

语法分析II

《编译原理和技术(H)》、《编译原理(H)》

张昱

0551-63603804, yuzhang@ustc.edu.cn 中国科学技术大学 计算机科学与技术学院

- □ 自上而下(top-down)
 - 从开始符号(根结点)出发, 自上而下、从左到右地为输入串寻找最左推导
 - 即便消除左递归、提取左公因子,仍然存在一些程序语言,它们对应的文法不是LL(1)

- □ 自下而上(bottom-up)
 - 针对输入串,尝试根据产生式规则归约(reduce,将与产生式右部匹配的串归约为左部符号),直至归约到开始符号
 - 比top-down方法更一般化



3.3 自上而下解析

- □自上而下解析、左递归及消除、提左因子
- □ LL(1)文法、LL(k)文法和ANTLR支持的文法
- □ 预测解析器: 递归下降的、非递归的(预测分析表)
- □ 错误恢复



自上而下解析的一般方法



□ 自上而下(top-down)解析

从开始符号出发,为输入串寻找最左推导

试探产生式的选择-失败回溯(效率低,代价高)

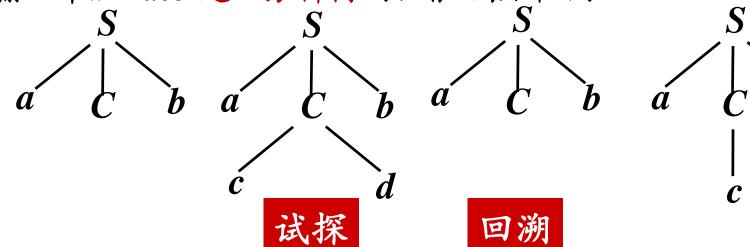
ANTLR: 引入带谓词的DFA使得回溯后不重新分析输入串

例 文法

$$S \rightarrow aCb$$

$$S \rightarrow aCb$$
 $C \rightarrow cd/c$

为输入串w = acb 建立分析树(从根结点开始)





自上而下解析: 左递归



□ 文法左递归

若文法中存在某非终结符A,使得 $A \rightarrow + A \alpha$,其中 α 是文法符号串

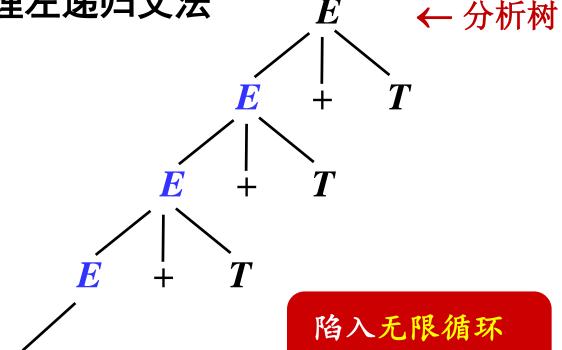
□ 自上而下解析不能处理左递归文法

算术表达式文法

$$E \rightarrow E + T \mid T$$

$$T \rightarrow T * F \mid F$$

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



(::文法左递归)

消除左递归(Eliminating left recursion)



- □ 文法左递归
 - 若文法中存在某非终结符A,使得 $A \rightarrow +A \alpha$,其中 α 是文法符号串
- □ 直接左递归 (immediate left recursion)

若文法中存在 $A \rightarrow A\alpha \mid \beta$, 其中文法符号串 β 不以A 开头

■ 由A推出的串的特点 $A \Rightarrow^+ \beta \alpha \dots \alpha$

$$A \Rightarrow^+ \beta \alpha \dots \alpha$$

□ 消除直接左递归

文法
$$A \rightarrow A\alpha \mid \beta$$
, 可写成 $A \rightarrow \beta A'$ $A' \rightarrow \alpha A' \mid \epsilon$

