北京大学信息科学技术学院 2022年春季学期《编译原理》



第4章语法分析(3)

Syntax Analysis

【对应教材 2.4, 4.3, 4.4】

内容提要



- □ 语法分析简介
- □上下文无关文法
- □ 文法的设计方法
- □ 自顶向下的语法分析
- □ 自底向上的语法分析
 - □ 简单LR分析: LR(0), SLR
 - □ 更强大的LR分析: LR(1), LALR
 - □ 二义性文法的使用

自底向上的语法分析



□ 为一个输入串构造语法分析树的过程;

- □ 从叶子(输入串中的终结符号,将位于分析树的底端)开始,向上到达根结点;
 - 在实际的语法分析过程中并不会真的构造出相应的 分析树,但是分析树概念可以方便理解;
- □ 自底向上语法分析的通用框架
 - "移进-归约"分析(shift-reduce)

归约(Reduce)



- □ 可以把自底向上语法分析过程看成从串 w "归约"为文法开始符号S的过程;
- □ 归约步骤:
 - 一个与某产生式右部相匹配的特定子串被替换 为该产生式左部的非终结符号;
- □ 问题:
 - 何时归约(归约哪些符号串)?
 - 归约到哪个非终结符号?

自底向上分析的例子



文法: S→aAS a A→SbA SS ba

句子: aabbaa

S

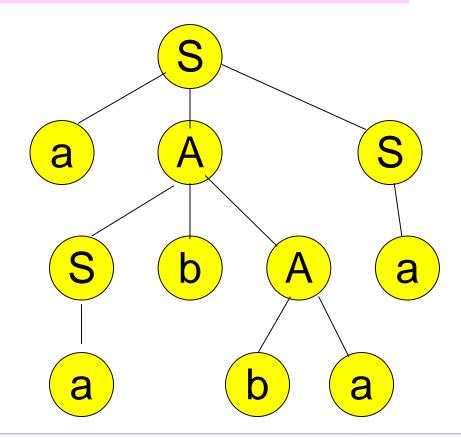
 \Rightarrow aAS

 \Rightarrow aA<u>a</u>

⇒aSbAa

⇒aSb<u>ba</u>a

⇒a<u>a</u>bbaa



归约的例子



- □ id * id的归约过程
 - id * id <= F*id <= T*id <= T*F <= T <= E</p>
- □ 对于句型T*id,有两个子串和某产生式右部 匹配
 - T是E→T的右部;
 - id是F→id的右部;
 - 为什么选择将id归约为F,而不是将T归约为E?
 - □ 原因: T归约为E之后, E*id不再是句型;
 - □ 问题:如何确定这一点?

句柄



- □ 对输入从左到右扫描,并进行自底向上语法分析,实际上可以反向构造出一个最右推导
- □ 句柄 (handle):
 - 最右句型中和某个产生式右部匹配的子串,对它的 归约代表了该最右句型的最右推导的最后一步;
 - 定义: 如果 $S \stackrel{*}{=} \alpha A w \stackrel{*}{=} \alpha \beta w$; 那么紧跟α之后的β是 (对应 $A \rightarrow \beta$ 的) 一个句柄
- □ 在一个最右句型中, 句柄右边只有终结符号
- □ 如果文法是无二义性的,那么每个句型都有且 仅有一个句柄。

句柄的例子



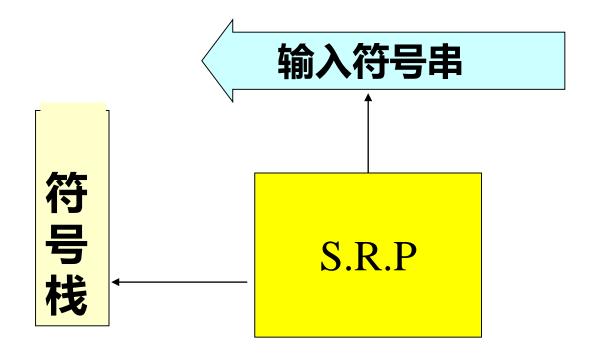
□ 输入: id *id

最右句型	句柄	归约用的产生式
$\mathbf{id}_1 * \mathbf{id}_2$	id_1	$F o \mathbf{id}$
$F*\mathbf{id}_2$	F	$T \to F$
$T * id_2$	id_2	$F o \mathbf{id}$
T * F	T * F	$T \rightarrow T * F$

移进-归约分析 (Shift-Reduce Parsing)



