离散数学：形式语言与自动机：语言及研究方向

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

自然语言

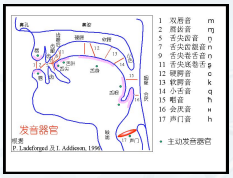
- 以呼吸器官**发声**为基础来**传递信息**的**符号系统**，人类最重要的交际工具和存在方式之一
- 大脑思维的符号化
- 自然语言：**自然**地随文化**演化**的语言
- 汉语、英语、法语、俄语……
- 全世界5000多种，使用者在5000万以上的有13种
- 联合国官方语言5种（汉、英、俄、法、西班牙）

人工语言

- 为特定的目的与用途，人为创造的语言
- 国际**辅助语言**：
- 世界语（Esperanto）
- **数学语言、计算机语言**
- 数学符号、逻辑语言
- 程序设计语言
- 作为**传递信息**的**符号系统**本质未改变

语音Phonetics：发音的体系

- 发音包括：**音素、音节、语调**
- 发音是语言最基础的部分
- 有些语言甚至没有文字符号
- 人类拥有共同的发音器官，但不同的语言发音却大不相同
- 同一语言的不同方言发音也大不相同
- **国际音标**建立了统一的发音记录符号体系



语形Morphology：书写的格式和规范

- 构字、词法、句法、文章格式
- 拼音文字：以发音为基础构字，**一维表音**串
- **bake, cake, fake, lake, make**
- 象形文字：以**二维表意**图形为基础构字
- 木，林，森，火，炎，焱
- **词法**由词典来规定，不在词典中的词为非法
- 比如不能用：**森炎**，这样的词
- 字和词的符号和意义都可以穷尽

句法Syntax（或语法）规定句子组成的规则

- 语句的符号和意义都不可穷尽
- 符号只能由规则描述其结构，每个部分应为哪些词类
- 主谓宾定状补：名词、动词、副词、形容词、数词、量词、介词、代词
- **小李**正在树下**读书**
- 语法对于意义无能为力
- **书**一下午**读**了3本**小李**
- （有趣的拼句游戏）

语义Semantics：词句的含义

- 从符号系统还原思维
- **思维→语言→(传递/翻译)→语言→思维**
- 共同理解和保持语义是人类交流的基础
- 语言交流的语义损耗（一图胜千言……）
- 艺术作品的研究
- 如何形式化表达语义是目前研究的热点难题
- 语义网、专家系统、数据挖掘、机器翻译、人工智能（符号系统描述符号系统）

语用Pragmatics：使用环境和功能

- 在不同的上下文环境中语句的应用，对语义的影响
- 语境对语句理解的影响
- **火，火！**
- A：**昨天后来怎么样？**B：**还好没耽误。**
- 语用的研究是自然语言处理NLP的重要内容

离散数学：形式语言与自动机：形式语言

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

语言的定义

- Chomsky:按照一定**规律**构成的**句子**和**符号串**的有限或者无限的集合
- 形式语言主要研究语言**描述**的问题
- **穷举法**：只适合句子数目有限的语言
- **语法描述**：通过有限的替换规则，**生成**语言中合格的句子
- **自动机**：对输入的句子进行**检验**，区别哪些是语言中的句子，哪些不是语言中的句子

基本概念：字符串、词

- 设V是有限集合，其中元素称为“**字符**”
- 由V中字符相连而成的**有限**序列称为“**字符串**”
- 不含任何字符的串称为“**空串**”，记做ε
- 字符串所包含的字符个数称为“**长度**”，记做|s|，|ε|=0
- 包括空串的V上的字符串**全体**记做V*
- 字符串**连接**：s=ab,t=gg连接st=abgg
- 字符串**n次幂**：s自身连接n次，$s^0=\varepsilon$

字符串集合的运算

- **乘积**：AB={st|s∈A,t∈B}
- **幂运算**：
- $A^0=\{\varepsilon\}$，$A^n=A^{n-1}A=AA^{n-1}$
- 例：
- A={aa,bb}, B={cc,dd,ee}
- AB={aacc, aadd, aaee, bbcc, bbdd, bbee}
- A^2={aaaa, aabb, bbaa, bbbb}
- **闭包**：$A^*=A^0\cup A^1\cup A^2\ldots$
- **正闭包**：$A^+=A^1\cup A^2\ldots=A^*-\{\varepsilon\}$

