**小组实验作业：(三个人1组，代码如果有抄袭，分数全为0，如果任何贡献都没有，为0)**

**提交截止时间：夏学期6月6日22:00**

**用lex写出一个tiger语言或者类C或者类PASCAL某个语言的词法分析器，用YACC的分析方法完成语法分析，并生成语法树和中间代码（如果生成目标代码：可加分）。 (30分)**

一、实验目的：掌握词法分析、语法分析、语义分析和代码生成方法。

二、实验环境：

Windows系统环境或者Linux环境。

1、Linux环境下的编译和运行

（1）Linux 2.6以上版本

（2）GCC3.4以上版本

（3）Bison 2.2以上版本

（4）Flex 2.5.33以上版本

发行版可以采用Ubantu, Gentoo, Fedora Core等。

2、Windows环境下的编译和运行

（1）Visual Studio 6.0

（2）Masm 6.0以上版本

（3）ParseGenerator 4.0 （Lex和Yacc的集成开发包）

三、实验要求：

1、实验内容

编译器包含词法分析、语法分析、语义分析、代码生成、代码优化、运行环境等阶段和环节。报告中要包括词法分析部分的正规表达式描述、数据结构、原理以及实现，语法分析部分的文法描述、数据结构、原理以及实现，语义分析的方法描述，中间代码的格式、数据结构描述以及中间代码生成的实现及部分优化（目标代码的生成）。

要明确组员分工情况。

2、如果生成语法树，要求结果能用可视的方法表示出来（图或者采用数据结构中学过的方法）

3、提交压缩包内的文件要求，压缩文件名为组号，如group1.zip

包括文件：

原代码文件：包括 lex文件，yacc文件以及其他程序等等

可执行文件

实验报告：文件名为report\_组号

4、实验报告建议格式 （不限于该格式）

报告封面 （见后一页）

序言 （包括对整个编译器的描述，对所提交的各个文件的说明，组员的分工）

1. 词法分析（正规表达式，实现原理和方法）
2. 语法分析 （上下文无关文法，实现原理和方法）
3. 语义分析 （实现方法）
4. 优化考虑 （每个阶段的优化考虑）
5. 代码生成 （所有语句的代码生成的处理）
6. 测试案例 （每个语句成分的测试案例，至少两个复杂语句组合后的测试案例）



**本科实验报告**

|  |  |
| --- | --- |
| 课程名称： | 编译原理 |
| 姓 名： | 范源颢（3180103574）  宋天泽（3180105221）  钱隆 （3180106068） |
| 学 院： | 计算机科学与技术学院 |
| 系： | 计算机科学与技术 |
| 专 业： | 计算机科学与技术 |
| 学 号： | 3180103574、3180105221 3180106068 |
| 指导教师： | 冯雁 陈纯 |

目录

[序言 5](#_Toc73799378)

[0.1 功能描述 5](#_Toc73799379)

[0.2 文件说明 5](#_Toc73799380)

[0.3运行环境 6](#_Toc73799381)

[0.4 分工 6](#_Toc73799382)

[第一章 词法分析 6](#_Toc73799383)

[1.1 约定 6](#_Toc73799384)

[关键字 6](#_Toc73799385)

[运算符 7](#_Toc73799386)

[标识符和字面量 7](#_Toc73799387)

[注释 7](#_Toc73799388)

[1.2 正规表达式 7](#_Toc73799389)

[1.3 lex实现原理和方法 8](#_Toc73799390)

[第二章 语法分析 10](#_Toc73799391)

[2.1 约定 10](#_Toc73799392)

[2.2 上下文无关文法 10](#_Toc73799393)

[2.2.1面向过程 11](#_Toc73799394)

[2.2.2变量定义 12](#_Toc73799395)

[2.2.3 函数定义 12](#_Toc73799396)

[2.2.3 复合语句结构 13](#_Toc73799397)

[2.2.4 语句 13](#_Toc73799398)

[2.2.4 表达式 14](#_Toc73799399)