□ 建立符号栈,用来记录分析的历史和现状, 并根据所面临的状态,确定下一步动作是移 进还是归约。



主要分析动作

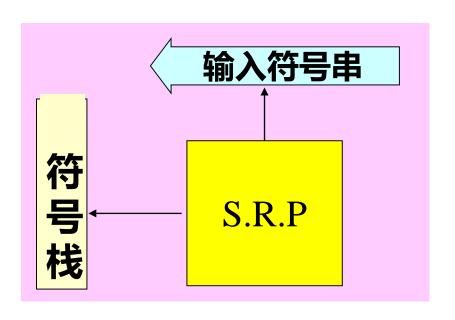


- □ 移进:将下一个输入符号移动到栈顶;
- □ 归约:将句柄归约为相应的非终结符号;
 - 句柄总是在栈顶。
 - 具体操作时弹出句柄,压入被归约到的非终结符号。
- □ 接受: 宣布分析过程成功完成;
- □ 报错:发现语法错误,调用错误恢复子程序

分析过程



- □ 开始时刻: 栈中只包含\$, 而输入为w\$;
- □ 成功结束时刻: 栈中\$S, 而输入\$;
- □ 在分析过程中,不断地移进符号,并在识别到 句柄时进行归约;
- □ 句柄被识别时总是 出现在栈的顶部



移进-归约分析过程的例子



			_
栈	输入	动作	
\$	$\mathbf{id}_1*\mathbf{id}_2\$$	移人	
$\mathbf{\$id}_1$	st \mathbf{id}_2 $\$$	按照 $F \rightarrow id$ 归约	
F	$*$ \mathbf{id}_2 $\$$	接照 $T \rightarrow F$ 归约	
T	$*$ \mathbf{id}_2 $\$$	移人	
T *	$\mathbf{id}_2\$$	移人	
$T * id_2$	\$	按照 $F o \mathbf{id}$ 归约	
T * F	\$	按照 $T \rightarrow T * F$ 归约	
$\ T$	\$	按照 $E \rightarrow T$ 归约	
\$E	\$	accept	

栈内符号串 + 未处理输入符号串 = 当前句型

句柄都在栈顶

图 4-28 一个移入 – 归约语法分析器 在处理输入 $id_1 * id_2$ 时经历的格局

为什么句柄总是在栈顶?



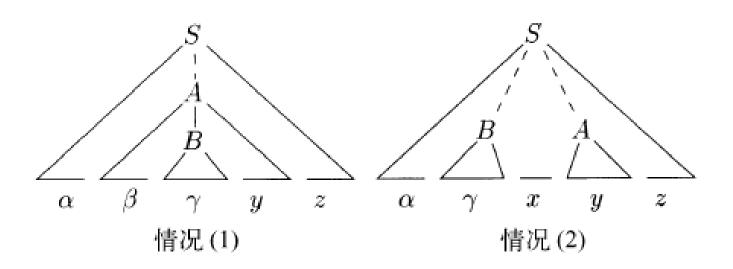
□ 分别考虑最右推导的两个连续步骤的两种情况

1)
$$S \stackrel{*}{\Longrightarrow} \alpha Az \stackrel{*}{\Longrightarrow} \alpha \beta Byz \stackrel{*}{\Longrightarrow} \alpha \beta \gamma yz$$

2)
$$S \stackrel{*}{\Longrightarrow} \alpha B x A z \stackrel{*}{\Longrightarrow} \alpha B x y z \stackrel{*}{\Longrightarrow} \alpha \gamma x y z$$

•
$$x, y, z \in V_T^*$$

•
$$\gamma \in (V_T \cup V_N^*)$$



移进-归约分析中的冲突



- □ 有时候,即使知道了栈中所有内容、以及下面k个输入符号,仍然无法知道是否该进行 归约,或者不知道按照什么产生式进行归约
- □ 分别对应两种冲突情形:
 - "移进-归约"冲突
 - "归约-归约"冲突
- 口 "移进-归约"冲突的例子:
 - 设栈中符号串是αβ,接下来的k个符号是x,产生移进/归约冲突的原因是存在y和y'使得aβxy是最右句型且β是句柄,而aβxy'也是最右句型,但是句柄还在右边。

归约/归约冲突的例子



- □ 输入为 id (id, id)
- □ 冲突时的格局:
 - 栈: ...id (id
 - 输入:,id)...

```
(1)
                  stmt \rightarrow id (parameter\_list)
                  stmt \rightarrow expr := expr
      parameter\_list \rightarrow parameter\_list, parameter
(4)
      parameter\_list \rightarrow parameter
(5)
                          \rightarrow id
           parameter
(6)
                          \rightarrow id ( expr\_list )
                  expr
(7)
                          \rightarrow id
                  expr
(8)
             expr\_list

ightarrow expr_list , expr
(9)
             expr\_list
                                 expr
```