例 算术表达式文法

$$E \rightarrow E + T \mid T$$

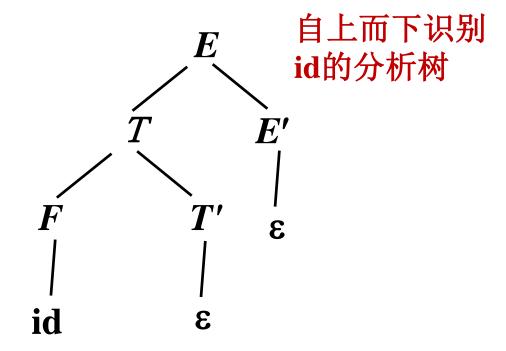
$$T \rightarrow T * F \mid F$$

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

消除左递归后的文法

$$E \rightarrow TE'$$
 $E' \rightarrow + TE' \mid \varepsilon$
 $T \rightarrow FT'$
 $T' \rightarrow *FT' \mid \varepsilon$
 $F \rightarrow (E) \mid id$

$$E \Rightarrow^{+} T + T \dots + T$$
$$T \Rightarrow^{+} F * F \dots * F$$





消除非直接左递归

□ 间接左递归

$$S \rightarrow Aa \mid b$$

$$A \rightarrow Sd \mid \varepsilon$$

□ 先变换成直接左递归

$$S \rightarrow Aa/b$$

$$A \rightarrow Aad/bd \mid \varepsilon$$

□ 再消除左递归

$$S \rightarrow Aa/b$$

$$A \rightarrow bdA'/A'$$

$$A' \rightarrow adA' / \epsilon$$

□ 隐藏左递归

$$S \rightarrow ABc$$

$$A \rightarrow \varepsilon$$

$$B \rightarrow ABe / Cd$$

$$C \rightarrow Sf$$

 $B \Rightarrow ABe \Rightarrow Be$ 是隐藏直接左递归,而 $S \Rightarrow ABc \Rightarrow Bc \Rightarrow Cdc \Rightarrow Sfdc$ 是隐藏间接左递归。

ANTLR4接受的文法包含 直接左递归,但不包含间 接或隐藏的左递归



提左因子(left factoring)



- □ 有左因子的(left-factored)文法: $A \rightarrow \alpha \beta_1 / \alpha \beta_2$
 - 自上而下解析时,不清楚应该用A的哪个选择来代换
- □ 提左因子

$$A \rightarrow \alpha A'$$

$$A' \rightarrow \beta_1 / \beta_2$$

例 悬空else的文法

 $stmt \rightarrow if expr then stmt else stmt$

if expr then stmt | other

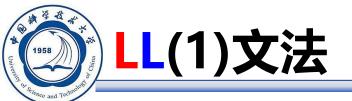
提左因子

 $stmt \rightarrow if \ expr \ then \ stmt \ optional_else_part \ | \ other \ optional_else_part \rightarrow else \ stmt \ | \ \epsilon$



3.3 自上而下解析

- □自上而下解析、左递归及消除、提左因子
- □ LL(1)文法、LL(k)文法和ANTLR支持的文法
- □ 预测解析器: 递归下降的、非递归的(预测分析表)
- □错误恢复



- L-scanning from left to right; L- leftmost derivation
- □ 对文法加什么样的限制可以保证没有回溯?
- □ 先定义两个和文法有关的函数
 - 符号串α的开始符号集合FIRST(α) FIRST(α) = { $a \mid \alpha \Rightarrow^* a..., a \in V_T$ } \cup { $\epsilon \mid \alpha \Rightarrow^* \epsilon$ }
 - $\uparrow \alpha$ 推出的非空串中首终结符和 α 推出空串时的 ϵ 组成的集合
 - A的后继符号集合FOLLOW(A)

FOLLOW(A) = $\{a \mid S \Rightarrow^* ... Aa..., a \in V_T\} \cup \{\$ \mid S \Rightarrow^* ... A\}$

 $\uparrow A$ 的后继符号集合:在句型中紧跟在A之后的终结符或\$组成的集合

这里约定\$为输入记号串的结束标志,S是开始符号

- □ 计算FIRST(X), $X \in V_T \cup V_N$
 - $X \in V_T$, FIRST(X) = {X}
 - $X \in V_N$ 且 $X \to Y_1 Y_2 ... Y_k$ 如果 $a \in FIRST(Y_i)$ 且ε 在FIRST (Y_1) , ..., FIRST (Y_{i-1}) 中,则将 a加入到 FIRST(X)

