编译器设计文档

——22373362杜启嵘

一、参考编译器介绍

二、编译器总体设计

> 使用cpp进行编译器实现

三、词法分析设计

3.1 初始设计

词法分析部分的主要任务是将读入的文件(字符串)进行一遍“字符串解析”，将读入的字符串中的单词按照种类识别出来。

单词名称	类别码	单词名称	类别码	单词名称	类别码	单词名称	类别码
Ident	IDENFR	else	ELSETK	void	VOIDTK	;	SEMICN
IntConst	INTCON	!	NOT	*	MULT	,	COMMA
StringConst	STRCON	&&	AND	/	DIV	(	LPARENT
CharConst	CHRCON	||	OR	%	MOD	)	RPARENT
main	MAINTK	for	FORTK	<	LSS	[	LBRACK
const	CONSTTK	getint	GETINTTK	<=	LEQ	]	RBRACK
int	INTTK	getchar	GETCHARTK	>	GRE	{	LBRACE
char	CHARTK	printf	PRINTFTK	>=	GEQ	}	RBRACE
break	BREAKTK	return	RETURNTK	==	EQL		
continue	CONTINUETK	+	PLUS	!=	NEQ		
if	IFTK	-	MINU	=	ASSIGN		

对于表格中给出的token种类，笔者使用一个枚举类型`TokenType`进行记录，该枚举类型内嵌在`Token`类中，`Token`类中有三个属性变量：token的字符串表示(`string`)，token的种类(`TokenType`)，token所属的行(`line_number`)，并定义其to_string方法用于输出。

```
public:
enum TokenType {
    IDENFR, INTCON, STRCON, CHRCON, MAINTK, CONSTTK, INTTK, CHARTK, BREAKTK, CONTINUETK,
    IFTK, ELSETK, NOT, AND, OR, FORTK, GETINTTK, GETCHARTK, PRINTFTK, RETURNTK, PLUS, MINU,
    VOIDTK, MULT, DIV, MOD, LSS, LEQ, GRE, GEQ, EQL, NEQ, ASSIGN, SEMICN, COMMA,
    LPARENT, RPARENT, LBRACK, RBRACK, LBRACE, RBRACE
};
```

词法分析由`Lexer`类完成，其**属性**定义为：

- source(string)：读入的程序字符串
- line_number(int)：当前分析到的行号
- pos(int)：当前分析的字符串索引位置
- errors(`vector<Error>`)：记录错误的数组(在词法分析阶段只有a类错误)

- tokens(`vector<TOken>`)：解析字符串得到的Token数组
- reverse_words(`unordered_map<std::string,Token::TokenType>`)：`SysY`保留字表

```
// lexer.h
private:
        std::string source;
        int pos;
        Token::TokenType token_type;
        int line_number;
        std::unordered_map <std::string, Token::TokenType> reserve_words;
        std::vector <Error> errors;  // 保存a类错误
        std::vector <Token> tokens;
```

`Lexer`类中的**方法**定义为：

- `initialize_reverse_word_map()`：建立保留字表，用于后续查找
- 构造方法/析构方法
- `next()`：按照读入的字符`ch=source[pos]`,从字符串中解析token，存入其自身属性tokens
  - `intcon()`：解析`INTCON`种类token的私有方法
  - `idenfr()`：解析`IDENFR`种类token的私有方法
  - `skip_single_line_comment()`：处理单行注释的私有方法
  - `skip_multi_line_comment()`：处理多行注释的私有方法(状态机设计)
  - `chrcon()`：解析`CHRCON`种类token的私有方法(尤其要注意转义字符的处理`\`)
- `run()`：按格式输出`tokens/errors`数组到文件中