某个语言中, 多维数组的 表示方法。

LR 分析简介



- □ 也称为LR(k)分析,是一种常用的自底向上分析
 - L 指的是从左向右扫描输入符号串
 - R 指的是构造最右推导的逆过程(即规范归约)
 - k 指的是决定动作时向前看的符号个数,通常取0或1
- □ LR分析的过程是进行规范归约,即每次归约的都 是句柄。
- □ LR 分析的主要特点:
 - 1) 适合文法类足够大:比LL(k)文法的范围大
 - 2) 分析效率高(无回溯)
 - 3) 报错及时
 - 4) 可以自动生成

LR(k) 项(LR(k) Item)



定义: $A \rightarrow \alpha \bullet \beta, \gamma$ 称为文法G的一个LR(k) 项。

- 其中, $A\rightarrow\alpha\beta$ 是G的一个产生式;
- γ是一个长度为 k 的由终结符和 \$ 构成的符号串, 称为搜索符串。
- 其中的圆点•表示语法分析程序已读到的符号串α,而构成圆点后面的符号串β尚未读到。
- 搜索串y用于辅助确定是否在栈顶形成了句柄。
- k取0时, LR(0)项简写为 $A \rightarrow \alpha \cdot \beta$ 。
- 产生式 $A\rightarrow \alpha$ 相应的项有 $|\alpha|+1$ 个。

LR(0)项集族的构造



- □ 拓广文法:
 - 增加一条产生式 $S' \rightarrow S$ 拓广文法G
 - 对文法中的产生式进行编号
- □ 构造项集 I用到的子函数
 - CLOSURE(I): I的项集闭包
 - □ 对应于DFA化算法的 ε-CLOSURE
 - GOTO(I,X): I的X后继
 - □ 对应于DFA化算法的转换边
- □ LR(0)项集族即LR(0)自动机的状态集合

LR(0)项



□ 产生式 $A \rightarrow XYZ$, 有下面四个项:

$$A \rightarrow \bullet XYZ \qquad A \rightarrow X \bullet YZ$$

$$A \rightarrow XY \bullet Z \qquad A \rightarrow XYZ \bullet$$

□ LR(0)项可分为几类:

■
$$A \rightarrow \alpha \bullet B\beta$$
, $B \in V_N$ 移进项

■
$$S' \rightarrow S$$
 • 接受项

- 对应于产生式 $A \rightarrow \epsilon$ 的唯一项是 $A \rightarrow \bullet$,它是归约项。
- □项也可以用一对整数表示。
 - (i,j)表示第i条规则,点位于右部第j个位置

计算 closure(I)



- 1. 把I中每一个项都加到初始为空的closure(I)中
- 2. 若项 $A \rightarrow \alpha \cdot B\beta \in closure(I)$ 且 $B \rightarrow \gamma \in P$ 则把 $B \rightarrow \bullet \gamma$ 加进 closure(I)中。
- 3. 反复执行(2)直到closure不再增大为止。

上面第二步的分析含义:

- 显然,要归约到B,首先要扫描归约到B的某个 产生式的右部:
- 因此对每个产生式 $B \rightarrow \gamma$, 加入 $B \rightarrow \bullet \gamma$;
 - 表示它期望能够扫描归约到γ。





```
SetOfItems CLOSURE(I) {
     J=I;
     repeat
           for (J中的每个项 A \to \alpha \cdot B\beta)
                 for (G 的每个产生式B \to \gamma)
                      将 B \rightarrow \cdot \gamma 加入J中;
     until 在某一轮中没有新的项被加入到J中;
     return J;
```