如果 ε 在FIRST (Y_1) , ..., FIRST (Y_k) 中,则将ε 加入到FIRST(X)

 $X \in V_N$ 且 $X \to \varepsilon$,则将 ε 加入到 FIRST(X)

FIRST集合只可能包含终结符和ε

□ 计算FIRST $(X_1 X_2 ... X_n)$,

$$X_i \in V_T \cup V_N$$
,它包含

- FIRST(X₁) 中所有的非 ε 符号
- FIRST(X_i) 中所有的非 ε 符号,如果ε 在FIRST(X_1), ...,
 FIRST(X_{i-1})中
- ε,如果ε在FIRST(X₁),..., FIRST(X_n)中

例
$$E \rightarrow TE'$$

$$E' \rightarrow + TE' \mid \varepsilon$$

$$T \rightarrow FT'$$

$$T' \rightarrow *FT' \mid \varepsilon$$

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

$$FIRST(E) = FIRST(T) = FIRST(F) = \{ (, id) \}$$

FIRST(
$$E'$$
) = {+, ε }

$$\mathbf{FRIST}(T') = \{*, \varepsilon\}$$



- □ 计算FOLLOW(A), $A \in V_N$
 - \$加入到FOLLOW(S)中
 - 如果 $A \rightarrow \alpha B\beta$, 则将 FIRST(β) 中所有终结符加入到 FOLLOW(B)
 - 如果 $A \rightarrow \alpha B$ 或 $A \rightarrow \alpha B\beta$ 且 ε ∈ FIRST(β), 则FOLLOW(A)中的 所有符号加入到FOLLOW(B)

 $FOLLOW(A) \subseteq V_T \cup \{\$\}$

例
$$E \rightarrow TE'$$
 $E' \rightarrow + TE' \mid \varepsilon$
 $T \rightarrow FT'$
 $T' \rightarrow *FT' \mid \varepsilon$
 $F \rightarrow (E) \mid id$

FIRST(E) = FIRST(T) = FIRST(F) = { (, id)}
FIRST(E') = {+,
$$\varepsilon$$
}
FRIST(T') = {*, ε }
FOLLOW(E) = FOLLOW(E') = {), \$}
FOLLOW(T) = FOLLOW(T') = {+,), \$}
FOLLOW(F) = {+, *,), \$}

□ LL(1)文法的定义

任何两个产生式 $A \rightarrow \alpha/\beta$ 都满足下列条件:

- $FIRST(\alpha) \cap FIRST(\beta) = \emptyset$

例 对于下面文法,面临a...时,第2步推导不知用A的哪个产生式选择

$$S \rightarrow A B$$

$$A \rightarrow a b \mid \varepsilon$$

$$B \rightarrow a C$$

$$C \rightarrow \dots$$

$$a \in FIRST(ab) \cap FOLLOW(A)$$

□ LL(1)文法的定义

任何两个产生式 $A \rightarrow \alpha/\beta$ 都满足下列条件:

- $FIRST(\alpha) \cap FIRST(\beta) = \emptyset$

□ LL(1)文法有一些明显的性质

- 没有公共左因子
- 不是二义的
- 不含左递归

□ LL(1)文法的定义可以扩展到LL(k)

任何两个产生式 $A \rightarrow \alpha/\beta$ 都满足下列条件:

- **■** FIRST(k, α) \cap FIRST(k, β) = \emptyset

FIRST (k, α) 表示文法符号串 α 的前k个终结符串集合 FOLLOW(k, A)表示跟在A后的前k个终结符串集合

- L- scanning from left to right 从左到右扫描输入
- L- leftmost derivation 产生最左推导
- k-决定解析器的每步动作时需要向前查看k个输入符号 (即输入指针所指向的符号)