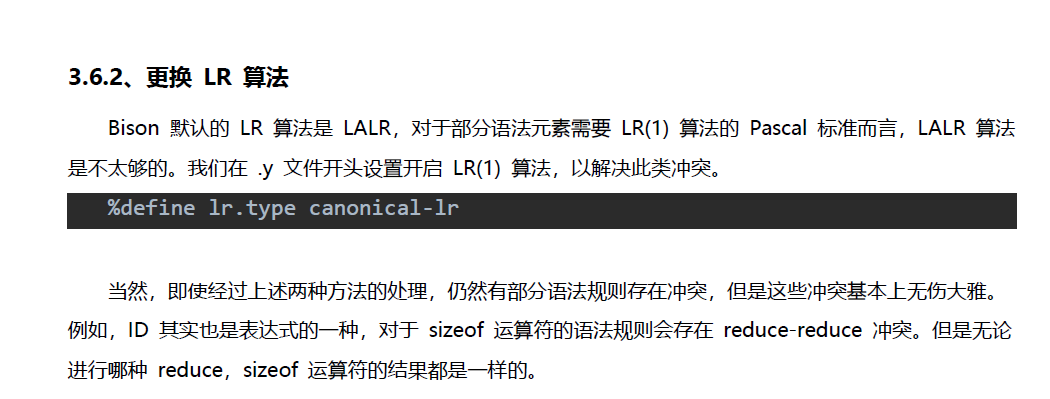
[2.2.5 函数调用 15](#_Toc73799400)

[2.3 语法树结点定义 15](#_Toc73799401)

[2.4 Yacc实现原理和方法 15](#_Toc73799402)

[2.5 测试 15](#_Toc73799403)

[简易测试和冲突处理 15](#_Toc73799404)

[ 16](#_Toc73799405)

[打印语法树函数的效果 16](#_Toc73799406)

[第三章 语义分析 16](#_Toc73799407)

[3.1 符号表设计 16](#_Toc73799408)

[数据结构设计 16](#_Toc73799409)

[符号表操作 16](#_Toc73799410)

[作用域 16](#_Toc73799411)

[3.2 语义分析（生成符号表和作用域） 16](#_Toc73799412)

[3.3 类型检查 17](#_Toc73799413)

[3.4 测试 17](#_Toc73799414)

[第四章 优化考虑 17](#_Toc73799415)

[4.1 词法设计优化 17](#_Toc73799416)

[4.2 语法设计优化 17](#_Toc73799417)

[4.3 llvm优化 17](#_Toc73799418)

[第五章 代码生成 17](#_Toc73799419)

[5.1 llvm模块和函数 18](#_Toc73799420)

[5.1.1 Module 18](#_Toc73799421)

[5.1.2 Function 18](#_Toc73799422)

[5.1.3 BasicBlock 18](#_Toc73799423)

[5.1.4 IRBuilder 18](#_Toc73799424)

[5.1.5 Value 18](#_Toc73799425)

[5.1.6 Name and Identity 18](#_Toc73799426)

[5.2 环境维护 19](#_Toc73799427)

[5.2.1代码生成的环境种类 19](#_Toc73799428)

[5.2.2标签信息的维护 19](#_Toc73799429)

[5.2.3 ID信息的维护 19](#_Toc73799430)

[5.2.4函数信息的维护 19](#_Toc73799431)

[5.2.5控制流信息的维护 19](#_Toc73799432)

[5.3 llvm 变量生成 19](#_Toc73799433)

[5.3.1 类型生成 19](#_Toc73799434)

[5.3.2 类型转换？ 19](#_Toc73799435)

[5.3.3常量信息 19](#_Toc73799436)

[5.4 llvm中间代码生成 19](#_Toc73799437)

[5.4.1表达式生成 19](#_Toc73799438)

[5.4.2语句生成 20](#_Toc73799439)

[5.4.3 函数生成 20](#_Toc73799440)

[5.4.4 效果测试： 20](#_Toc73799441)

[5.5 llvm 目标代码 20](#_Toc73799442)

[5.5.1 .asm结构 20](#_Toc73799443)

[5.5.2 .obj结构 21](#_Toc73799444)

[5.5.3 .out, Linux可执行文件 21](#_Toc73799445)

[第六章 测试案例 21](#_Toc73799446)

[6.1 构建工程 21](#_Toc73799447)

[6.2 编译器运行测试 21](#_Toc73799448)

[6.3 各语句测试 21](#_Toc73799449)

[6.4 复合语句测试 21](#_Toc73799450)

# 序言

## 功能描述

本报告是编译原理期末大作业小组实验的实验报告，此次编译原理期末大作业的目标是实现一个编译器，可使用的工具包括lex、Yacc、llvm等等，具体的实现形式可以是一个类C、类Pascal或其他程序语言的词法、语法分析器，需要生成语法树和中间代码，语法树需要可视化展示，但具体展示形式不限，中间代码的形式并无规定，从中间代码到目标机器的目标代码的转化为选做模块。