闭包构造的例子



例: (0) E'→E

$$(1) \quad E \rightarrow E + T$$

$$(2) \mid E \rightarrow T$$

$$(3) \mid T \rightarrow T*F$$

$$(4) \mid T \rightarrow F$$

$$(5) \overline{F \rightarrow (E)}$$

$$(6) \mid F \rightarrow id$$

项集{[E'→•E]}的闭包

$$E' \rightarrow \bullet E$$

$$E \rightarrow \bullet E + T$$

$$E \rightarrow \bullet T$$

$$T \rightarrow T*F$$

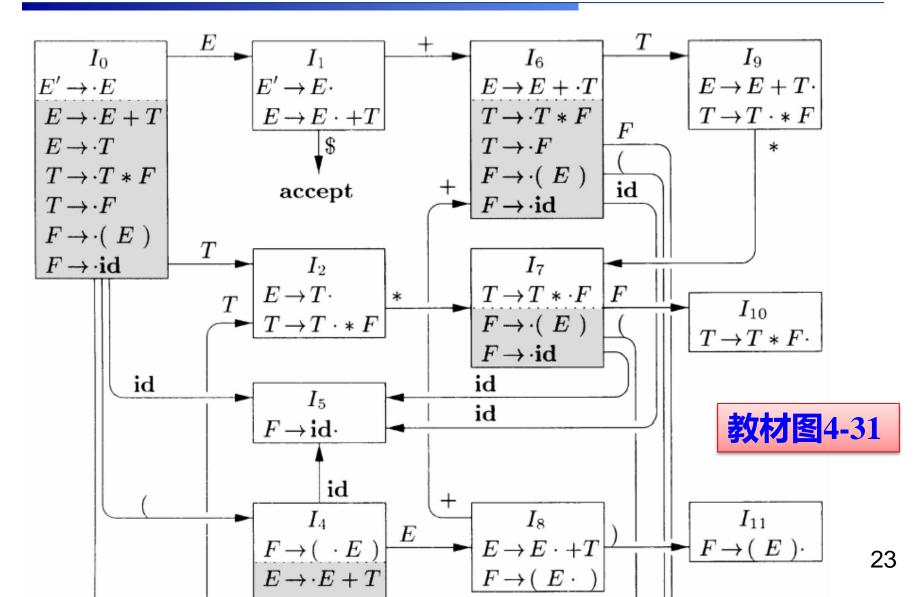
$$T \rightarrow \bullet F$$

$$F \rightarrow \bullet (E)$$

$$F \rightarrow \bullet id$$

构造LR(0)项集族与LR(0)自动机





LR(0)项集中的内核项和非内核项



- □ 在算法中,在加入某个项B→•γ时,所有以 B为左部的产生式所对应的(点在最左边的)项都会被加入
- □ 内核项:初始项S'→·S、以及所有点不在最 左边的项
- □ 非内核项:除了 $S' \rightarrow S$ 之外、点在最左边的项;
- □ 实现算法时可以考虑只保存内核项,而在要使用非内核项时调用CLOSURE函数重新计算。

GOTO函数



- □ GOTO(I,X): 项集{ $[A \rightarrow \alpha X \cdot \beta] \mid [A \rightarrow \alpha \cdot X \beta] \in I$ } 的闭包
 - 根据项的含义,GOTO(I,X)表示读取输入中的X或者归约到一个X之后的情况。
 - GOTO(I,X)定义了LR(0)自动机中状态I在X之上的 转换。

□ 例如:

- $= I = \{ [E' \rightarrow E \bullet], [E \rightarrow E \bullet + T] \};$
- GOTO(I,+)计算如下:
 - □ I中只有一个项的点后面跟着+,对应的项为[E→E+•T]
 - $\Box CLOSURE(\{[E \rightarrow E + \bullet T]\}) = \{[E \rightarrow E + \bullet T], [T \rightarrow \bullet T * F], [T \rightarrow \bullet F], [F \rightarrow \bullet (E)], [F \rightarrow \bullet id]\}.$

求LR(0)项集族的算法



```
\mathbf{void}\ items(G')\ \{\ C = \mathrm{CLOSURE}(\{[S' 	o \cdot S]\});\ \mathbf{repeat}\ \mathbf{for}\ (C + \mathbf{phh eta} \wedge \mathbf{phh} )\ \mathbf{for}\ (\mathbf{a} \wedge \mathbf{phh} \wedge \mathbf{phh} )\ \mathbf{for}\ (\mathbf{a} \wedge \mathbf{phh} \wedge \mathbf{phh} \wedge \mathbf{phh} )\ \mathbf{for}\ (\mathbf{a} \wedge \mathbf{phh} \wedge
```

