计算语言学

第4讲 形式语言与自动机

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

什么是语言

- 代数学的定义方法（本次课介绍）
  - 确定性定义方法
  - 语言是句子的集合
- 统计学的定义方法（下次课介绍）
  - 不确定性定义方法
  - 语言就是一个概率分布，又称为语言模型
  - 一种语言中，每一个句子都对应一个出现概率

如何描述一种语言

- 枚举
  - 给出语言中的所有句子
  - 对于含无限多个句子的语言不合适
- 语法
  - 给出生成语言中所有句子的方法
  - 当且仅当能够用该方法产生的句子才属于该语言
- 自动机
  - 给出识别该语言中句子的机械方法

形式语法 1

- 终结符（Terminals）的有限集合V_T
  - 终结符是句子中实际出现的符号
  - 相当于单词表（有时也称为字母表）
- 非终结符（Non-terminals）的有限集合V_N
  - 非终结符在句子中不实际出现
  - 但在推导中起变量作用
  - 相当于语言中的语法范畴

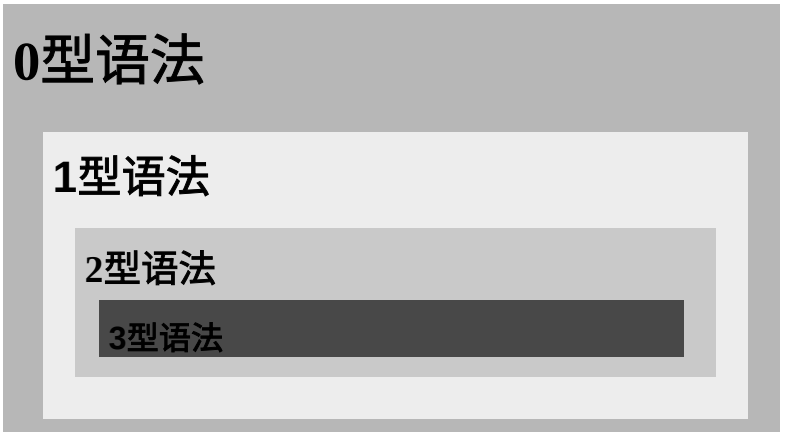
形式语法 2

- 起始符S
  - S属于V_N
  - 相当于句法范畴中的句子
- 重写式规则（Rewriting Rules）的有限集合P
  产生式规则（Production Rules）的有限集合P
  - 基本形式：$\alpha \rightarrow \beta$
  - 含义：将α改写成β
  - α和β是终结符和非终结符组成的串
  - α非空，β可以为空

形式语法 3

- 形式语法：四元组 $G = < V_T, V_N, S, P >$
- 直接推导：$\alpha x \beta \Rightarrow \alpha y \beta$
  如果 x→y是P中的规则
- 推导：$\alpha \overset{*}{\Rightarrow} \beta$
  如果α可以经过多次直接推导得到β
- 语言：$L(G)=\{\alpha \mid \alpha \in V_T^*；S \overset{*}{\Rightarrow} \alpha\}$

乔姆斯基的语法层级



乔姆斯基0型语法

- 短语结构语法，无限制重写语法
  PSG：Phrasal Structure Grammar
- 对规则形式的约束
  - 对于规则形式没有任何限制

乔姆斯基1型语法

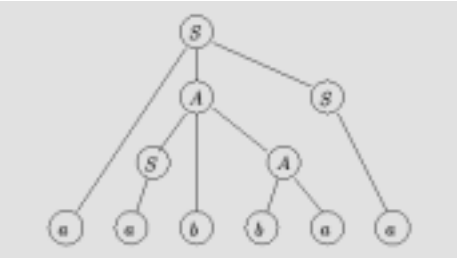
- 上下位有关语法，上下位敏感语法
  CSG：Context Sensitive Grammar
- 对规则形式的约束：
  - $\alpha \rightarrow \beta$
    α, β是任意串，且α的长度小于β的长度
  - $\alpha A \gamma \rightarrow \alpha\beta\gamma$
    A是非终结符， α,β,γ是任意串
  - 以上两种形式等价
  - 敏感：在一定的上下文环境下A可改写为y