3.3 自上而下解析

- □自上而下解析、左递归及消除、提左因子
- □ LL(1)文法、LL(k)文法和ANTLR支持的文法
- □ 预测解析器: 递归下降、非递归(预测分析表)
- □错误恢复

- □ 预测解析
 - 根据面临的输入符号, 预测用非终结符的哪个选择进行展开(推导)
- □ 递归下降(recursive-descent)的预测解析
 - 为每一个非终结符写一个解析函数
 - 这些函数可能是递归的

例 下面的文法产生Pascal语言的类型子集

```
type → simple

| ↑ id
| array [simple] of type
simple → integer
| char
| num dotdot num
```

递归:分析函数可能是递归的

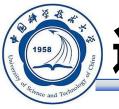
下降: 自上而下



递归下降的预测解析器



```
type \rightarrow simple \mid \uparrow id \mid array [simple] of type
void match (terminal t) {
                                                      simple \rightarrow integer | char | num dotdot num
    if (lookahead == t) lookahead = nextToken( );
    else error( );
void type( ) {
    if ( (lookahead == integer) || (lookahead == char) || (lookahead == num) )
         simple();
    else if ( lookahead == '\uparrow' ) { match('\uparrow'); match(id); }
    else if (lookahead == array) {
         match(array); match('['); simple();
         match(']'); match(of); type();
    else error( );
```



递归下降的预测解析器



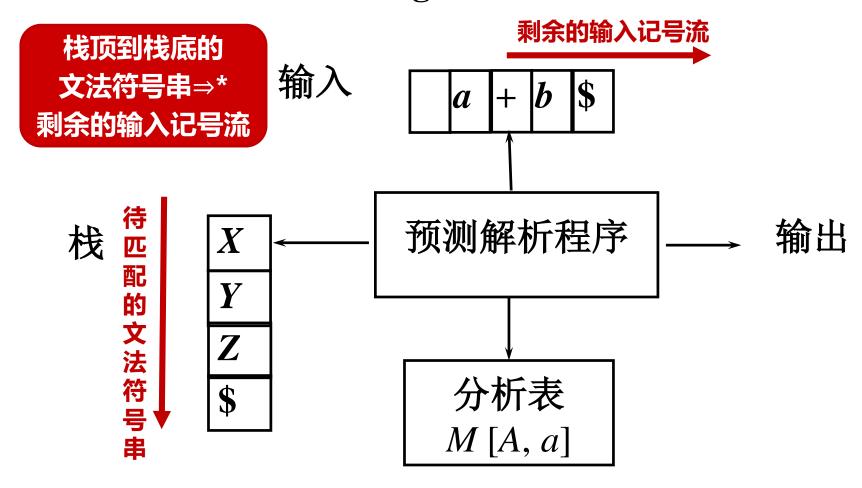
```
void simple() {
    if ( lookahead == integer) match(integer);
    else if (lookahead == char) match(char);
    else if (lookahead == num) {
        match(num); match(dotdot); match(num);
    }
    else error();
}
```

 $type \rightarrow simple \mid \uparrow id \mid array [simple] of type$ $<math>simple \rightarrow integer \mid char \mid num \ dotdot \ num$