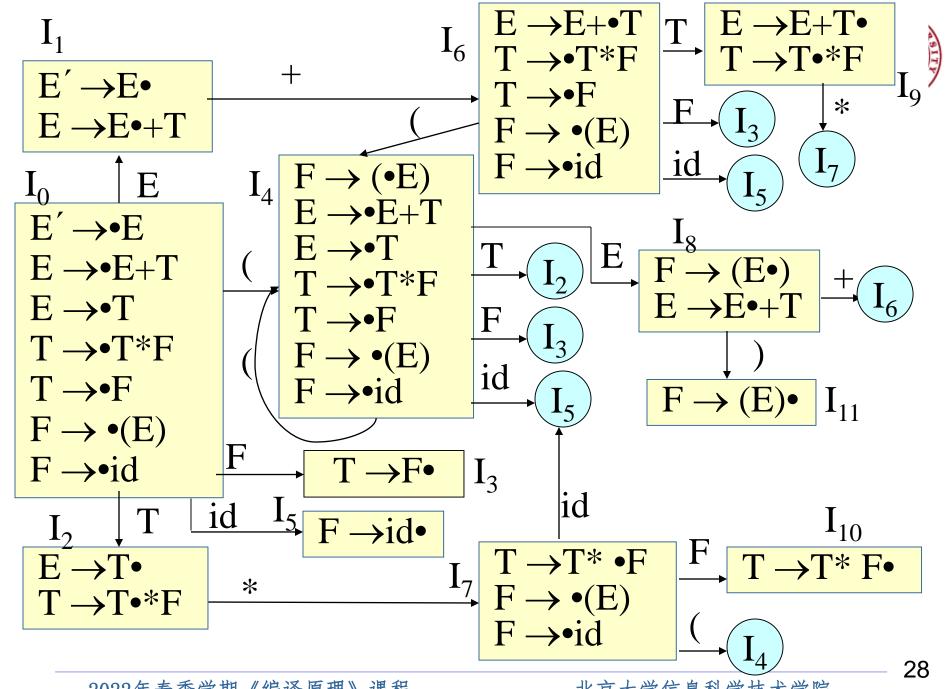
- □ 从初始项集开始,不断计算各种可能的后继,直到生成所有的项集。
- □ 可以和NFA的子集构造算法类比

LR(0)自动机的构造



□ 构造方法:

- LR(0)项集族中的项集可以作为LR(0)自动机的 状态
- GOTO(I,X)=J,则从I到J有一个标号为X的转换
- 初始状态为CLOSURE({S'→·S})对应的项集
- 接受状态:包含形如A → α•的项集对应的状态



LR(0)自动机的作用(1)



- □ 假设文法符号串γ使LR(0)自动机从开始状态运行到状态(项集)j
- □ 如果j中有一个形如A→ α •的项。那么
 - 在γ之后添加一些终结符号可以得到一个最右句型,
 - α是γ的后缀,且A→α是这个句型的句柄
 - 表示可能找到了当前最右句型的句柄
- □ 如果j中存在一个项B→ α • $X\beta$ 。那么
 - 在γ之后添加Xβ,然后再添加一个终结符号串可得 到一个最右句型。
 - 在这个句型中B→αXβ是句柄
 - 此时表示还没有找到句柄, 需要移进

LR(0)自动机的作用(2)



- □ LR(0)自动机的使用
 - 移进-归约时, LR(0)自动机被用于识别句型
 - 已经归约/移进得到的文法符号序列对应于LR(0)自动机的一条路径;
 - 如果路径到达接受状态,表明栈上端的某个符号串可能是句柄。
- □ 不需要每次用归约/移进得到的串来运行LR(0)自动机
 - 文法符号可以省略,因为由LR(0)状态可以确定相应文法 符号
 - 在移进后,根据原来的栈顶状态即可知道新的状态;
 - 在归约时,根据归约产生式的右部长度弹出相应状态,仍然可以根据此时的栈顶状态计算得到新状态。

LR(0)的作用演示:分析id*id



行号	栈	符号	输入	动作
(1)	0	\$	id * id \$	移入到 5
(2)	0.5	\$ i d	$*\operatorname{id}\$$	接照 $F \rightarrow id$ 归约
(3)	0.3	\$F	* id $$$	按照 $T \rightarrow F$ 归约
(4)	0 2	$\$ T	* id $$$	移入到 7
(5)	027	\$T*	$\mathbf{id}\$$	移入到 5
(6)	0275	$\$T*\mathbf{id}$	\$	接照 $F \rightarrow id$ 归约
(7)	0 2 7 10	\$T*F	\$	接照 $T \rightarrow T * F$ 归约
(8)	0 2	\$T	\$	接照 $E \rightarrow T$ 归约
(9)	01	\$E	\$	接受

