第三章 词法分析

§3.1 设计扫描器时应考虑的问题

§3.2 正则文法和状态转换图

§3.3 有限自动机

§3.4 正则表达式和正规集

§3.5 词法分析程序的实现

教学任务

- 重点：
  1. 正则文法和状态转换图
  2. 有限自动机
  3. 正则表达式和正规集
  4. 词法分析程序的实现
- 难点

  正则文法和状态转换图、有限自动机

§ 3.1 设计扫描器时应考虑的几个问题

§ 3.1.1 词法分析器的功能与实现方式

1、功能

输入:符号串形式的源程序

输出:单词符号串

2 、实现方式

(1)词法分析作为单独的一遍

特点①大部分编译时间花在扫描字符上,

独立出来便于集中处理.

②单词的词法规则简单，可建立特别适用于这种文法的有效技术，实现词法分析的自动生成.

③整个编译程序结构简单，清晰，产生中间文件，占内存.

(2) 词法分析作为一个独立的子程序，供语法分析程序调用

特点：

①语法分析调用时，返回一个单词符号.

②无中间文件，省内存，编译效率高.

§ 3.1.2 源程序的输入及预处理

1、预处理

---构造预处理子程序（输入缓冲区）

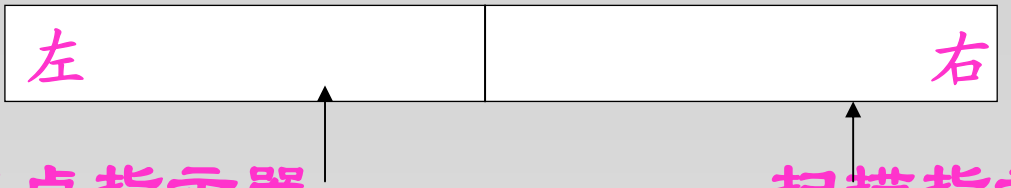
(1)作用：消除无用空白、回车、注释行、区分标号区、续行号（FORTRAN）等.

(2)功能：词法分析器调用时，在输入缓冲区内处理出确定长的字符串放入扫描缓冲区.