递归下降的预测解析程序的构造 与 文法结构直接对应 →编程简单



Nonrecursive Predictive Parsing



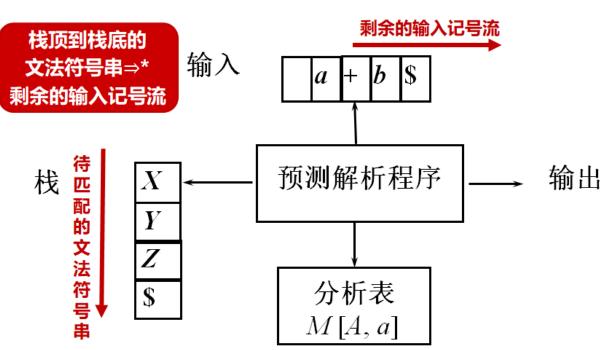


非递归的预测解析算法



```
让输入指针ip指向剩余输入记号流w$的第一个符号a;
令X等于栈顶符号;
while (X != $ ) { /* 栈非空 */
 if(X \to a) 把X从栈顶弹出并把ip推进到指向下一个符号;
 else if (X是终结符) error ();
 else if (M[X, a]是出错入口) error ();
 else if (M[X, a] = X \rightarrow Y_1Y_2 \dots Y_k) {
   输出产生式X \rightarrow Y_1 Y_2 \dots Y_k;
    从栈中弹出X;
   把Y_k, Y_{k-1}, ..., Y_1依次压入栈,Y_1在栈顶;
```

 $\diamondsuit X$ 等于栈顶符号;







- □ 行: 非终结符; 列: 终结符或\$; 单元: 产生式
- □ 教材 表3.1

非终结符	输入符号						
	id	+	*	()	\$	
E	$E \rightarrow TE'$			$E \rightarrow TE'$			
E'		$E' \rightarrow +TE'$			$E' \rightarrow \varepsilon$	$E' \rightarrow \varepsilon$	
\overline{T}	$T \rightarrow FT'$			$T \rightarrow FT'$			
$\overline{T'}$		$T' \rightarrow \varepsilon$	$T' \rightarrow *FT'$		$T' \rightarrow \varepsilon$	$T' \rightarrow \varepsilon$	
\overline{F}	$F \rightarrow id$			$F \rightarrow (E)$			



栈	输入	输出
\$ <i>E</i>	id * id + id\$	



预测解析举例

预测解析器接受输入id * id + id的前一部分动作

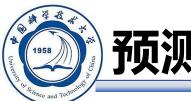
栈	输入	输出
\$ <i>E</i>	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$



预测解析举例

预测解析器接受输入id * id + id的前一部分动作

栈	输入	输出
\$ <i>E</i>	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$
\$E 'T 'F	id * id + id\$	$T \rightarrow FT'$



预测解析举例

预测解析器接受输入id * id + id的前一部分动作

栈	输入	输出
\$E	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$
\$E 'T 'F	id * id + id\$	$T \rightarrow FT'$
<i>E 'T'</i> id	id * id + id\$	$F o \mathrm{id}$



栈	输入	输出
\$ <i>E</i>	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$
\$E 'T 'F	id * id + id\$	$T \rightarrow FT'$
<i>\$E 'T '</i> id	id * id + id\$	$F \rightarrow id$
\$E 'T'	* id + id\$	



栈	输入	输出
\$ <i>E</i>	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$
\$E 'T 'F	id * id + id\$	$T \rightarrow FT'$
\$E 'T' id	id * id + id\$	$F \rightarrow id$
\$E 'T'	* id + id\$	
\$E 'T 'F *	* id + id\$	$T' \rightarrow *FT'$



栈	输入	输出
\$ <i>E</i>	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$
\$E 'T 'F	id * id + id\$	$T \rightarrow FT'$
<i>\$E 'T '</i> id	id * id + id\$	$F \rightarrow id$
\$E 'T'	* id + id\$	
\$E 'T 'F *	* id + id\$	$T' \rightarrow *FT'$
\$E 'T 'F	id + id\$	





栈	输入	输出
<i>\$E</i>	id * id + id\$	
\$E 'T	id * id + id\$	$E \rightarrow TE'$
\$E 'T 'F	id * id + id\$	$T \rightarrow FT'$
<i>\$E 'T '</i> id	id * id + id\$	$F \rightarrow id$
\$E 'T'	* id + id\$	
\$E 'T 'F *	* id + id\$	$T' \rightarrow *FT'$
\$E 'T 'F	id + id\$	
<i>\$E 'T '</i> id	id + id\$	$F \rightarrow id$

预测分析表的构造



□ 预测分析表(predictive parsing table)

