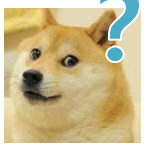
Lecture 2: Language and Grammar

Xiaoyuan Xie 谢晓园
xxie@whu.edu.cn
计算机学院E301

2.1 语言的定义

2.1 语言的定义

- **什么是语言？**
  - 作用：沟通，交流，传递信息
  - 自然语言：英语，中文，日语……
- **你如何定义语言？**
  - 无限的集合
  - 抽象地来看：字符组成了单词; 单词组成了句子; 句子携带了信息
  - “doge”（it’s not a word): reserved words and some rules for words（词法）



  - “Long time no see”（It’s not a correct sentence): rules for sentences（语法）
  - “I ate sky for my cat”（It does not make any sense): 正确的语义信息

2.1 语言的定义

- **Languages as Infinite Sets, there are problems**
  - How can finite recipes generate enough infinite sets of sentences?
  - If a sentence is just a sequence and has no structure and if the meaning of a sentence derives, among other things, from its structure, how can we assess the meaning of a sentence?
- **计算机视角-抽象，再抽象**
- **Grammars as a Generating Device**
  - 有限的规则来描述的无限集

2.1 语言的定义

- **符号(Symbol/Character): 语言中不可再分的单位**
- **字母表: 符号的非空有穷集合**
  - Σ，V或其它大写字母
  - V_1 = {a, b, c}, V_2 = {+, -, 0, 1, …,9}, Σ= {x|x∈ASCII字符}
- **符号串(字符串): 某字母表上的符号的有穷序列**
  - a, b, c, abc, bc,…：V_1上的符号串; 1250, +2, -1835,…：V_2上的符号串
  - 空串（ε）：不含任何符号的串

2.1 语言的定义

- **语句: 字母表上符合某种构成规则的符号串序列**
  - He is a good student. Peanut eats monkey.
  - for(int i = 0; i<10; i++) {call_func(i);}
- **语言 L：某字母表上的语句的集合**

  用a, b, c,…表示符号;

  用α, β, γ…表示符号串;

  用L, M,…表示符号或符号串的集合

2.1 语言的定义

- **符号串连接:**
  - x和y的连接xy是把y的所有符号顺序地接在x的符号之后所得到的符号串
- **符号串方幂:**
  - 设x是字母表∑上的符号串，把x自身连接n次得到的符号串z, 即z = xx…xx(n个x),称作符号串x的n 次幂，记作 $z = x^n$
- **符号串前缀后缀:**
  - 设x、y、z是某一字母表上的符号串，x = yz，则y是x 的前缀，z 是x的后缀; z≠ε 时y是x的真前缀, y≠ε时z 是 x 的真后缀
- **符号串子串:**
  - 非空字符串x，删去它的一个前缀和一个后缀后所得到的字符串称为 x 的子字符串，简称子串。如果删去的前缀和后缀不同时为ε，则称该子串为真子串

2.1 语言的定义

- **符号串集合(语言)的积**
  - 设串集$L=\{\alpha_1, \alpha_2, \ldots\}$，$M=\{\beta_1, \beta_2, \ldots\}$，二者的笛卡尔积$LM=\{\alpha\beta \mid \alpha\in L, \beta\in M\}$
  - E.g. L={ab, abb}, M={ced, cd}，那么 LM = {abced, abcd, abbced, abbcd}
- **字符串集合(语言)的方幂**
  - $L^0=\{\varepsilon\}$，$L^1=L$，$L^n=LL^{n-1}$
  - 若 $|L|=m$，那么，$|L^0|=1$，$|L^1|=m$，$|L^n|=m^n$
- **字符串集合(语言)的Kleene闭包**
  - $L^*=L^0\cup L^1\cup L^2\cup\ldots$
- **字符串集合(语言)的正闭包**
  - $L^+=L^1\cup L^2\cup\ldots=L^*-\{\varepsilon\}$

