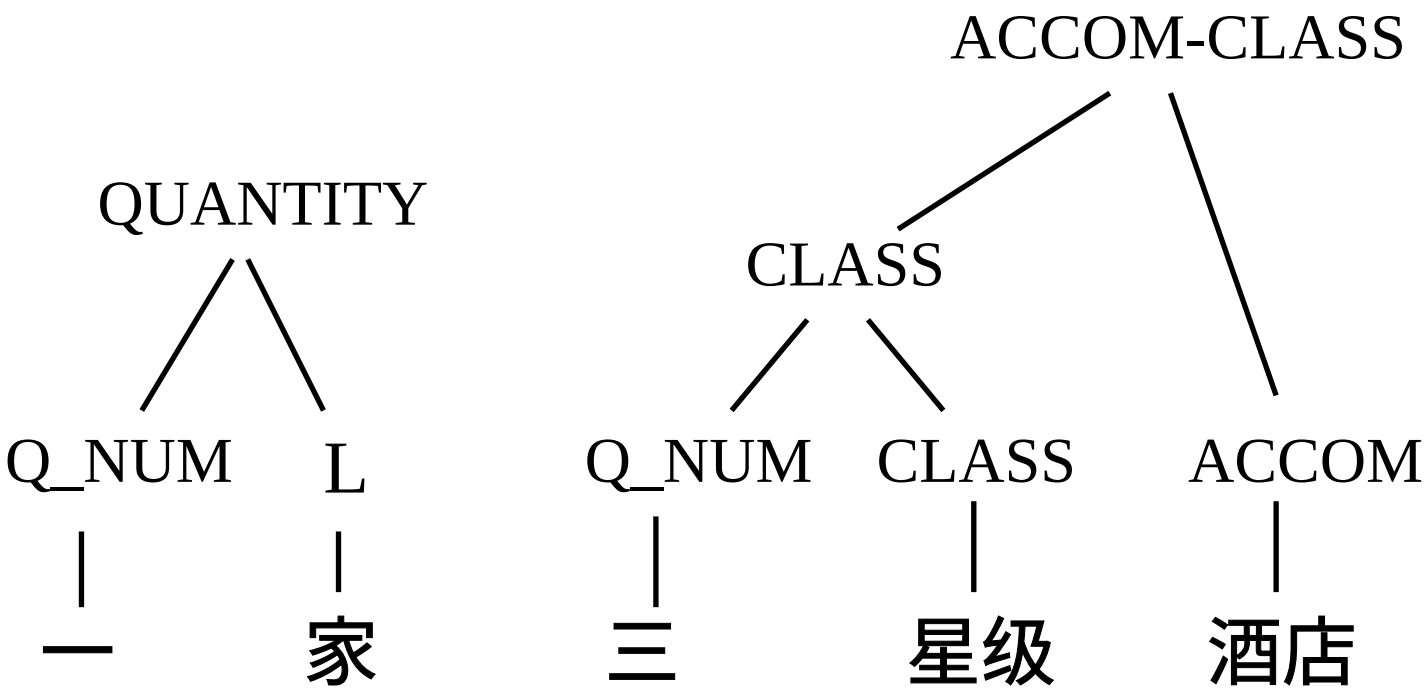
ROOM_INFO → *Q_NUM ROOM_UNIT ROOM_TYPE* *1*

两个单人间

3 例子-1

- 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

3 例子-1



3 例子-1

- 统计语义解析方法的核心－HMM
  - 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

3 例子-1

- HMM 参数需要从语料中训练获得
  - 输入：语块序列
  - 状态：IF的关键词线性序列
  - 观察：语块概念序列

a: give-information+existence+accommodation
(experiencer=we, accommodation-spec=
(quantity=1,
(accom-class=three_star)))

- 将线性的IF表示层次化处理

4 例子-2

- 构造一个汉英翻译系统

◆方法选择：

- 基于规则的方法
- 基于统计的方法
- 基于实例的方法

……

4 例子-2

◆基于规则的方法

- 汉英对照词典

  #汉语词, 词性(必要的语义信息), 英语对应词, …

- 汉语词法分析

  分词、时态分析、单复数分析

- 汉语句法规则

  NP → NN NN, p

  NP → ADJ NN, p

4 例子-2

- 汉语句法分析器
  - Chart parser
  - CYK parser ……
- 汉英句法分析树转换
  - 结构变换规则
- 英语句子生成
  - 词汇选择，时态变化，名词单复数变化,…

4 例子-2

◆基于统计的方法

- IBM 模型1-4？
- 基于短语的翻译模型(phrase-based)
- 基于句法的模型(syntax-based)

… …

4 例子-2

选择基于短语的翻译模型(phrase-based)

$$p(e|c) = p_T(c|e)^{\lambda_t} \times p_L(e)^{\lambda_l} \times p_D(e,c)^{\lambda_d}$$

- 获取英汉短语 $(c|e)$ 对及其概率
- 计算英语句子语言模型
- 解码器

4 例子-2

(1)语料准备和预处理

工具：EGYPT- TokenizeE.perl.tmpl / 汉语分词

(2)工具准备：

- 语言模型：SRI LM/ CMU-Cambridge LM
- 翻译模型工具：GIZA++结果→ 抽取短语对
- 解码器：Pharaoh / ReWrite

(3)系统调试、性能测试

Thanks

谢谢!