乔姆斯基2型语法

- 上下位无关语法，上下位自由语法
  CFG：Context Free Grammar
- 对规则形式的约束：
  - $A \rightarrow \alpha$ ：A是非终结符， α是任意串
  - 在任何上下文环境下A可改写为α

上下文无关语法－例子

- $S \rightarrow aAS$
- $S \rightarrow a$
- $A \rightarrow SbA$
- $A \rightarrow ba$



$\mathbf{S} \Rightarrow aA\mathbf{S} \Rightarrow a\mathbf{A}a \Rightarrow a\mathbf{S}bAa \Rightarrow aab\mathbf{A}a \Rightarrow aabbaa$

乔姆斯基3型语法

- 正规语法，正则语法
  RG：Regular Grammar
- 对规则形式的约束
  - A → Bx 或者 A → x，A,B是非终结符，x是终结符
- 一部正则语法可以表示为一个正则表达式
  例子：{a{b|c}*}+[d|e]{f|g|h}+

乔姆斯基语法层级－例子

- P={S→A1, A→A0, A→0}
  - L(G)={$0^m1|m>=1$}
  - 是正则语法
- P={S→0S1, S→01}
  - L(G)={$0^n1^n|n>=1$}
  - 是上下位无关语法，但不是正则语法
- P={S→0SBC, S→0BC, CB→BC,
  0B→01,1B→11, 1C→12, 2c→22}
  - L(G) ={$0^n1^n2^n|n>=1$}
  - 是上下位有关语法，但不是上下位无关语法语法

乔姆斯基层级以外的语法

- 介于CFG和CSG之间的语法
  - 索引语法
    IG：Index Grammar
  - 可以生成{$a^nb^nc^n$}形式的语言
  - 树粘接语法
    TAG：Tree Adjoining Grammar
- 与乔姆斯基语法层级相交叉的语法

索引语法 1

- 索引语法是一个五元组（V_N, V_T, V_I, P, S)
- V_N, V_T, S与前面的定义相同
- I是索引的有限集合
- P是重写规则的有限集合，规则形式为：

  1) $A \rightarrow \alpha$

  2) $A \rightarrow B(f)$

  3) $A(f) \rightarrow B$

  $A, B \in V_N,\ f \in V_I,\ \alpha \in (V_N \cup V_T)^*$

索引语法 2

- 直接推导($\Rightarrow$)的定义
  - 如果$A \rightarrow X_1X_2\ldots X_k$是规则集中具有1)形式的规则，那么：

    $\beta A(\delta)\gamma \Rightarrow \beta X_1(\delta_1)\ X_1(\delta_1)\ X_2(\delta_2)\ldots\ X_k(\delta_k)\gamma$

    其中，$X_i \in V_N$时，$\delta_i = \delta$；$X_i \in V_T$时，$\delta_i = \varepsilon$
  - 如果$A \rightarrow B(f)$是规则集中具有2)形式的规则，那么：

    $\beta A(\delta)\gamma \Rightarrow \beta B(f\delta)\ \gamma$
  - 如果$A(f) \rightarrow X_1X_2\ldots X_k$是规则集中具有3)形式的规则，那么：

    $\beta A(f\delta)\gamma \Rightarrow \beta X_1(\delta_1)\ X_1(\delta_1)\ X_2(\delta_2)\ldots\ X_k(\delta_k)\gamma$

    其中，$X_i \in V_N$时，$\delta_i = \delta$；$X_i \in V_T$时，$\delta_i = \varepsilon$
- 推导($\overset{*}{\Rightarrow}$)的定义和语言的定义与前面类似