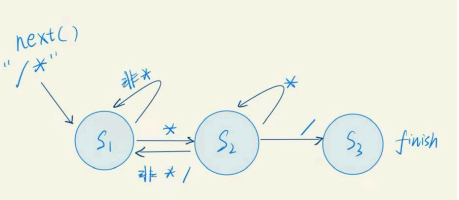
```
    public:
        Lexer(std::string source);
        ~Lexer();
        void next();
        void run();

    private:
            void intcon();
            void idenfr();
            void strcon();
            void chrcon();
            void skip_single_line_comment();
            void skip_multi_line_comment();
            void initialize_reverse_word_map();
```

处理多行注释的状态机设计：借鉴了课程组提供的词法分析`ppt`中的思路，但由于笔者的设计中只有连续读到`/,*`符号才会进入多行注释处理程序，因此只需要一个三状态状态机。

```
// multi_line_comment_fsm.h
enum State {
    S1,S2,S3
};
```

状态转移图如下图：



词法分析阶段的错误处理：在词法分析阶段只会有a类错误：将`&& / ||`记为`& / |`，为了可扩展性考虑，笔者建立了`Error`类来管理错误，其属性为：行号(`line_number`)，错误类型(error_type)，并定义相应的to_string方法用于输出Error类实例对象的字符串格式。

3.2 编码后修改

在完成语法分析部分时，词法分析部分读到错误的token时，不仅要报告错误，**还要返回正确的token类型**，需要对原设计进行修改。

四、语法分析部分设计

语法分析部分的任务主要为在读取Token流的同时建立抽象语法树，我的设计主要分为两部分：一是抽象语法树节点的设计，二是语法成分分析程序的设计(包括处理多产生式、处理左递归产生式)。

4.1 抽象语法树(AST)

> 该部分主要在ast.h/ast.cpp中实现

抽象语法树的根节点为编译程序单元(CompUnit)，各树节点为非终结符，叶子节点为终结符语法成分。对于各个节点，我们可以建立一个基类`Node`，其中包含各种树节点的一个公共属性所属行(`line_number`)，一个公共打印方法(`print`)。

```
struct Node {
    int line_number;

    Node() = default;
    Node(int line_number) : line_number(line_number) {}
    virtual void print(std::ostream &os) = 0;
};
```

而后各种语法成分继承基类，重载`print`方法，并对应各自的构造函数，例如：

```
struct CompUnit : public Node {
    std::vector<std::unique_ptr<FuncDef>> func_defs;
    std::vector<std::unique_ptr<Decl>> decls;
    std::unique_ptr<MainFunc> main_func;

    CompUnit(std::vector<std::unique_ptr<FuncDef>> func_defs,
std::vector<std::unique_ptr<Decl>> decls, std::unique_ptr<MainFunc>
main_func);
    void print(std::ostream &os) override;
};
```

同时，对于一种非终结符所管理的其他语法成分，我们使用C++11标准中提供的特性unique_ptr智能指针，该指针保证一个指针只能指向内存中一个内存实体，或者说所有权只能归一个指针所属，通过对这块对象的所有权传递(`std::move()`)来保证在函数传参等过程中不会发生参数的复制，减少内存开销。

对于多产生式的推导规则，我采用了两种方法，对于推导规则中不包含递归式的，例如`Stmt`：

```
using Stmt = std::variant<AssignStmt, ExpStmt, BlockStmt, IfStmt, ForStmt,
BreakStmt,
                          ContinueStmt, ReturnStmt, GetIntStmt, GetCharStmt,
PrintfStmt>;
```

对于这种情况，我们可以使用C++17标准中提供的`std::variant`来实现，该类型保证了实际存储类型中只能是声明中的任一种类型，并通过`get_if<T>`方法来返回指定T类型的指针(如果存储的是T类型，则返回T*类型，否则返回nullptr)，或者通过`std::visit()`方法访问其中的内容(`visit`内存需要传入一个匿名函数)。