正则表达式Regular Expression

- **RE1**:ε是正则式，对应正则集{ε}
- **RE2**:x∈V，x是正则式，对应正则集{x}
- **RE3**:如果α、β是正则式，则αβ是正则式，对应正则集AB（字符串集合乘积）
- **RE4**:如果α、β是正则式，则(α|β)是正则式，对应正则集A∪B
- **RE5**:如果α是正则式，则(α)*是正则式，对应正则集A*

正则表达式对应字符串集合

- V上的正则表达式**对应和描述**了V*的一个子集（正则子集）
- 例：V={a,b}，下列正则表达式：
- ab*b**描述**了{ab, abb, abbb, ...}
- ab***描述**了{a, ab, abb, abbb, ...}
- a*b***描述**了{ε, a, b, ab, aab, abb, ...}
- (ab)***描述**了{ε, ab, abab, ababab, ...}

离散数学：形式语言与自动机：短语结构语法

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

短语结构语法Phrase Structure Grammar

- 短语结构语法是一个四元组G=<V,S,v0,├>
- V是**字符集**
- S⊆V，称作**终结符**，N=V-S称作**非终结符**
- ├称作**产生式关系**（二元关系），由**w├w'**这样的**产生式**（二元组）构成，表示**w可以替换成w'**，分别称为左部和右部
- v_0∈N，称作**初始符**（句子符），是替换的起点

短语结构语法

- V*上的二元关系：
- 直接导出关系（**x→y**）定义为：
- x=lwr, y=lw'r, 且w├w'是产生式, l,r∈V*
- →关系的**传递闭包→∞**（见和划分中**t(→)**）
- w∈S*是**语法正确**的句子当且仅当v_0 →∞ w

语法例子

- **终结符**S={张三，李四，深情地，狂野地，歌唱，奔跑}
- **非终结符**N={<句子>,<主语>,<谓语>,<动词>,<副词>}
- **产生式关系**：
- <句子>├<主语><谓语>
- <主语>├张三；<主语>├李四
- <谓语>├<副词><动词>
- <副词>├深情地；<副词>├狂野地
- <动词>├歌唱；<动词>├奔跑

直接导出关系生成句子

- <句子>
- → <主语><谓语>
- → 张三<谓语>
- → 张三<副词><动词>
- → 张三<副词>奔跑
- → 张三狂野地奔跑
- 传递闭包：<句子> →∞ 张三狂野地奔跑
- 导出的结果称作**符合语法G的句子**

离散数学：形式语言与自动机：语言及语法表达

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

语法产生的语言

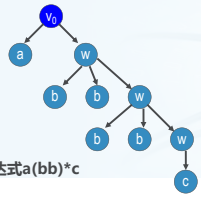
- 利用语法G产生的**所有**的正确构造的句子的集合称作为**G的语言**，记做L(G)
- 不同的语法可能产生不同的语言
- 也可能产生相同的语言

导出树(derivation tree)

- 用多元有向树表示**语言导出过程**
- 起始符v_0作为树根
- 每个子树的**树根**是某个生成式的**左部**
- **子节点**分别是生成式**右部**包含的符号
- 适合**所有**产生式的左部**仅有一个非终结符**情形

语法例子（1）

- V={v_0,w,a,b,c}，S={a,b,c}
- 产生式：v_0├aw; w├bbw; w├c
- L(G)⊆S*



- 对应的正则表达式a(bb)*c

语法例子（2）

- V,S同上
- 产生式：v_0├av_0b; v_0b├bw; abw├c
- 语句分析：
- 第一个产生式产生形如$a^n v_0 b^n$这样的串
- 应用第二个产生式结果形如$a^m abw b^m$
- 应用第三个产生式消除非终结符，结果L(G)形如$a^m c b^m$
- L(G)**不能**用正则表达式表示
- 可见语言之间具有某种类别上的**差异**