语言L就是其字母表上闭包的子集

2.1 语言的定义

- **练习：**

  L: { A, B, …, Z, a, b, …, z }, D: { 0, 1, …, 9 }

  则$L \cup D$, LD, L^6, L^*, $L(L \cup D)^*$, D^+ 分别是什么？

2.1 语言的定义

- **文法(G, Grammar): 四元组G = (V_N,V_T, S, P)，其中**
  - V_N：一个非空有限的非终结符号集合，它的每个元素称为非终结符，一般用大写字母表示，它是可以被取代的符号；
  - V_T：一个非空有限的终结符号集合，它的每个元素称为终结符，一般用小写字母表示, 是一个语言不可再分的基本符号；
  - S：一个特殊的非终结符号，称为文法的开始符号或识别符号，$S \in V_N$。开始符号S必须至少在某个产生式的左部出现一次；
  - P：产生式的有限集合。所谓的产生式，也称为产生规则或简称为规则，是按照一定格式书写的定义语法范畴的文法规则。
  - 设V是文法G的符号集，则有：$V = V_N \cup V_T$ ， $V_N \cap V_T = \emptyset$

2.1 语言的定义

- **产生式形式**
  - a→b 或 a:: = b
  - a称为产生式的左部, a∈V+，并且至少含有一个非终极符；
  - b称为产生式的右部，b∈V * ；
  - “→” “:: = ” 读作“定义为”或“由…组成” ；
  - “|”是或操作

2.1 语言的定义

- **推导：使用产生式的右部取代左部的过程**
  - 文法产生句子
  - 最左推导和最右推导称为规范推导。
- **归约：推导的逆过程，用产生式的左部取代右部的过程**
  - 最左归约和最右归约称为规范归约。

2.1 语言的定义

- **自然语言文法示例**
  - 产生式

<句子>→<主语><谓语><宾语>
<主语>→<形容词><名词>
<谓语>→<动词>
<宾语>→<形容词><名词>
<形容词>→young | pop
<名词>→men | music
<动词>→like

还能推导出什么句子？

2.1 语言的定义

- **句型**
  - 从文法开始符号S开始，每步推导（包括0步推导）所得到的字符串α: S→α，其中$\alpha \in (V_N \cup V_T)^*$
- **句子**
  - 仅含终结符的句型
- **语言**
  - 由S推导所得的句子的集合L(G)={α|S→α，且$\alpha \in V_T^*$}，G为文法

2.1 语言的定义

- **文法规则的递归定义**
  - 非终结符的定义中包含了非终结符自身
  - 设Σ={0,1}; <整数>→<数字><整数>|<数字>; <数字>→0 | 1
- **使用递归定义时要谨慎，要有递归出口，否则可能永远产生不出句子**

2.1 语言的定义

- **扩充的BNF表示**
  - () ——提因子: U → ax | ay | az 改写为U → a (x | y | z)
  - {} ——重复次数的指定: <标识符>→<字母>{<字母>|<数字>}$_0^5$
  - [] ——任选符号: <整数>→[+|-]<数字>{<数字>}

2.1 语言的定义

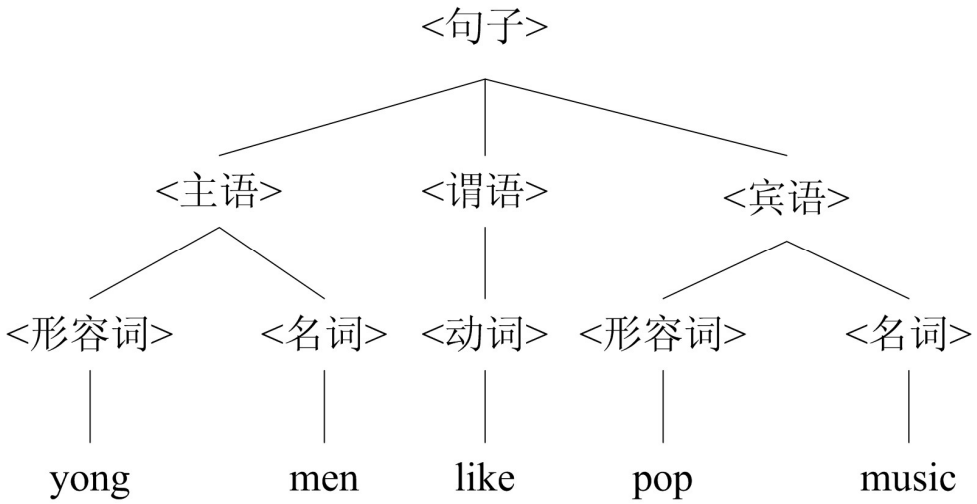
- **自然语言文法示例**
  - 产生式

<句子>→<主语><谓语><宾语>
<主语>→<形容词><名词>
<谓语>→<动词>
<宾语>→<形容词><名词>
<形容词>→young | pop
<名词>→men | music
<动词>→like