行: 非终结符; 列: 终结符 或\$; 单元: 产生式 M[A,a]产生式 $A \to \alpha$ 表示在面临a 时, 将栈顶符号A替换为 α

□ 构造方法

- (1) 对文法的每个产生式 $A \rightarrow \alpha$,执行(2)和(3)
- (2) 对FIRST(α)的每个终结符a, 把 $A \rightarrow \alpha$ 加入M[A, a]
- (3) 如果 ε 在FIRST(α)中,对FOLLOW(A)的每个终结符b (包括\$),把 $A \to \alpha$ 加入M[A,b]
 - (4) M中其它没有定义的条目都是error



例 stmt → if expr then stmt e_part | other

 $e_part \rightarrow else\ stmt \mid \varepsilon$

 $expr \rightarrow b$

非终结符	输入符号						
	other	b	else	if	then	\$	
stmt	$stmt \rightarrow other$			$stmt \rightarrow if$			
e_part			$e_part \rightarrow else \ stmt$ $e_part \rightarrow \epsilon$			$e_part \rightarrow \varepsilon$	
expr		$expr \rightarrow b$					

多重定义条目意味着文法左递归或者是二义的



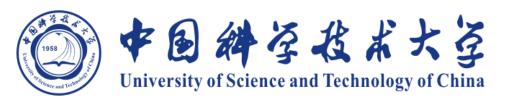


例 删去 $e_part \rightarrow \epsilon$, 这正好满足 else和最近的then配对

LL(1)文法 ⇔ 预测分析表无多重定义的条目

非终	输入符号					
结符	other	b	else	if	then	\$
stmt	$stmt \rightarrow other$			$ stmt \rightarrow if$		
e_part			$\begin{array}{c} e_part \rightarrow \textbf{else} \ stmt \\ e_part \rightarrow \varepsilon \end{array}$			$e_part \rightarrow \varepsilon$
expr		$expr \rightarrow b$				

