

第二章 乔布斯基形式文法

1. 词

是最小的能够独立运用的语言的单位

文法 (grammar) : 是对语言知识的规范化表示。是建立语言形式模型的基础, 它使采用**规范的规则体系**对**非规范**的自然语言做分析处理成为可能。

本章主要介绍 :

- **语法驱动的文法 (层次分析法)**

对句子结构层面分析, 理论工具形式语言

- **依存文法**

句结构由词间关系决定

- **格文法**

对句子语义层面分析, 分析句子语义成分

这四条公理相当于对依存图和依存树的形式约束为 :

- ◆ 单一父结点(single headed)
- ◆ 连通(connective)
- ◆ 无环(acyclic)
- ◆ 可投射(projective)

格文法 : 施事格, 受事格, 工具格

格语法

格语法有三部分组成 : 基本规则, 词汇部分 和 转换部分

下为乔布斯基形式文法:

用G表示形式语法，G定义为四元组：

$$G = (V_n, V_t, S, P)$$

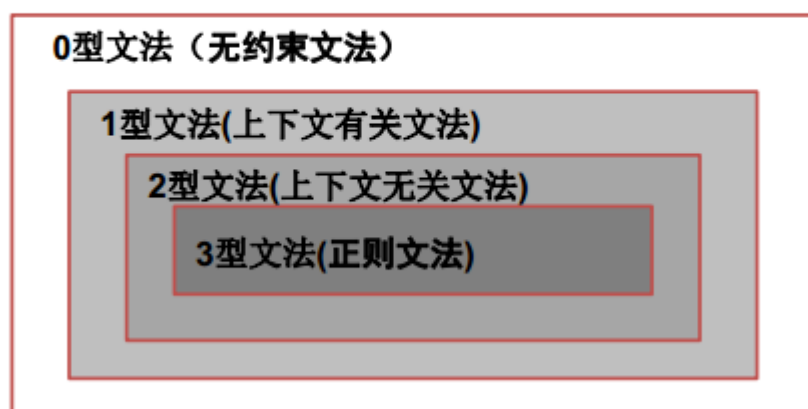
V_n ：非终结符(non-terminals)的有限集合，不能处于生成过程的终点，即在实际句子中不出现。在推导中起变量作用，相当于语言中的语法范畴。

V_t ：终结符(terminals)的有限集合，只处于生成过程的终点，是句子中实际出现的符号，相当于单词表。

S ： V_n 中的初始符号，相当于语法范畴中的句子。

P ：重写规则(rewriting rules)，又称生成规则(production rules)，一般形式为 $\alpha \rightarrow \beta$ ，其中 α 和 β 都是符号串，至少含有 V_n 中的一个符号。

Chomsky根据重写规则的形式，把形式语法分为4级：



第五章 语言模型

对语句合理性判断：

规则法：判断是否合乎语法、语义（定性分析）

统计法：通过可能性（概率）的大小来判断（定量计算）

主要来自解决语音识别问题

语言模型是用来计算一个句子概率的概率模型

统计自然语言处理的基础模型

2. n 元文法(n-gram)

n 元文法(n-gram) : 一个词由前面的 $n-1$ 个词决定

❖ 当 $n=1$ 时, 即出现在第 i 位上的基元 w_i 独立于历史。一元文法

也被写为 uni-gram 或 monogram ;

❖ 当 $n=2$ 时, 2-gram (bi-gram) 被称为1阶马尔可夫链 ;

❖ 当 $n=3$ 时, 3-gram(tri-gram)被称为2阶马尔可夫链, 依次类推。

参数估计 (模型训练): 获得模型中所有的条件概率 (**模型参数**)

用最大似然估计计算参数

语言模型对于训练文本的类型、主题和风格等都十分敏感

1. 数据平滑的基本思想 :

调整最大似然估计的概率值,使零概率增值,使非零概率下调,

“**劫富济贫**”, 消除零概率, 改进模型的整体正确率。

2. 数据平滑方法 :

◆ **加1法(Additive smoothing)**

◆ **减值法/折扣法 (Discounting)**

- 1) Good-Turing 2) Back-off (Katz)
- 3) 绝对减值(H. Ney) 4) 线性减值

◆ **删除减值法 : 低阶代替高阶**

评价：实用方法和理论方法（困惑度）

改进：基于缓存的语言模型、基于混合方法的语言模型

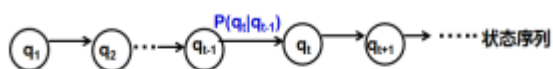
第六章 概率图模型

- **产生式模型**：由数据学习联合分布 $P(X, Y)$ ，然后以求出的条件概率分布 $P(Y|X)$ 作为预测模型，即生成模型：

$$P(Y|X) = \frac{P(X, Y)}{P(X)} \quad \text{如HMM, BNs, MRF.}$$

- **判别式模型**：由数据直接学习决策函数 $f(X)$ 或条件概率分布 $P(Y|X)$ 作为预测模型，即判别模型，关注的是对于给定的输入 X ，应该预测什么样的输出 Y 。如SVM, CRF, MEMM。