对于推导规则中包含递归式的，例如算数表达式的推导，我采用结构体来实现，并在语法分析阶段递归创建结构体。

```
struct MulExp : public Node {
    //  MulExp → UnaryExp | MulExp ('*' | '/' | '%') UnaryExp
    std::unique_ptr<MulExp> mulexp; //!需要使用指针来避免无限递归
    std::unique_ptr<Token> op;
    std::unique_ptr<UnaryExp> unaryexp;

    void print(std::ostream &os) override;
    MulExp(std::unique_ptr<MulExp> mulexp, std::unique_ptr<Token> op,
std::unique_ptr<UnaryExp> unaryexp);
    MulExp(std::unique_ptr<UnaryExp> unaryexp);
};
```

4.2 语法分析(Parser)

对于语法分析程序，主要需要处理的点为多产生式的处理和左递归产生式的处理。

4.2.1 多产生式的处理

对于一些多产生式规则，各个产生式的FIRST集交集为空，这样只需要当前读到的Token就可以知道属于哪一种产生式，而无需进行预读，例如`InitVal`,`ConstInitVal`，对于各个产生式的FIRST交集不为空的情况，可以采取预读的策略，在判断结束后再回退读到的token，我在parser中实现了该功能。

```
        std::deque<Token> buffer; // 正常的token缓存区
        std::deque<Token> backbuf; // 保存预读的token
        std::deque<Token> recoverybuf; // 保存stmt处理中读LVal的token 这部分实际上
是可能解析错误的, 见parser.cpp中的注释
        bool is_recovering = false; // 是否正在恢复
```

其中buffer为正常读入的token缓冲区，backbuf中保存预读的token，而后再从backbuf中读取(这一规则在后续详细介绍)。

而对于最复杂的一种，例如`Stmt`中多产生式的处理，

```
语句 Stmt → LVal '=' Exp ';' // 每种类型的语句都要覆盖
| [Exp] ';' //有无Exp两种情况
| Block
| 'if' '(' Cond ')' Stmt [ 'else' Stmt ] // 1.有else 2.无else
| 'for' '(' [ForStmt] ';' [Cond] ';' [ForStmt] ')' Stmt // 1. 无缺省，1种情况
2.
ForStmt与Cond中缺省一个，3种情况 3. ForStmt与Cond中缺省两个，3种情况 4. ForStmt与
Cond全部
缺省，1种情况
| 'break' ';' | 'continue' ';'
| 'return' [Exp] ';' // 1.有Exp 2.无Exp
| LVal '=' 'getint''('')'';'
| LVal '=' 'getchar''('')'';'
| 'printf''('StringConst {','Exp}')'';' // 1.有Exp 2.无Exp
```

在首先利用FIRST集不相交的部分判断一些规则之后，以下几条规则中FIRST集有相交

- LVal '=' Exp ';'
- [Exp] ';'（**其中规则二排除了FIRST不相交的情况后起始成分最终推导也为LVal**）
- LVal '=' 'getint''('')'';'
- LVal '=' 'getchar''('')'';'

可以按照先解析一个LVal成分(一种步伐比较大的预读)，而后继续向下，这样可以判断出他们的分别，但是问题在于，这种解析方法对于规则2中的Exp并不适用(即使可以进行parse,但并不满足语法推导，得到的LVal不是合法的语法成分)，也就是说我们需要设计一种“回退”机制，可以回退解析LVal过程中读入的Token，对于Exp这种情况使用Exp的解析程序重新解析，得到符合语法推导的语法成分，本文设计了一种**恢复模式**，在后文中介绍。