2、扫描缓冲区的结构：

两个半区，两个指示器，互补使用

扫描缓冲区



（读字符指示器）　（超前读字符指示器）

起点指示器：

指向当前正在识别的单词的开始位置

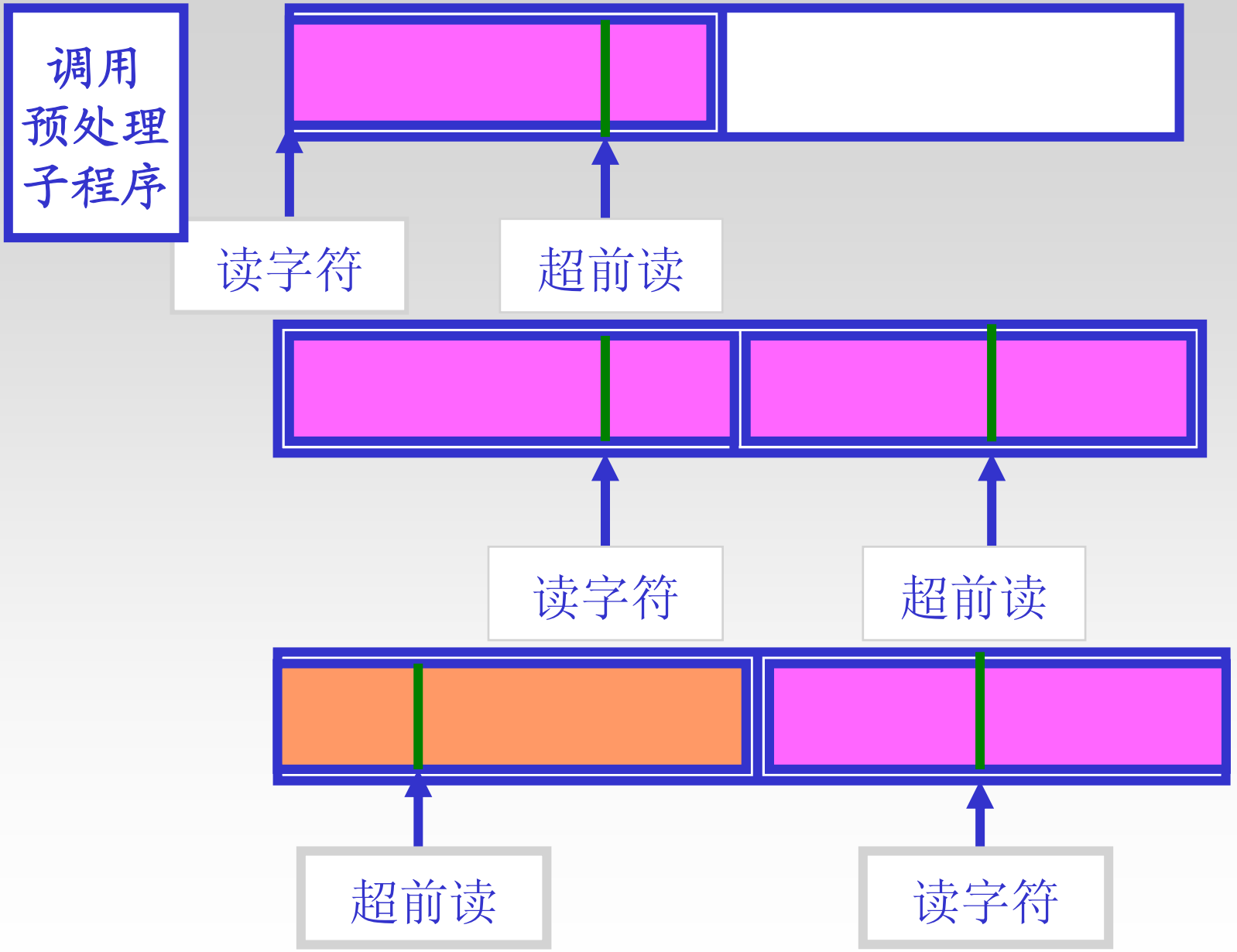
扫描指示器：

用于向前搜索以寻找单词的终点

两个半区互补使用：

规定单词的长度不能超过半区的长度。

例如 60


调用
预处理
子程序
读字符
超前读
读字符
超前读
超前读
读字符

§3.1.3 单词符号的内部表示

——词法分析器的输出形式

(1) 单词符号的种类

①保留字：如for, while, do, 等用户不能使用

②标识符：由用户定义

③无符号整数：如124

④单字符分界符：+, －, *, /, ;, ,, ,
(,), :, ,>,<,=

⑤双字符分界符：//, /*, **, :=, >=, <=,
<>, ==, ++等

(2) 单词符号的表示形式——二元式

二元式：（单词类别，单词自身值）

①单词类别：说明单词属于哪一类，

一般用整数编码表示.

例：标识符用4 表示

②单词自身值(2种情形)

- 一种类只有一个单词，不必给出单词自身值.

  因为种别编码能完全确定.

- 一种类含有多个单词，必须给出单词自身值

  予以区别。

一般：

①保留字，运算符和分界符各是一符一种，不需单词自身值.

如 ′ + ′ 类别8,

′ + ′ 的二元式 (8, -)

②标识符为一种类,其单词自身值采用自身的字符串编码表示.

符号表中的地址

如标识符类别为5, AB的二元式(5, AB)

③常数按类型分类别：单词自身值采用自身的二进制形式.

如整数类别为20, 4的二元式(20, 100)

§ 3.1.4 识别标识符的若干约定和策略

1、约定：

(1) 标识符中的字符个数超过最大允许长度，截去尾部.

(2) 不超过最大长度的标识符，则按“尽可能长”的原则匹配（Greedy Method:贪欲法）.

如：IENTIFIER（如果最大长度为6）

则识别为IENTIF标识符

〉和〉=

2. 超前搜索技术

所谓超前搜索技术（超前读字符）：是仅向前读取字符，和判别该字符是什么，不作处理. 当判明后再回过来处理已读过的字符。

例：FORTRAN（非标准）只设关键字，无保留字

①DO 99 K=1，10　（do语句）

②DO 99 K=1. 10　（do语句）

③IF（5 .EQ.M）　（逻辑IF语句）

④IF（5）=55　（赋值语句）

〉和 〉=

a4y=9

2. 零层等号和零层逗号

根据可嵌套的括号由外向内进行编号。

作用：超前搜索技术中作为某些语句的判定条件来使用

如：含有零层等号的有赋值语句、DO语句、语句函数定义句、某些逻辑IF语句等

基本思想：结合语句各自的特征寻找判定条件