马尔可夫模型



$$p(S_0, S_1, \dots, S_T) = \prod_{t=1}^T p(S_t|S_{t-1}) p(S_0)$$

模型输入：状态序列

模型输出：状态序列的概率值

模型参数： $P(q_t|q_{t-1})$

马尔可夫模型又可视为**随机有限状态自动机**，该有限状态自动机的每一个状态转换过程都有一个相应的概率，该概率表示自动机采用这一状态转换的可能性。

◆ **隐马尔可夫模型作用**：

输入：观察序列

输出：观察序列的概率值

隐状态序列

参数： $P(q_t|q_{t-1})$, $P(O_t|q_t)$

HMM的三个假设

对于一个随机事件，有一观察值序列： $O=O_1, O_2, \dots, O_T$

该事件隐含着—个状态序列： $Q = q_1, q_2, \dots, q_T$.

假设1：马尔可夫性假设（状态构成—阶马尔可夫链）

$$P(q_i|q_{i-1} \dots q_1) = P(q_i|q_{i-1})$$

假设2：不动性假设（状态与具体时间无关）

$$P(q_{i+1}|q_i) = P(q_{j+1}|q_j), \text{ 对任意 } i, j \text{ 成立}$$

假设3：输出独立性假设（输出仅与当前状态有关）

$$p(O_1, \dots, O_T | q_1, \dots, q_T) = \prod p(O_t | q_t)$$

第七章 词法分析

词法分析任务：

将句子转换成**词序列**并标记句子中的词的**词性**。


- **英文的词法分析（曲折语）**
 - 英文词识别、词形还原
 - 未登录词处理
 - 英文词性标注
- **中文的词法分析（孤立语）**
 - 分词
 - 未登录词识别
 - 词性标注

英文词法分析基本任务：

1. 单词识别(Tokenization)
2. 词形还原(Lemmatization)
3. 词性标注：POS (Part-of-Speech) Tagging

中文自动分词

涉及问题：

1. 分词标准（切到什么粒度？）
 2. 切分歧义问题
 3. 自动分词算法
 4. 未登录词处理
- 

分词基本原则：切分、合并原则

切分歧义：交集型歧义、组合型歧义

自动分词算法：基于规则的方法[正向、逆向最大匹配]、基于统计的方法[全切分、最少分词法]

词性 (part-of-speech) 是词汇的基本语法属性，通常称为**词类**

一般词性标注集应遵守以下原则：

- **标准性**: 普遍使用和认可的分类标准和符号集；
- **兼容性**: 与已有资源标记尽量一致，或可转换；
- **可扩展性**：扩充或修改。

◆ 词性标注方法：

1. 基于规则的词性标注方法
2. 基于统计模型的词性标注方法
3. 基于错误驱动的机器学习方法
4. 规则和统计方法相结合的词性标注方法
5. 基于神经网络的词性标注方法

词法分析评价：P\R\F

两种测试

- 封闭测试 / 开放测试
- 专项测试 / 总体测试

第八章 句法分析

完全句法分析算法

PCFG(概率无关上下文文法)

计算分析树概率的基本假设

- **位置不变性**：子树的概率与位置无关，即对于任意的 k , $p(A_{k(k+C)} \rightarrow w)$ 一样。
- **上下文无关性**：子树的概率与上下文无关，即 $p(A_{kl} \rightarrow w)$ 任何超出 k 和 l 的上下文无关。
- **祖先无关性**：子树的概率与祖先无关。

分析算法有三种策略:

- ◆ 自底向上 (Bottom-up)
- ◆ 从上到下 (Top-down)
- ◆ 从上到下和从下到上结合

1. 线图分析法 (Chart) [节点和边组成]

活动边集(ActiveArc) 、线图 Chart (非活动边)

代理表(待处理表)(Agenda)、输入缓冲区

Chart parsing 算法评价

◆ 优点 :

- 算法简单, 容易实现, 开发周期短。

◆ 弱点 :

- 算法效率低, 时间复杂度为 Kn^3 ;
- 需要高质量的规则, 分析结果与规则质量密切相关;
- 难以区分歧义结构。

2. CYK 分析算法

(CYK) 算法思想 : 通过构造识别矩阵进行分析

◆ CYK 算法的评价

➤ 优点

- 简单易行, 执行效率高

➤ 弱点

- 必须对文法进行范式化处理
- 无法区分歧义

3. 完全句法分析评估

P、R、F、词性标注准确率、交叉括号数和交叉准确率

局部句法分析任务 : $\left\{ \begin{array}{l} \text{①语块边界分析;} \\ \text{②语块之间的关系分析。} \end{array} \right.$ **特点 :** 不需要句法模型

浅层句法分析方法 (局部句法分析)