对于`Stmt`的解析逻辑如下：

```
std::unique_ptr<Stmt> Parser::parse_stmt() {
    if (get_curtoken().get_type() == Token::LBRACE) { // Block
        auto block = parse_block();
        return std::make_unique<Stmt>
(std::in_place_type<BlockStmt>,BlockStmt(std::move(block)));
    } else if (get_curtoken().get_type() == Token::IFTK) { //IF
        auto if_stmt = parse_ifstmt();
        return std::make_unique<Stmt>
(std::in_place_type<IfStmt>,std::move(*if_stmt));
    } else if (get_curtoken().get_type() == Token::FORTK) { //FOR
        auto for_stmt = parse_forstmt();
        return std::make_unique<Stmt>
(std::in_place_type<ForStmt>,std::move(*for_stmt));
    } else if (get_curtoken().get_type() == Token::BREAKTK) { //BREAK
        auto break_stmt = parse_breakstmt();
        return std::make_unique<Stmt>
(std::in_place_type<BreakStmt>,std::move(*break_stmt));
    } else if (get_curtoken().get_type() == Token::CONTINUETK) { //
CONTINUE
        auto continue_stmt = parse_continuestmt();
        return std::make_unique<Stmt>
(std::in_place_type<ContinueStmt>,std::move(*continue_stmt));
    } else if (get_curtoken().get_type() == Token::RETURNTK) { // RETURN
        auto return_stmt = parse_returnstmt();
```

```
        return std::make_unique<Stmt>
(std::in_place_type<ReturnStmt>,std::move(*return_stmt));
    } else if (get_curtoken().get_type() == Token::PRINTFTK) { // PRINTF
        auto printf_stmt = parse_printfstmt();
        return std::make_unique<Stmt>
(std::in_place_type<PrintfStmt>,std::move(*printf_stmt));
    } else if (get_curtoken().get_type() == Token::LPARENT ||
               get_curtoken().get_type() == Token::PLUS ||
               get_curtoken().get_type() == Token::MINU ||
               get_curtoken().get_type() == Token::NOT ||
               get_curtoken().get_type() == Token::INTCON ||
               get_curtoken().get_type() == Token::CHRCON) { // EXP 注意这里
不能有IDENFR
        auto exp = parse_exp();
        if (get_curtoken().get_type() == Token::SEMICN) {
            next_token();
        } else { // i error
            unget_token();
            report_error(get_curtoken().get_line_number(), 'i');
            next_token();
        }
        return std::make_unique<Stmt>
(std::in_place_type<ExpStmt>,ExpStmt(std::move(exp)));
    } else if (get_curtoken().get_type() == Token::SEMICN) { // [EXP];(无
exp)
        next_token();
        return std::make_unique<Stmt>
(std::in_place_type<ExpStmt>,ExpStmt(nullptr));
    } else { // IDENTFR : may be rule 1,2,8,9
        Token t1 = get_curtoken();
        next_token();
        Token t2 = get_curtoken();
        unget_token();
        if (t1.get_type() == Token::IDENFR && t2.get_type() ==
Token::LPARENT) { // rule2 ident()
            auto exp = parse_exp();
            if (get_curtoken().get_type() == Token::SEMICN) {
                next_token(); // 读到LVal的起始token
            } else { // i error
                unget_token();
                report_error(get_curtoken().get_line_number(), 'i');
                next_token();
            }
            return std::make_unique<Stmt>
(std::in_place_type<ExpStmt>,ExpStmt(std::move(exp)));
        } else { // 1,2,9,10
            // 2余下的情况中一定以LVal开头, 1,9,10一定以LVal开头
            start_recovery(); // 将LVal读到recoverybuf中
            auto lval = parse_lval();
            if (get_curtoken().get_type() == Token::ASSIGN) { // rule
1,9,10
                abort_recovery(); // 不再恢复
                next_token(); // 跳过=
                if (get_curtoken().get_type() == Token::GETINTTK ||
                    get_curtoken().get_type() == Token::GETCHARTK ) { //
rule 9,10
```

```
65                      bool getint_flag = (get_curtoken().get_type() ==
Token::GETINTTK);
66                      next_token(); // 跳过getint/getchar
67                      next_token(); // 跳过(
68                      if (get_curtoken().get_type() == Token::RPARENT) {
69                          next_token();
70                      } else { // j error
71                          unget_token();
72                          report_error(get_curtoken().get_line_number(),
'j');
73                          next_token();
74                      }
75                      if (get_curtoken().get_type() == Token::SEMICN) {
76                          next_token();
77                      } else { // i error
78                          unget_token();
79                          report_error(get_curtoken().get_line_number(),
'i');
80                          next_token();
81                      }
82                      if (getint_flag) {
83                          return std::make_unique<Stmt>
(std::in_place_type<GetIntStmt>,GetIntStmt(std::move(lval)));
84                      } else {
85                          return std::make_unique<Stmt>
(std::in_place_type<GetCharStmt>,GetCharStmt(std::move(lval)));
86                      }
87                  } else { // rule 1
88                      auto exp = parse_exp();
89                      if (get_curtoken().get_type() == Token::SEMICN) {
90                          next_token();
91                      } else { // i error
92                          unget_token();
93                          report_error(get_curtoken().get_line_number(),
'i');
94                          next_token();
95                      }
96                      return std::make_unique<Stmt>
(std::in_place_type<AssignStmt>,AssignStmt(std::move(lval),
std::move(exp)));
97                  }
98              } else { // rule 2
99                  done_recovery(); // token恢复到backbuf中重新读取
100                 auto exp = parse_exp();
101                 if (get_curtoken().get_type() == Token::SEMICN) {
102                     next_token();
103                 } else { // i error
104                     unget_token();
105                     report_error(get_curtoken().get_line_number(), 'i');
106                     next_token();
107                 }
108                 return std::make_unique<Stmt>
(std::in_place_type<ExpStmt>,ExpStmt(std::move(exp)));
109             }
110         }
111     }
```

```
}
```