没有通用的方法来指导多重定义条目的删除



3.3 自上而下解析

- □自上而下解析、左递归及消除、提左因子
- □ LL(1)文法和递归下降的预测解析器
- □ 非递归的预测解析器(预测分析表)
- □ 错误恢复

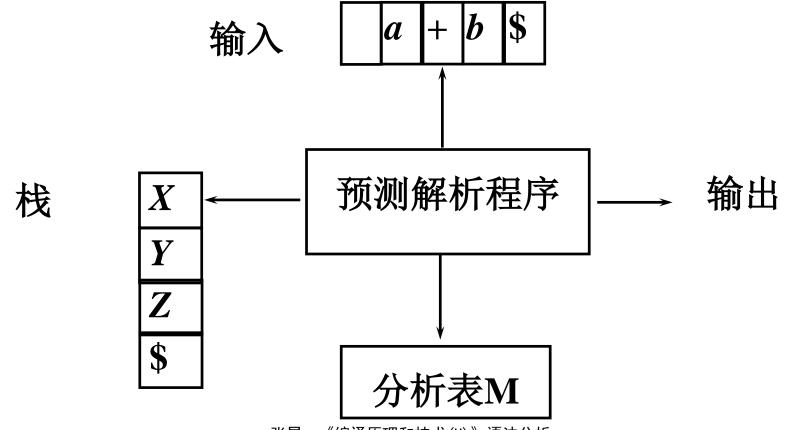
□ 编译器的错误处理

- 词法错误,如标识符、关键字或算符的拼写错
- 语法错误,如算术表达式的括号不配对
- 语义错误,如算符作用于不相容的运算对象
- 逻辑错误,如无穷的递归调用

□ 解析器对错误处理的基本目标

- 清楚而准确地报告错误的出现,并尽量少出现伪错误
- 迅速地从每个错误中恢复过来,以便诊断后面的错误
- 它不应该使正确程序的处理速度降低太多

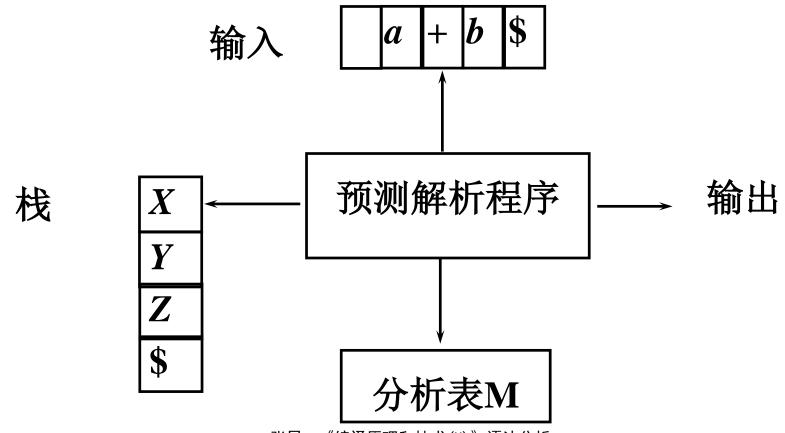
- □ 非递归预测解析在什么场合下发现错误
 - 栈顶的终结符和下一个输入符号不匹配







- □ 非递归预测解析在什么场合下发现错误
 - 栈顶是非终结符A,输入符号是a,而M[A,a]是空白



□ 非递归预测解析

采用紧急方式(panic mode)的错误恢复

■ 发现错误时, 抛弃输入记号直到其属于某个指定的同步记号(synchronizing tokens)集合为止

□ 同步(synchronizing)

- 同步: 词法分析器当前提供的记号流能够构成的语法构造, 正是语法分析器 所期望的
- 不同步的例子

语法分析器期望剩余的前缀构成过程调用语句,而实际剩余的前缀形成的是赋值语句





□ 同步记号集合的选择

■ 把FOLLOW(A)的所有终结符放入非终结符A的同步记号集合 if expr then stmt 同步记号

出错

(then和分号等记号是expr的同步记号)

■ 把高层构造的首终结符加到低层构造的同步记号集中

a = expr; if ...

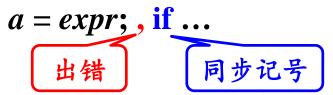
出错

同步记号

(语句的首终结符作为表达式的同步记号,以免表达式出错又遗漏分号时忽略if语句等一大段程序)

□ 同步记号集合的选择

- 把FOLLOW(A)的所有终结符放入非终结符A的同步记号集合
- 把高层构造的首终结符加到低层构造的同步记号集中
- 把FIRST(A)的终结符加入A的同步记号集合



(语句的开始符号作为语句的同步符号,以免多出一个逗号时会把if语句忽略了)

□ 同步记号集合的选择

- ν FOLLOW(A)的所有终结符放入非终结符A的同步记号集合
- 把高层构造的首终结符加到低层构造的同步记号集中
- 把FIRST(A)的终结符加入A的同步记号集合
- 如果出错时栈顶是存在有产生空串选择的非终结符,则可以使用其推出空串 的产生式选择



错误恢复举例

例 栈顶为T′, 面临id时出错

非终 结符		输入	符号	
结符 	id	+	*	• • •
E	$E \rightarrow TE'$			
E'		$E' \rightarrow +TE'$		
T	<i>T</i> → <i>FT</i> ′			
T '	出错	$T' \rightarrow \varepsilon$	$T' \rightarrow *FT'$	
• • •				



例 栈顶为T′, 面临id时出错

非终结符		输入	符号	
	id	+	*	• • •
$oldsymbol{E}$	$E \rightarrow TE'$			
E'		$E' \rightarrow +TE'$		
T	$T \rightarrow FT'$			
T '	出错	$T' \rightarrow \varepsilon$	$T' \rightarrow *FT'$	
	用 $T' \rightarrow ε$			
• • •				

□ 同步记号集合的选择

- 把FOLLOW(A)的所有终结符放入非终结符A的同步记号集合
- 把高层构造的首终结符加到低层构造的同步记号集中
- 把FIRST(A)的终结符加入A的同步记号集合
- 如果出错时栈顶是存在有产生空串选择的非终结符,则可以使用其推出空串 的产生式选择
- 如果终结符在栈顶而不能匹配,弹出此终结符



下期预告: ANTLR、自下而上的分析