- (1) 基于规则的方法
- (2) 基于统计的方法
- (3) 统计和规则相结合的方法

随着语料库的不断完善，统计的方法越来越站主导地位，
在基于统计的方法中大部分方法是将组块的分析转化成**序列标注**问题。

基于最大熵的组块分析：定义组块、标签集、训练语料处理、最大熵建模、
语料和特征模板训练模型、应用模型求解

依存句法分析

目前依存句法分析主要是 统计依存句法分析方法

统计依存句法分析要素

- 1.语料-中文依存树库
- 2.统计算法

统计依存句法分析方法

- ◆ 生成式的分析方法(generative parsing)
- ◆ 判别式的分析方法(discriminative parsing)
- ◆ 决策式的(确定性的)分析方法(deterministic parsing)

依存评估

无标记依存正确率(unlabeled attachment score, UA)

带标记依存正确率(labeled attachment score, LA)

依存正确率(dependency accuracy, DA)

根正确率(root accuracy, RA)

完全匹配率(complete match, CM)

第十章 篇章分析

篇章分析应用领域：

- 统计机器翻译(Statistical Machine Translation)
- 自动文摘(Text Summarization)、
- 自动问答系统(Question Answering System)
- 信息抽取(Information Extraction)
- 情感分析(Sentiment Analysis)
- 自然语言理解 (Natural Language understanding)
- 自然语言生成 (Natural Language Generation)

篇章概念

篇章：由一个以上的句子 (sentence) 或语段 (utterance) 构成的**有组织、有意义的**自然语言文本整体。一篇文章、一段会话等都可以看成篇章。构成篇章的句子 (或语段) 彼此之间在形式上相互衔接，在意义上前后连贯。

篇章结构：微观、宏观篇章结构



衔接性：强调构成成分 (主要是词或短语) 之间的形式的关联

连贯性：强调通过句子意义 (内容) 表示的关联

基于RST的篇章结构分析主要包括**两个子任务**：
基本篇章单位EDUs的划分 和 篇章结构的生成。

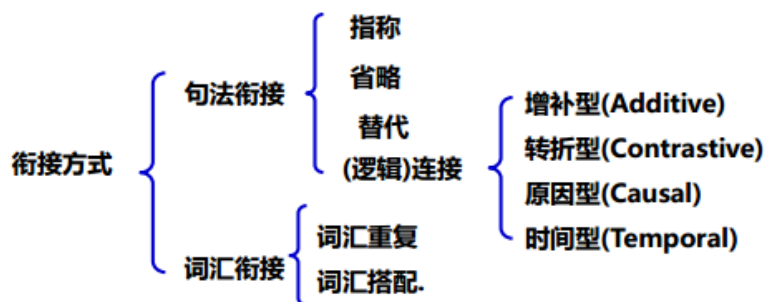
基于RST的篇章结构分析步骤

1.篇位切分:

将整个语篇切分成若干篇位(EDUs)。篇
Mann&ThomsPon 以小句(Clause)作

2.确定结构段

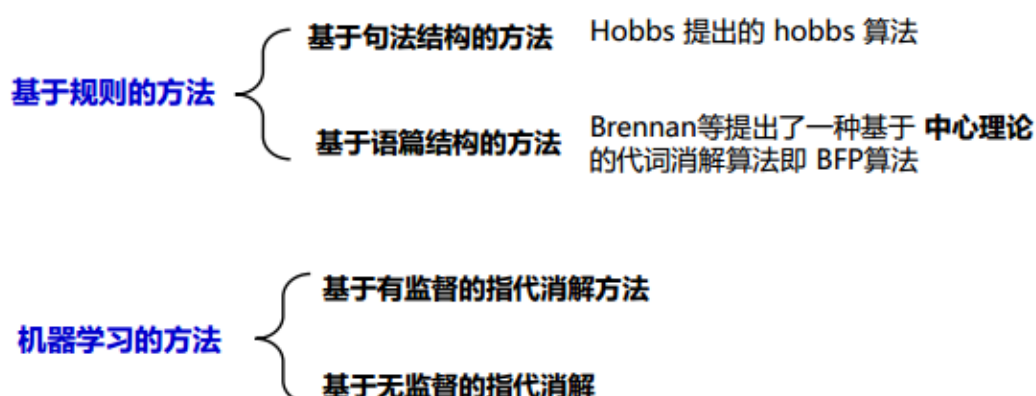
RST 特点：语篇核心、关联词语和关系命题、从属连接



指代消解

指代：篇章中的一个语言单位（通常是词或短语）与之前出现的语言单位存在特殊语义关联，其语义解释依赖于前者。

指代消解方法:



NLP中常用中心理论做指代消解理论

中心理论 (Centering Theory)

Grosz and Sidner (1983)创立,是一种关于语篇结构的理论。该理论认为篇章由三个分离的但相互联系的部分组成：**话语序列结构**（语言结构），**目的结构**（说话者意图）和**关注焦点状态**（说话者注意力状态）

中心理论话题关系主要有四种：

延续话题(continue),**保持**话题(retain),**小幅度转换**(smooth shift),**大幅度转换**(rough shift)。

中心理论：每个语篇单位有三个中心