4.2.2 左递归产生式的处理

对于算数表达式中存在的左递归文法，我的做法是首先改写为消除左递归的形式，例如

```
MulExp → UnaryExp | MulExp ('*' | '/' | '%') UnaryExp;
// 改写为
MulExp → UnaryExp {('*'|'/'|'%') UnaryExp};
```

在解析过程中，首先解析一个UnaryExp，并构造为MulExp，而后根据后续是否读到运算符来递归构造MulExp,来满足文法中的要求。

```
std::unique_ptr<MulExp> Parser::parse_mulexp() {
    std::cout << "MulExp : cur token is " << get_curtoken().get_token() << std::endl;
    std::unique_ptr<MulExp> mul_exp = std::make_unique<MulExp>(parse_unaryexp());
    while (get_curtoken().get_type() == Token::MULT ||
            get_curtoken().get_type() == Token::DIV ||
            get_curtoken().get_type() == Token::MOD) {
        Token op = get_curtoken();
        next_token();
        auto unaryexp = parse_unaryexp();
        mul_exp = std::make_unique<MulExp>(std::move(mul_exp), std::make_unique<Token>(op),std::move(unaryexp));
    }
    return std::move(mul_exp);
}
```

4.2.3 回退模式与恢复模式

我设计的回退模式实际上都是回退模式与恢复模式实际上都是为多产生式的预读动作服务的，当我们一次只预读一个token的时候，可以直接使用恢复模式来解决，**但当我们遇到例如`Stmt`中去除不相交FIRST集后仍然起始同为LVal的四条规则，我们预读的token数量比较多，需要预读一个LVal才能判别，这时我设计了一个新的buffer用来存储解析LVal过程中保存的token,用来回退。**

```
LVal → Ident ['[' Exp ']'] //1.普通变量、常量 2.一维数组
```

buffer定义如下：

```
std::deque<Token> buffer; // 正常的token缓存区
std::deque<Token> backbuf; // 保存预读的token
std::deque<Token> recoverybuf; // 保存stmt处理中读LVal的token 这部分实际上是可能解析错误的, 见parser.cpp中的注释
bool is_recovering = false; // 是否正在恢复
```