这次编译原理实验，我们完成了一个使用C++语言，利用lex、Yacc、llvm工具的类C（mytinyC）语言的编译器，这种类C语言在文法上基本参照C语言文法，在功能上去除了指针、结构体、联合体等C语言的特性，此外对于函数和语句的定义有比起C语言更加严格的限制，对于这种类C语言，我们设计的编译器可以有效的实现词法分析、语法分析、语义分析、代码生成以及代码优化，最终可以展示语法树、llvm中间代码和目标机器的目标代码以及可执行文件。

## 文件说明

ASTNodes.cpp 实验中编写，用于实现构建和展示语法树功能的C++语言代码文件；

ASTNodes.hpp 实验中编写，用于实现构建和展示语法树功能的C++语言头文件；

compile.sh 实验中编写，用于在Ubuntu 16.04+系统中运行完整运行编译器的批处理脚本文件；

lexer.l  实验中编写，用于对类C语言进行词法分析的lex文件

parser.y 实验中编写，用于对类C语言进行语法分析的yacc文件

lex.yy.c flex根据lexer.l生成的，可用于实现词法分析的C源文件

parser.tab.c bison根据parser.y生成的，可用于实现语法分析的C源文件

parser.tab.h bison根据parser.y生成的，可用于实现语法分析的C头文件

our\_syntax.md 实验中编写，用于提示我们设计的类C语言语法规则的Markdown文件，不参与编译器工程的构建。

c\_minus\_syntax.md 上交时删除

other file may also be included, and the file-organizing may be modified before handing-in.

## 0.3运行环境

使用cmake构建，依赖包括：

* Linux 16.04+
* gcc5.4.0+， g++ 5.4.0+
* flex 2.6.0+
* bison 2.6.0+, (推荐3.7.6+)
* llvm 3.8.0+
* clang 3.8.0+

(其中‘+’表示当前版本及以上)

## 0.4 分工

范源颢：词法分析 语法分析

宋天泽：语法分析（语法树构建） 语义分析

钱隆：中间代码和目标代生成

# 词法分析

## 1.1 约定

我们实现的类C语言mytinyC对于C语言的语法做了一些简化，因此我们有必要约定简化的内容，如下所述：

### 关键字

mytinyC关键字如下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| int | float | char | double | bool |
| void | if | else | for | while |
| return | extern |  |  |  |

其中bool用于声明mytinyC中的布尔类型的变量，bool类型仅包含true和false两种，分别可以被声明为true/True /TRUE和false/False/FALSE，运算规则和C++一致。其余的关键字和标准C语言中的关键字意义相同。

于是，我们的mytinyC支持的基本类型包括：整形int、浮点数float、字符char、布尔类型bool、双精度浮点数double。在函数的返回类型方面，我们支持虚返回void。我们的mytinyC语法支持的控制语句包括if-else分支语句， while/for循环语句，函数返回语句return。

注意比起标准C语言，我们在关键字便不再支持结构体（struct）、联合体（union）； unsigned、signed等复杂类型不再支持，我们不支持使用typedef自定义类型，也不支持使用const、static约束变量和函数的作用域。在流程控制方面，我们没有实现switch case 语法，也不支持使用break关键字跳出循环。在底层构建方面，我们也不支持使用sizeof语法查看数据类型的长度。

我们保留了extern关键字，这是为了在使用llvm的过程中，利用extern可以直接llvm中实现的函数，例如，假如在mytinyC文件中按照如下语句声明了输出函数printf：

extern int printf(char[] str)

我们就可以在测试在mytinyC文件中调用printf进行输出。

### 运算符

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| = | += | -= | \*= | /= |
| %= | |= | &= | ^= | >>= |
| <<= | || | && | | | & |
| ^(位抑或) | > | < | >= | <= |
| == | != | << | >> | + |
| - | \* | / | % | ! |
| ~（位取反） | ++（仅后缀） | --（仅后缀） | ( | ) |
| { | } | [ | ] | , |
| ; | ‘’ |  |  |  |

基本上涵盖了所有的C语言操作符。

mytinyC约定++和—只有作为后缀表达式出现，同时，我们没有指针运算，因此我们将\*仅理解为乘法，&理解为位与。

### 标识符和字面量

我们对于标识符的约定和C语言基本相同，用户自定义标识符需满足：字母（大小写均可）或者下划线开头，之后包含若干字母、数字和下划线的组合，推荐总长度在20字符以下。mytinyC不得将true/True/TRUE/false/False/False作为标识符。