□ 根据LR(0)自动机状态和符号的对应关系,可以得到符号栏中的符号串。

使用图4-31的LR(0)自动机

LR分析的结构

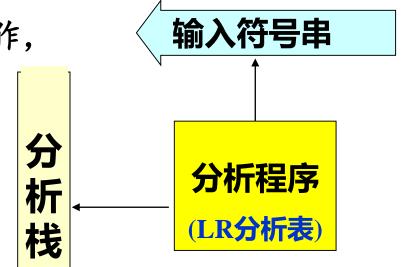


- □ LR分析表(依赖于具体文法)
 - 由两个矩阵组成,其功能是指示分析器的动作是移 进还是归约,根据不同的文法类要采用不同的构造 方法
- □ 驱动程序

□ 执行分析表所规定的动作,

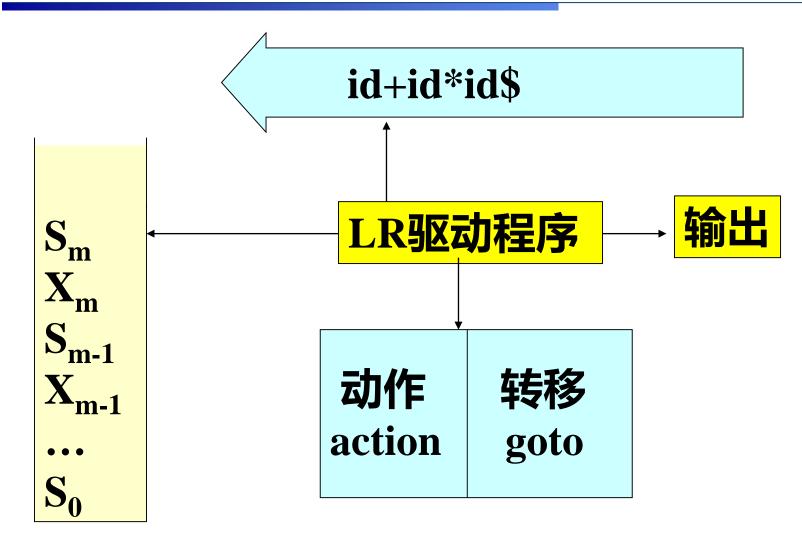
对栈进行操作。

- 口 分析栈
 - □ 放置分析器状态 和文法符号。



LR分析器的结构





LR分析表



□ 分析表的第一栏是状态、第二栏是action部分,由(|T|+1)列构成、第三栏是goto部分,由|V|列构成:

由
$$|V|$$
列构成:
$$action[s,a] = \begin{cases} 8 \text{ 进}S_i & \text{状态}i(字符a) \text{进栈} \\ \text{归约 }r_j: & \text{出栈k项} \\ \text{接受 } & \text{Angoto}[s',A] \text{进栈} \\ \text{出错} \end{cases}$$
 其中, s 是状态, a 是读入的终结符(单词)或 s :

其中,s是状态,a是读入的终结符(单词)或\$; k是j号产生式 $A \rightarrow \beta$ 右部(β)的长度; s'是出栈k项后新的栈顶元素中的状态。 表达式文法的SLR分析表

1. E→E+T 2. E→ T

3. T→T*F

4. T→**F**

5. F→(E)

6. **F**→ id

			ACT	TION				GOTO		
卷	id	+	*	()	\$	Е	Т	F	
0	S ₅			S ₄			1	2	3	
1		S ₆				acc				
2		r ₂	S ₇		r_2	r_2				
3		r ₄	r ₄		r ₄	r ₄				
4	S ₅			S ₄			8	2	3	
5		r ₆	r ₆		r ₆	r ₆				
6	S ₅			S ₄				9	3	
7	S ₅			S ₄					10	
8		S ₆			S ₁₁			_1.5		
9		r ₁	S ₇		r ₁	r ₁		疋-	片图4-3	7
10		r ₃	r ₃		r ₃	r ₃		使用	图4-31	的
11		r ₅	r ₅		r ₅	r ₅		LR((0)自动	机

2022年春季学期《编译原理》课程

北京大学信息科学技术学院

LR分析器的驱动程序



- □ 把状态 0 和符号 \$ 压入初始为空的栈里。
- □ 设栈顶元素中的状态为 s, 当前读入的符号为 a。
- □ 反复执行以下各动作,直到分析成功或发现语法错误为止:
 - 1 (移进) 若 $action[s, a]=S_i$,则把状态i(a)压进栈,读下一个输入符号到a中。
 - 2(归约)若action[s, a]= r_j (j: A $\rightarrow x_{m-k+1}x_{m-k+2}...x_m$),则出栈k项,把A和 s_{new} =goto[s', A] 进栈。其中 s'是出栈k项后新的栈顶元素中的状态。
 - 3(接受) 若action[s,\$]=accept, 则分析成功, 结束。
 - 4(出错) 若action[s,a]=error, 则转出错处理程序。

LR分析过程示例

使用龙书图4.37的分析表

栈	输入	动作
\$0	id+id*id\$	S_5
\$0id5	+id*id\$	$r_6 \rightarrow id$
\$0F3	+id*id\$	$r_4 \rightarrow F$
\$0T2	+id*id\$	$\mathbf{r_2} \ \mathbf{E} \rightarrow \mathbf{T}$
\$0E1	+id*id\$	S_6
\$0E1+6	id *id\$	S_5
\$0E1+6id5	*id\$	$r_6 \rightarrow id$
\$0E1+6F3	*id\$	$r_4 \rightarrow F$
\$0E1+6T9	*id\$	S ₇