离散数学：形式语言与自动机：形式语法分类

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

乔姆斯基形式语法分类1956

- 对于短语结构语法G，讨论其**产生式集合**
- 如果对产生式没有任何约束，称作**0型语法**
- 无限制语法，短语结构语法
- PSG:Phrase Structure Grammar
- 产生**递归可枚举**语言
- 被**图灵机**识别

乔姆斯基形式语法分类1956

- 如果所有产生式形如αAβ├αγβ
- A是非终结符，α,β,γ是任意串，但γ不能为空串
- 称作**1型语法**（上下文相关语法CSG）
- CSG:Context Sensitive Grammar
- 产生**上下文相关**语言
- 被**线性有界自动机**识别

乔姆斯基形式语法分类1956

- 如果所有产生式左部是一个非终结符，形如A├γ（γ可以是任意串）
- 称作**2型语法**（上下文无关语法CFG）
- CFG:Context Free Grammar
- 产生**上下文无关**语言
- 被**下推自动机**识别
- 为大多数**程序设计语言**的语法提供理论基础

乔姆斯基形式语法分类1956

- 如果所有产生式**左**部是**一个**非终结符
- **右**部最多有**一个**非终结符，且只能在**最右**端
- 称作**3型语法**（正则语法RG）
- RG：Regular Grammar
- 产生**正则**语言，被**有限状态自动机**识别
- 也可以用**正则表达式**表示
- 通常用来定义检索**模式**或者程序设计语言中的**词法**结构

形式语言分类

- 对应于形式语法，如果某个语言可以用某个3型语法产生，则称作3型语言
- 如果某个语言不能用任何3型语法产生
- 但可以用某个2型语法产生，则称作2型语言
- L3 ⊂ L2 ⊂ L1 ⊂ L0

离散数学：形式语言与自动机：语法分析

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

语法分析Parsing

- 语法分析是语言导出的**逆**过程
- 从一个**句子**得到**导出树**或导出过程
- 涉及到对句子**结构**的分析

语法分析Parsing

- 在程序设计语言的编译中有重要的应用
- 从高级语言源代码到机器指令序列

语法分析Parsing

- 在自然语言处理中用于分析句子结构，助于理解自然语句
- 每种类型语言的语法分析要借助于对应的自动机

离散数学：形式语言与自动机：BNF范式

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

BNF(Backus-Naur Form)范式

- 1960年提出，用于描述ALGOL60语言
- 形式简单，仅包含三个符号：
- “::=”，定义为
- “|”，或
- “<>”，用来区分非终结符
- 如：<A> ::= a | bc | c<B>

BNF(Backus-Naur Form)范式

- BNF可以表示2、3型语法（上下文无关文法、正则文法）
- BNF产生式左部仅有一个非终结符
- 相同左部的产生式合并用“|”隔开
- 非终结符用“<>”标记

BNF例子

- BNF例子：“中文句子”
- <句子>::=<主语><谓语>
- <主语>::=张三|李四
- <谓语>::=<副词><动词>
- <副词>::=深情地|狂野地
- <动词>::=歌唱|奔跑

BNF例子

- BNF例子：“a(bb)*c”
- <v0>::=a<w>
- <w>::=bb<w>|c
- 递归出现一次，并在最右，称为**正规**的
- 3型语法中的递归产生式都是**正规**的

BNF例子

- BNF例子：“十进制数”
- <十进制数>::=<无符号整数>|<小数>|<无符号整数><小数>
- <无符号整数>::=<数字>|<数字><无符号整数>
- <小数>::=.<无符号整数>
- <数字>::=0|1|2|3|4|5|6|7|8|9

BNF例子

- BNF例子：“标识符”
- 字母开始的字母数字串，用于程序设计语言中的变量、函数等名称
- <标识符>::=<字母>|<字母><后部>
- <后部>::=<字母>|<数字>|<字母><后部>|<数字><后部>
- <字母>::=a|b|c|d|e|.....|z
- <数字>::=0|1|2|3|4|5|6|7|8|9