索引语法 3

- 例子（规则集）
  - $S \rightarrow S(\alpha)$
  - $S \rightarrow ABC$
  - $A(\alpha) \rightarrow aA$
  - $B(\alpha) \rightarrow bB$
  - $C(\alpha) \rightarrow cC$
  - $A \rightarrow a$
  - $B \rightarrow b$
  - $C \rightarrow c$
- 推导
  - $S \Rightarrow S(\alpha)$
  - $\Rightarrow S(\alpha\ \alpha)$
  - $\Rightarrow S(\alpha\ \alpha\ \alpha)$
  - $\Rightarrow A(\alpha\ \alpha\ \alpha)\ B(\alpha\ \alpha\ \alpha)\ C(\alpha\ \alpha\ \alpha)$
  - $\Rightarrow aA(\alpha\ \alpha)\ B(\alpha\ \alpha\ \alpha)\ C(\alpha\ \alpha\ \alpha)$
  - $\Rightarrow aaA(\alpha)\ B(\alpha\ \alpha\ \alpha)\ C(\alpha\ \alpha\ \alpha)$
  - $\Rightarrow aaaaB(\alpha\ \alpha\ \alpha)\ C(\alpha\ \alpha\ \alpha)$
  - $\Rightarrow \ldots\ldots \Rightarrow aaaabbbbcccc$
- 可以生成$\{a^n b^n c^n\}$形式的语言，不是CFG

语法的判定复杂度

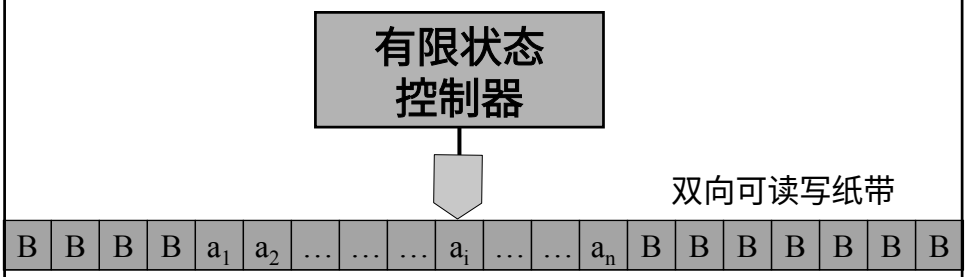
- PSG：半可判定

  对于一个属于0型语言的句子L，总可以在确定步内判断出“是”；但对于一个不属于0型语言的句子L’，不存在一个算法，可以在确定步内判断出“否”。
- CSG：可判定，复杂度：NP完全
- CFG：可判定，复杂度：多项式
- RG：可判定，复杂度：线性

用什么语法描述自然语言

- 正则语法描述能力太弱、上下文有关语法计算复杂度太高，上下文无关语法使用最为普遍
- 从描述能力上说，上下文无关语法不足以描述自然语言——自然语言中上下文相关的情况非常常见
- 从计算复杂度来说，上下文无关语法的复杂度是多项式的，其复杂度可以忍受
- 为弥补上下文无关语法描述能力的不足，需要加上一些其他手段扩充其描述能力

图灵机—直观描述



在每一个时刻，可以定义图灵机的格局为(q,a,i)

其中q为当前状态，a为当前纸带上的字符串，i为当前读写头的位置

图灵机－形式定义

- 图灵机是一个七元组
  M= (Q, Σ, Γ, δ, q_0, B, F)
  - Q为自动机状态的有限集合
  - Γ为一个有限的字符集合
  - B为空白字符，B∈Γ
  - Σ ⊆ Γ-{B}, 为输入字符集合
  - δ 是一个状态转移函数：Q× Σ→Q× Σ×{R,L,S}
    R,L,S分布表示读写头左移、右移或者不动
  - q_0 ∈Q为初始状态
  - F⊆Q为终止状态集

图灵机接受的字符串

- 开始时，纸带中间有n个字符构成输入串，余下的无穷多个字符为空白字符，空白字符不是输入符号
- 开始时，读写头处于输入串的最左端，图灵机的状态为q_0
- 如果图灵机M对于字符串α在执行过程中进入某个终止状态，称为M接受字符串α；如果执行过程中遇到一个格局在状态转移函数中没有定义，那么称M不接受字符串α