我们对于字面量的约定和C语言也基本相同，开头或有字符-（负号），之后仅有数字组成的字符串我们认为是整形数字；开头或有负号，之后含有若干数字，在此之后包含小数点.，小数点之后若有字符则全为数字的字符串我们认为是浮点数（注意若小数点之后没有字符，也认为是浮点数，比如4.0和4.均为浮点数，而4为整形数）；‘’中包含的单个字符是C语言的char类型；三种不同大小写风格的true和三种不同大小写风格的false是布尔类型的字面量。

### 注释

我们实现了C语言风格的注释，包括//开头的单行注释，以及/\*和\*/之间的多行注释。

## 1.2 正规表达式

对于C语言关键字、操作符以及字面量false、true，我们都可以使用穷举的方式列出他们的正规表达式，例如，true的正规表达式可以写成：

true|True|TRUE

标识符和整形数、浮点数的正规表达式亦不难写出，我们给出整形、浮点型和标识符的正规表达式，如下：

int  [\-]?[0-9]+

float [\-]?[0-9]+\.[0-9]\*

identifier [a-zA-Z\_][a-zA-Z0-9\_]\*

注意，我们没有使用正规表达式的思路进行注释的词法分析，因为我们知道多行注释的正规表达式将比较复杂，为此，我们将使用下文中提到的lex的另一特性实现C语言的注释。

## 1.3 lex实现原理和方法

lex主要是通过有限状态机的思路实现词法分析的，我们不在赘述有限状态机的知识，而将注意力集中在如何通过书写lexer.l实现代码的分析。

在.l文件的第一部分（第一个%%之前）我们首先通过正规表达式定义一些语词：

 /\*空白符\*/

macro  #.\*

whitespace [ \t\r\n]+

linecomment \/\/.\*\n

 /\*字面量\*/

int  [\-]?[0-9]+

float [\-]?[0-9]+\.[0-9]\*

identifier [a-zA-Z\_][a-zA-Z0-9\_]\*

char  \'.\'

true  true|True|TRUE

false false|False|FALSE

之后，我们在第二部分定义lex在遇到这些语词的过程中需要执行的操作，若分析的结论是关键字和运算符，直接返回token即可：

"if"        return IF;

"else"      return ELSE;

"return"    return RETURN;

"for"       return FOR;

"while"     return WHILE;

"||"    return LOR;

"&&"    return LAND;

若是类型名称，我们需要保存内容，并且通过token表示其类型，这样将有助于日后的类型检查。

"int"  {

    strcpy(yylval.type\_id, yytext);

    return INT\_TYPE;

}

若是字面量，则我们需要保存字面量的值，并且通过token表示其类型。

{int} {

    yylval.type\_int = atoi(yytext);

    return INT;

}

{float} {

    yylval.type\_float = atof(yytext);

    return FLOAT;

}

{true} {

    yylval.type\_bool = 1;

    return TRUE;

}

{identifier} {

    strcpy(yylval.type\_id, yytext);

    return ID;

}

最后我们讲述注释的处理，如上所述，一旦匹配到//，则此行的内容均视为注释。（其后若出现/\*，标准C也不认为是多行注释的开始，下一行除非再有注释符号，否则即使上一行有未闭合的/\*，下一行的语句不会被编译器作为注释跳过）因此，我们可以直接用正则表达式匹配单行注释。

多行注释我们直接利用lex作为有限自动机的机制进行实现，lex在.l文件的第一部分支持直接定义状态，我们可以使用语句：

%x C\_COMMENT

定义一个名为C\_COMMENT的状态，

在.l文件的第二部分，我们定义一旦遇到字符/\*，使用lex定义的宏BEGIN进入C\_COMMENT状态：

"/\*" BEGIN(C\_COMMENT);

处于C\_COMMENT状态中时，无论遇到何种字符均进行空操作，

<C\_COMMENT>.  {}

注意lex使用尖括号标识遇到字符串时有限状态机的状态，如无标识，则推定此时是有限状态机处于lex内置的初始状态INITIAL中。

最后，如果处于C\_COMMENT状态中遇到\*/，则有限状态机应当返回初始状态INITIAL：

<C\_COMMENT>"\*/" BEGIN(INITIAL);

注意书写时，<C\_COMMENT>"\*/"语句应该先于<C\_COMMENT>.，因为lex优先匹配文件中先行的规则，如若带有.的规则先行，则之后的规则便失去意义，因为之后规则中的字符都会被.优先匹配掉。