LR分析过程示例(续)

使用龙书图4.37的分析表

栈	输入	动作
\$0E1+6T9*7	id \$	S_5
\$0E1 +6T9*7 id5	\$	$r_6 \rightarrow id$
\$0E 1+6T9*7 F <u>10</u>	\$	$r_3 \rightarrow T^*F$
\$0E1+6T9	\$	$r_1 \xrightarrow{E \to E + T}$
\$0E1	\$	accept

LR分析的关键是构造LR(k)分析表。

LR 分析表的种类



- □ LR(0) 分析表
 - 过于简单,只能用于最简单的文法,不实用
- □ SLR分析表 (Simple LR)
 - 构造简单,最易实现,大多数上下文无关文法都可以构造出SLR分析表,具有较高的实用价值
- □ LR(1)分析表 (Canonical LR)
 - 适用文法类最大,几乎所有上下文无关文法都能构造 出LR分析表,但其分析表体积太大,实用价值不大
- □ LALR(1)分析表 (LookAhead LR)
 - 这种表适用的文法类及其实现上难易在上面两种之间,较为实用

构造LR(0)分析表



- □ 根据DFA构造LR(0)分析表M:
 - DFA中的每个状态对应分析表中的一行
 - 对于DFA中的每一个从状态 i 到状态j的转移
 - \square 如果转移符号为终结符a:在相应的表项 M[i,a] 中填写移进动作 S_i
 - □ 如果转移符号为非终结符A:在相应的表项 M[i, A] 中填写要转移到的状态 j
 - 对于包含归约项A→α•的状态i
 - \square 对于所有终结符 a,在相应的表项 M[i,a] 中填写归约动作(r_k),其中 k 是产生式A
 ightarrow lpha的编号。

如果按照上面的步骤填写的分析表中没有冲突(即每个单元格中只包含一个动作),那么得到的就是合法的 LR(0)分析表

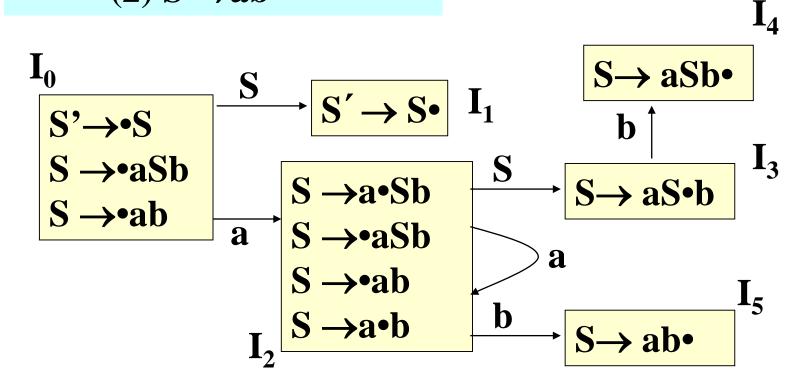
例: LR(0)分析表的构造



例: (0) S´→S

 $(1) S \rightarrow aSb$

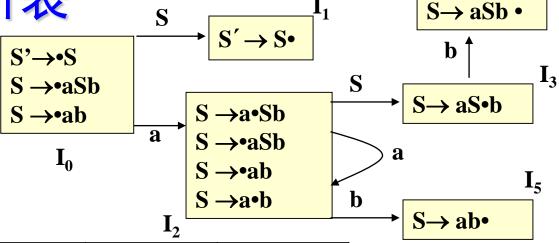
 $(2) S \rightarrow ab$



例: LR(0)分析表

文法: (0) S´→S

- $(1) S \rightarrow aSb$
- $(2) S \rightarrow ab$



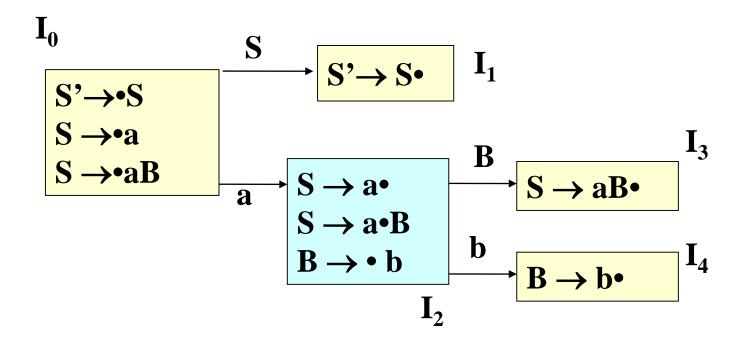
	a	b	\$	S
0	S ₂			1
1			acc	
2	S_2	S_5		3
3		S_4		
4	\mathbf{r}_1	\mathbf{r}_1	\mathbf{r}_1	
5	\mathbf{r}_2	\mathbf{r}_2	\mathbf{r}_2	