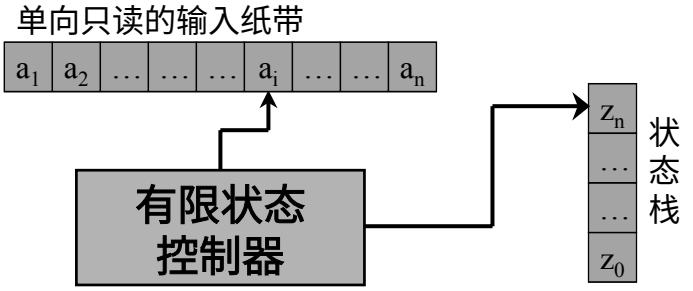
图灵机识别的语言

- 图灵机识别的语言等价于0型语法所生成的语言
  - 某个0型语法所生成的语言必定能被某个图灵机所识别
  - 一个图灵机所能识别的语言必定可以用某种0型语法来生成

线性有界自动机

- 线性有界自动机的构造与图灵机完全一致
- 对图灵机的限制：纸带存在一个左右边界（用两个特殊符号来标识），图灵机的执行过程中读写头位置不能超出边界
- 线性有界自动机所识别的语言等价于1型语法所生成的语言

下推自动机—直观描述



在每一个时刻，可以定义下推自动机的格局为(q,w,z)

其中q为当前状态，w为当前纸带上未读入的字符串，z为状态栈中的字符串

下推自动机—形式描述

- 下推自动机是一个七元组

  $M=(Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$
  - Q为自动机状态的有限集合
  - Σ为输入纸带上字符的有限集合
  - Γ为堆栈字符的有限集合
  - $q_0 \in Q$为初始状态
  - $Z_0 \in \Gamma$是堆栈中的一个特殊符号，表示栈底
  - $F \subseteq Q$为终止状态集
  - δ 是一个状态转移函数：$Q\times (\Sigma \cup \{\varepsilon\}) \rightarrow Q\times \Gamma$

有限状态自动机－形式定义

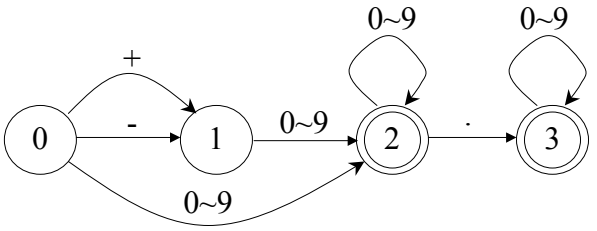
- 有限状态自动机是一个七元组

  M=(Q, I, U, δ, σ, q_0, F)
  - Q为自动机状态的有限集合
  - I为输入字符的有限集合
  - O为输出字符的有限集合
  - δ 是一个状态转移函数：Q× I→Q
  - σ 是一个输出函数：Q× I→U
  - q_0 ∈Q为初始状态
  - F⊆Q为终止状态集

有限状态接收机

- 有限状态接收机（Acceptor）是一个五元组

  M=(Q, I, δ, q_0, F)



识别一个十进制实数的自动机

有限状态转录机

- 有限状态转录机（Transducer）是一个六元组 M=(Q, I, U, δ, σ, q_0)

有限状态自动机的应用

- 有限状态自动机具有简单、直观、高效的特点，在很多领域中得到了广泛的应用
  - 词典构造（Xerox Europe）
  - 机器翻译（Alshiwi's work）
- 有限状态机自动机通过递归（输入另一个自动机的识别结果）可以实现上下文无关语法的描述能力
- 有限状态转录机可以进行翻译

语法、自动机和语言

	语法	自动机	语言	复杂度
0型	无约束短语结构语法	图灵机	递归可枚举语言	半可判定
1型	上下文有关语法	线性有界自动机	上下文有关语言	NP完全
2型	上下文无关语法	下推自动机	上下文无关语言	多项式
3型	正则语法	有限自动机	正则语言	线性

复习思考题

- 给出能够识别以下语言的语法
  $\{a^i b^j c^k d^l \mid i, j, k, l \geq 1\}$
  $\{a^j b^k c^k d^j \mid j, k \geq 1\}$
  $\{a^j b^j c^k d^k \mid j, k \geq 1\}$
  $\{a^k b^k c^k \mid k \geq 1\}$
- 写一部简单的上下文无关语法分析以下句子，并给出其各种分析结果（画出句法分析树）
  - 咬死了猎人的狗
  - Time flies like an arrow.
  - The boy saw the girl with a telescope in the park.
- 写一部正则语法能够识别一个英语的简单名词短语（BaseNP），并画出其有限状态自动机