扩展BNF

- **可选项**：用“[]”表示
- <if语句>::=if <逻辑表达式> then <语句> [else <语句>] end if;
- **重复项**：用“{}”表示重复0次或者多次
- <标识符>::=<字母>{<字母>|<数字>}
- 便于区分单个符号的终结符，加**引号**
- <语句序列>::=<语句>{ “;” <语句> }
- 有时也用**粗体字**表示终结符，非终结符不加尖括号，可读性更好

用BNF定义BNF ☺

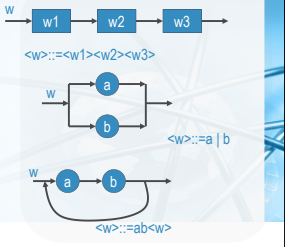
- <BNF> ::= <规则> | <规则><BNF>
- <规则> ::= <非终结符>::=<表达式>
- <非终结符> ::= <<单词>>
- <单词> ::= <字母> | <字母><单词>
- <表达式> ::= <项> | <项>|<表达式>
- <项> ::= <因子> | <因子><项>
- <因子> ::= <单词> | <非终结符>

离散数学：形式语言与自动机：语法图

陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

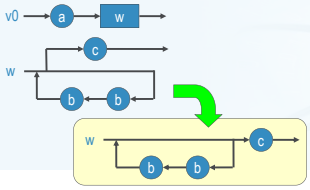
语法图（Syntax Diagram）

- 用**方框**表示非终结符，用**圆**表示终结符
- 用带**箭头线**表示连接关系和连接顺序
- 类似程序流程图
- **顺序**结构（连接）
- **可选分支**结构（BNF或）
- **循环**结构（非终结符递归）

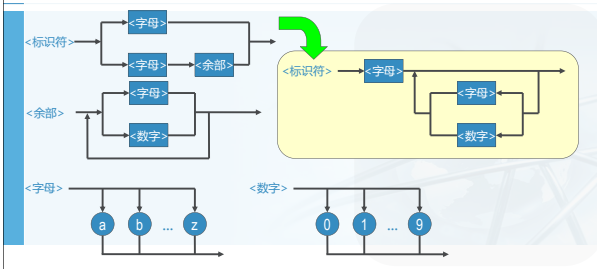


语法图的化简与等价

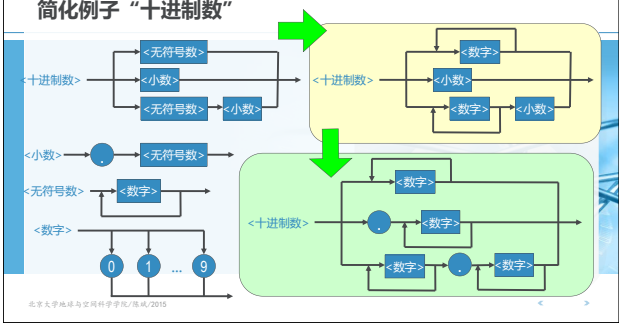
- <v0>::=a<w>
- <w>::=bb<w>|c



简化例子“标识符”



简化例子“十进制数”



离散数学：形式语言与自动机：正则语法

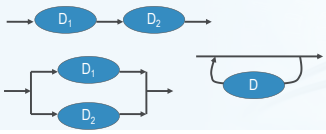
陈斌 北京大学地球与空间科学学院 gischen@pku.edu.cn

正则集合与正则语法

- S是有限集合，L⊆S*
- 则L是正则集合，**当且仅当：**
- 对某个正则语法G，有L=L(G)
- 我们可以从正则语法G**构造**对应正则集合的正则表达式（而正则表达式对应正则集合）
- 将G的语法**组合**为一个大图，其中仅包含v0和终结符，称作主导图（master diagram）
- 语法图中终结符**对应**正则表达式中的终结符

主导图翻译成正则表达式

- **串联**子图，对应子表达式$\alpha_1\alpha_2$
- **并联**子图，对应子表达式$\alpha_1|\alpha_2$
- **回路**子图，对应子表达式α^*



主导图翻译成正则表达式：例子

- a|b|c*
- (a|b)cd*
- (a|bc)*