还能推导出什么句子？

2.1 语言的定义

- **自然语言文法示例**



2.1 语言的定义

- **也可以用字符串定义语言**

例：设文法$G_2=(\{S\},\{a,b\},P,S)$，其中P为:

(0) $S \rightarrow aSb$

(1) $S \rightarrow ab$

等价于 $L(G_2)=\{a^nb^n|n>=1\}$

2.2 Chomsky语法类型

2.2 Chomsky 语法类型

- **Chomsky 0型文法: 短语文法或无限制文法**
  - P：$\alpha \rightarrow \beta$，其中$\alpha \in V^+$ 并至少含有一个非终结符, $\beta \in V^*$.
  - 是对产生式限制最少的文法；
  - 对0型文法的产生式作某些限制，可以得到其他类型的文法
- **识别0型语言的自动机称为图灵机 (TM)**

2.2 Chomsky 语法类型

- **Chomsky 1型文法: 长度增加文法/上下文有关文法 ）**
  - P：$\alpha \rightarrow \beta$，除可能有$S \rightarrow \varepsilon$ 外均有$|\beta| >= |\alpha|$；若有$S \rightarrow \varepsilon$，规定S不得出现在产生式右部。或
  - P中产生式$\alpha \rightarrow \beta$，除可能有$S \rightarrow \varepsilon$外均有$\alpha A \beta \rightarrow \alpha \gamma \beta$，其中$\alpha, \beta \in V^*$，$A \in V_N$，$\gamma \in V^+$
  - 1型文法对非终结符进行替换时必须考虑上下文
  - 除文法开始符号外不允许将其它的非终结符替换成ε
- **识别1型语言的自动机称为线性界限自动机(LBA)**

2.2 Chomsky 语法类型

- **Chomsky 2型文法:上下文无关文法**
  - P：$A \to \beta$，其中$A \in V_N$ ，$\beta \in V^*$。
  - 所有的产生式左边只有一个非终结符，产生式右部可以是V_N、V_T或ε
  - 非终结符的替换不必考虑上下文，故也称作上下文无关文法。
- **识别2型语言的自动机称为下推自动机(PDA)。**

2.2 Chomsky 语法类型

- **Chomsky 3型文法:正规文法**
  - P中产生式具有形式$A\rightarrow \alpha B$， $A\rightarrow \alpha$ (左线性)，或者$A\rightarrow B\alpha$，$A\rightarrow \alpha$(右线性)， 其中A，$B\in V_N$, $\alpha\in V_T^*$。
  - 也称为正规文法RG、线性文法：若所有产生式均是左线性，则称为左线性文法；若所有产生式均是右线性，则称为右线性文法。
  - 产生式要么均是右线性产生式，要么是左线性产生式，不能既有左线性产生式，又有右线性产生式。
- **识别3型语言的自动机称为有限状态自动机(FA)。**

2.2 Chomsky 语法类型

- **由文法产生语言（3型）**

例：设文法$G_1=(\{S\},\{a,b\},S,P)$，其中P为:

(0) $S \rightarrow aS$

(1) $S \rightarrow a$

(2) $S \rightarrow b$

答：$L(G_1) = \{a^i(a \mid b) \mid i >= 0\}$

2.2 Chomsky 语法类型

- **由文法产生语言（2型）**

例：设文法$G_2=(\{S\},\{a,b\},P,S)$，其中P为:

(0) $S \rightarrow aSb$

(1) $S \rightarrow ab$

答：$L(G_2)=\{a^nb^n|n>=1\}$

2.2 Chomsky 语法类型

- **由文法产生语言（1型）**

例：设文法G_3＝({S, Q},{a,b},P,S)
其中P为:
S ---> abc | aSQ
bQc ---> bbcc
cQ ---> Qc

答：

$L(G_2)=\{a^nb^nc^n|n>=1\}$

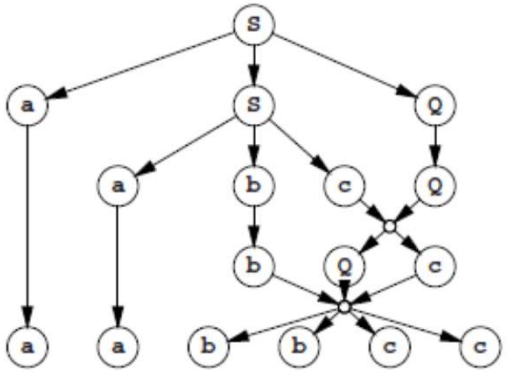


Fig. 2.8. Derivation of **aabbcc**

Step	Queue				Result
1	S				
2	abc	aSQ			abc
3	aSQ				
4	aabcQ	aaSQQ			
5	aaSQQ	aabQc			
6	aabQc	aaabcQQ	aaaSQQQ		
7	aaabcQQ	aaaSQQQ	aabbcc		
8	aaaSQQQ	aabbcc	aaabQcQ		
9	aabbcc	aaabQcQ	aaaabcQQQ	aaaaSQQQQ	aabbcc
10	aaabQcQ	aaaabcQQQ	aaaaSQQQQ		
11	aaaabcQQQ	aaaaSQQQQ	aaabbccQ	aaabQQc	
...	...				

Fig. 2.17. The first couple of steps in producing for $a^nb^nc^n$

2.2 Chomsky 语法类型

- **由语言构造文法**

例：设$L_1=\{a^{2n}b^n|n>=1$且$a, b \in V_T\}$，试构造生成L_1的文法G_1。

解：n=1， L_1 =aab

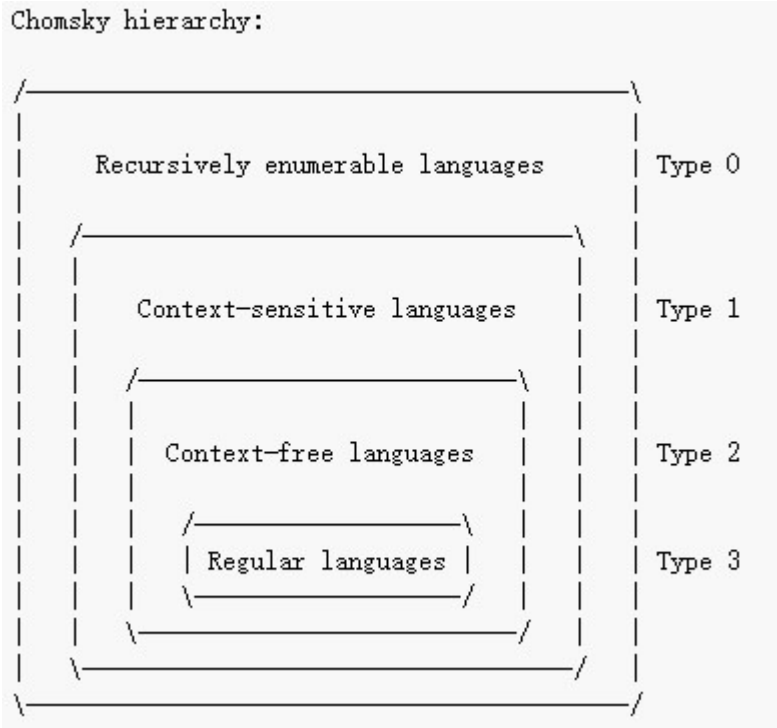
n=2， L_1 =aaaabb

n=3， L_1 =aaaaaabbb

得：$S \rightarrow aaSb$

$S \rightarrow aab$

2.2 Chomsky 语法类型



本章小结

语言的定义

Chomsky四种类型文法

问题与作业

- 教材P78: 3.3.3
- $\{a, b\}^*=\{?\}$, $\{a, b\}^+=\{?\}$
- 自然语言文法产生式示例

<句子>→<主语><谓语><宾语>
<主语>→<形容词><名词>
<谓语>→<动词>
<宾语>→<形容词><名词>
<形容词>→young | pop
<名词>→men | music
<动词>→like

给出所有能推导出的句子

思考

- 给定产生式，如何证明给定的语言是几型文法？--- 根据定义即可
- 如何通过字符串定义证明给定的语言是几型文法？
  - 写出产生式，然后判断文法类型，or
  - 利用pumping lemma
- 延伸阅读 (optional)
  - pumping lemma (uvvwxxy 和 uvw定理)
  - 思考：利用uvw定理证明不存在适用于于语言L(G) = {$a^i b^i$}的三型文法.

作业

- 判断chomsky语言类型:
  - S -> aSb; S -> ab (Type-?)
  - aSb -> aaSbb; S -> ab (Type-?)
  - S-> aS; S->ab (Type-?)

Thank you!