 I_4

LR(0)分析表中的冲突



- □ 假设我们为 $S \rightarrow a \mid aB \mid B \rightarrow b$ 构造 LR(0)项集。
 - (0) S' \rightarrow S (1) S \rightarrow a (2) S \rightarrow aB (3) B \rightarrow b
- □ I₂状态中出现了移进-归约冲突。



解决办法: SLR分析表



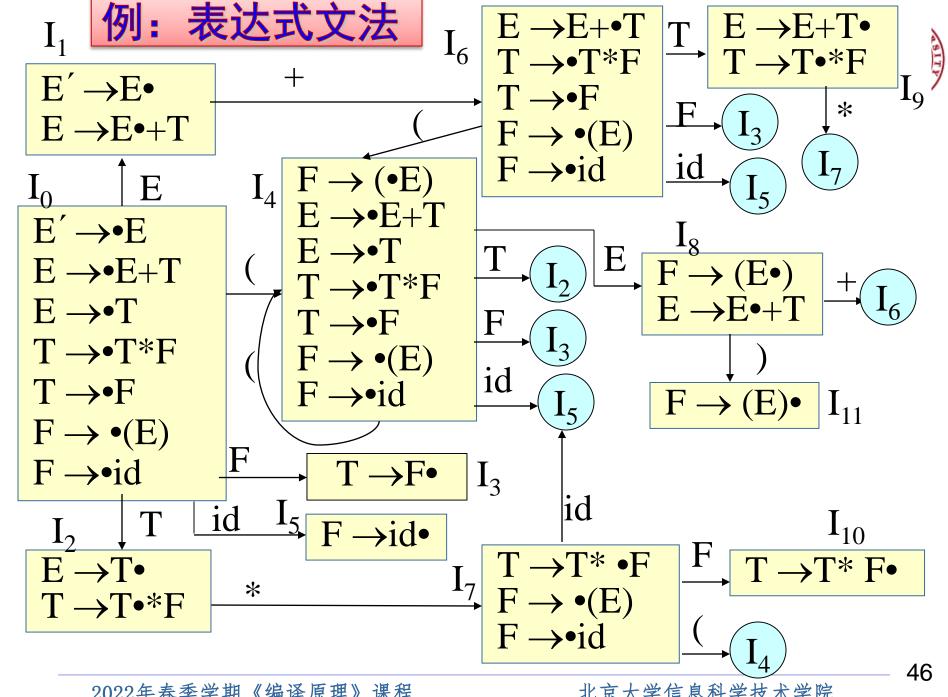
- □ 根据Follow集来选择是否要进行归约。
- □ 如果 $I=\{X \rightarrow \alpha \cdot b \beta, A \rightarrow \alpha \cdot, B \rightarrow \alpha \cdot\}$ 若 $\{b\}$, Follow(A), Follow(B)两两不交,则面对当前读入符号a,状态I的解决方法:
 - 1. 若a=b, 则移进。
 - 2. 若a∈ Follow(A), 则用A →α 进行归约。
 - 3. 若a∈ Follow(B), 则用B →α进行归约。
 - 4. 此外,报错。
- □ 这种解决方法比较简单,因此称作SLR分析, 由此构造的分析表,称作SLR分析表。

构造SLR分析表



- □ 根据LR(0)的DFA构造SLR分析表M:
 - 为所有非终结符计算相应的Follow集合
 - 对于包含归约 $A \rightarrow \alpha$ 。的状态 i
 - □ 为所有终结符 $a \in Follow(A)$, 在相应的表项 M[i, a] 中填写归约动作 r_i , 其中 i 是产生式 $A \rightarrow \alpha$ 的编号。
 - 所有其他表项都按照与LR(0)分析表相同的方式

如果按照上面的步骤填写的分析表中没有冲突(即每个单元格中只包含一个动作),那么得到的就是合法的SLR分析表



计算Follow集



存在冲突的项集:

$$I_{1}:\{E'\rightarrow E\bullet E\rightarrow E\bullet +T\}$$

$$I_{2}:\{E\rightarrow T\bullet T\rightarrow T\bullet *F\}$$

$$I_{9}:\{E\rightarrow E+T\bullet T\rightarrow T\bullet *F\}$$

文法的Follow集:

该文法的SLR分析表参见龙书图4.37。