buffer是正常读取的token缓存区，backbuf中存储预读之后回退的token，recoverybuf在恢复模式启用时启动，读到的token将保存到recoverybuf中，以下为关于三种buffer协作的方法：

```
void Parser::next_token() {
```

```
    auto& buf = is_recovering ? recoverybuf : buffer; //! 注意这里要引用 否则会创建副本
    if (!backbuf.empty()) { // 将预读的token从backbuf移回buffer
        buf.push_back(backbuf.back());
        backbuf.pop_back();
    } else if (lexer.has_next()) {
        buf.push_back(lexer.next());
    } else {
        std::cout << "Error! No Lexer Read!" << std::endl;
    }

    // backbuf是不断消耗的 不需要关心内存
    // buffer需要不断从头部移除旧元素来控制内存
    if (buffer.size() > 5) {
        buffer.pop_front();
    }
}

Token Parser::get_curtoken() {
    return (is_recovering ? recoverybuf : buffer).back();
}

void Parser::unget_token() {
    auto& buf = is_recovering ? recoverybuf : buffer;
    std::cout << "BackBuf in unget_token : " << buf.back().to_string() << std::endl;
    backbuf.push_back(buf.back());
    buf.pop_back();
}

void Parser::start_recovery() {
    is_recovering = true;
    if (!buffer.empty()) { // 将buffer中的最新token复制一个到recoverybuf中(LVal的起始token)
        recoverybuf.push_back(buffer.back());
    }
}

void Parser::done_recovery() { // 将recoverybuf中的token恢复到backbuf中重新解析
    is_recovering = false;
    while (!recoverybuf.empty()) { // 逆序填入backbuf, buffer读取的是正序
        backbuf.push_back(recoverybuf.back());
        recoverybuf.pop_back();
    }
    if (!buffer.empty()) {
        backbuf.pop_back(); // 这个元素是start时从buffer复制到recoverybuf的
    }
}

void Parser::abort_recovery() { // 将recoverybuf中的token放到buffer中，相当于正常解析过了，不再重新解析(不再恢复)
    is_recovering = false;
    if (!buffer.empty()) {
        recoverybuf.pop_front(); // 这个元素是start时从buffer复制到recoverybuf的
    }
    while (!recoverybuf.empty()) { // 顺序填入buffer
```

```
        buffer.push_back(recoverybuf.front());
        recoverybuf.pop_front();
    }
    while (buffer.size() > 5) { // 控制buffer的大小
        buffer.pop_front();
    }
}
```

- `next_token`：获取下一个token,当前为恢复模式时，向recoverybuf中装填，否则向buffer中装填，**同时需要注意判断backbuf中是否为空，如果不为空，说明其中包含预读回退的token，需要优先装填backbuf中的token**(每次next时从buffer尾部装填，从头部移除元素控制buffer大小)。
- `get_curtoken`：获取当前token，依据是否处在恢复模式选择从哪个buffer中取(注意这里的取只是获取一份复制，而不是从buffer中删除)。
- `unget_token`：回退预读的token，用于回退模式，将预读的token装填入backbuf，用于后续重新读取。
- `start_recovery`：启动恢复模式，将标志位设为true，并将buffer中最新的token复制到recoverybuf。
- `done_recovery`：结束恢复模式，取消标志位，将recoverybuf中的token装填到backbuf中重新解析，这里适用于Stmt中Exp那一条规则。
- `abort_recovery`：中止恢复模式，取消标志位，将recoverybuf中的token装填入buffer中(这时相当于成功解析，不需要重新解析)，适用于Stmt中以LVal开头的那三条规则。

4.3 打印输出(Print)

打印输出部分即为树的后序遍历，同时注意输出非终结符语法成分，本文中实现为对Node基类中的print方法进行重写。

```
void CompUnit::print(std::ostream &os) { // CompUnit → {Decl} {FuncDef} MainFuncDef
    for (auto &decl : this->decls) {
        std::visit(
            [&os](auto &arg){
                arg.print(os);
            },
            (*decl) //!需要注意的是数组中保存的是unique_ptr类型的元素, visit不能访问unique_ptr, 必须直接拿到元素类型，所以解引用
        );
    }
    for (auto &func_def : this->func_defs) { // 对容器中元素的引用 避免unique_ptr拷贝
        func_def->print(os);
    }
    main_func->print(os);
    os << "<CompUnit>";
}
```