# 语法分析

（上下文无关文法， 实现的原理和方法）

## 2.1 约定

我们在设计mytinyC的语法的时候亦有一些语法上的限制：

1. 我们在大括号中语句需要严格满足所有局部变量的定义先于其他语句；我们不支持for语句的小括号内第一个分句直接定义域内变量

2. 我们支持基本变量和一维数组在定义时初始化，但是不支持多维数组的初始化。

3. 同类型变量的声明支持使用逗号分隔，但所谓的“同类型”仅包括基本类型，不包括等复杂类型数组，例如： int a,b,c;视为合法，int a, b[10];视为不合法。

4. 和标准C一样，多维数组作为参数调用的时候，除标识符最右侧的方括号内可以为空之外，其余括号均需指明数值，例如一下的函数k,i,j中，仅有j是正确的定义。

void k(int a[5][]);

void i(int a[][]);

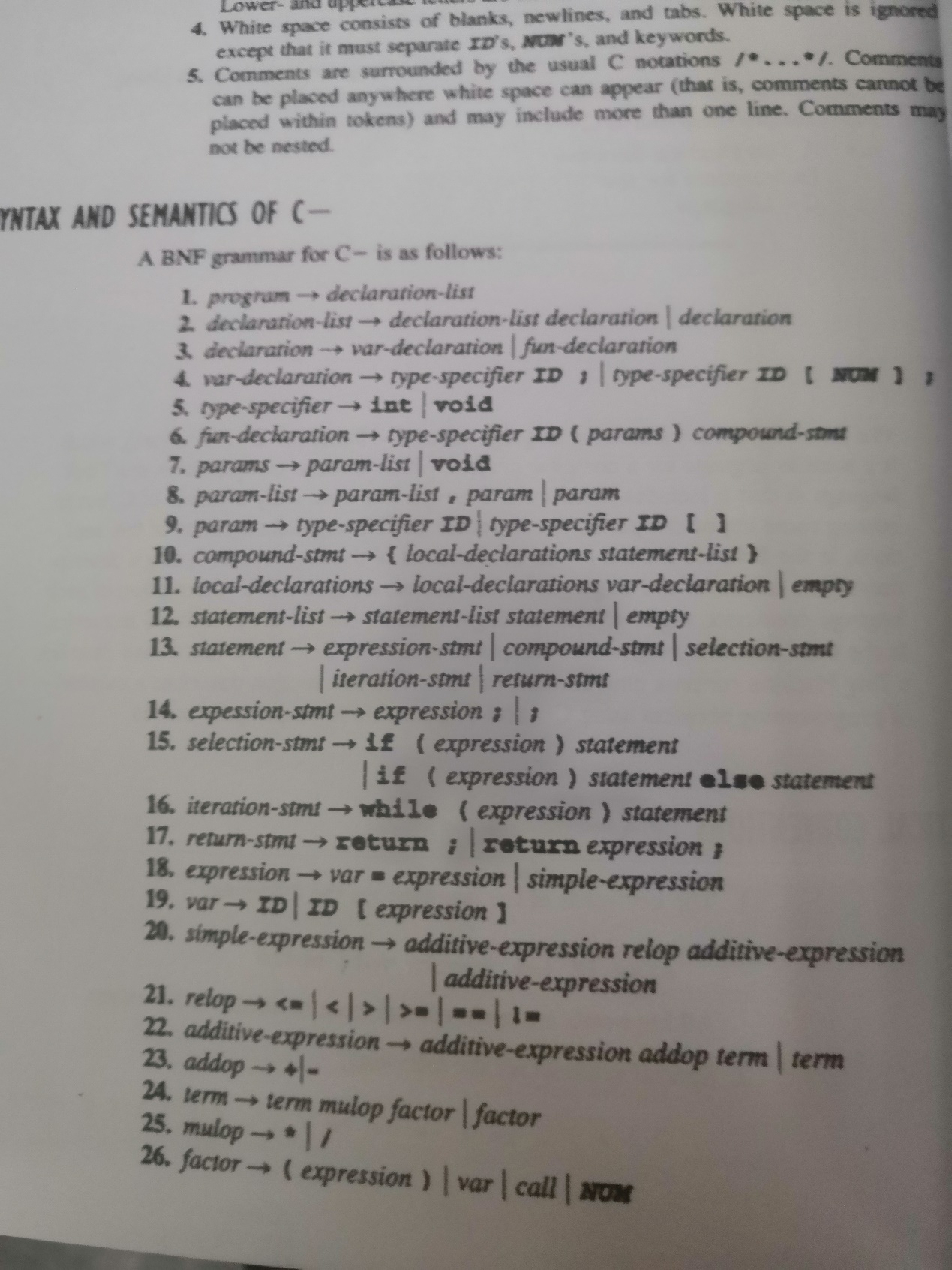
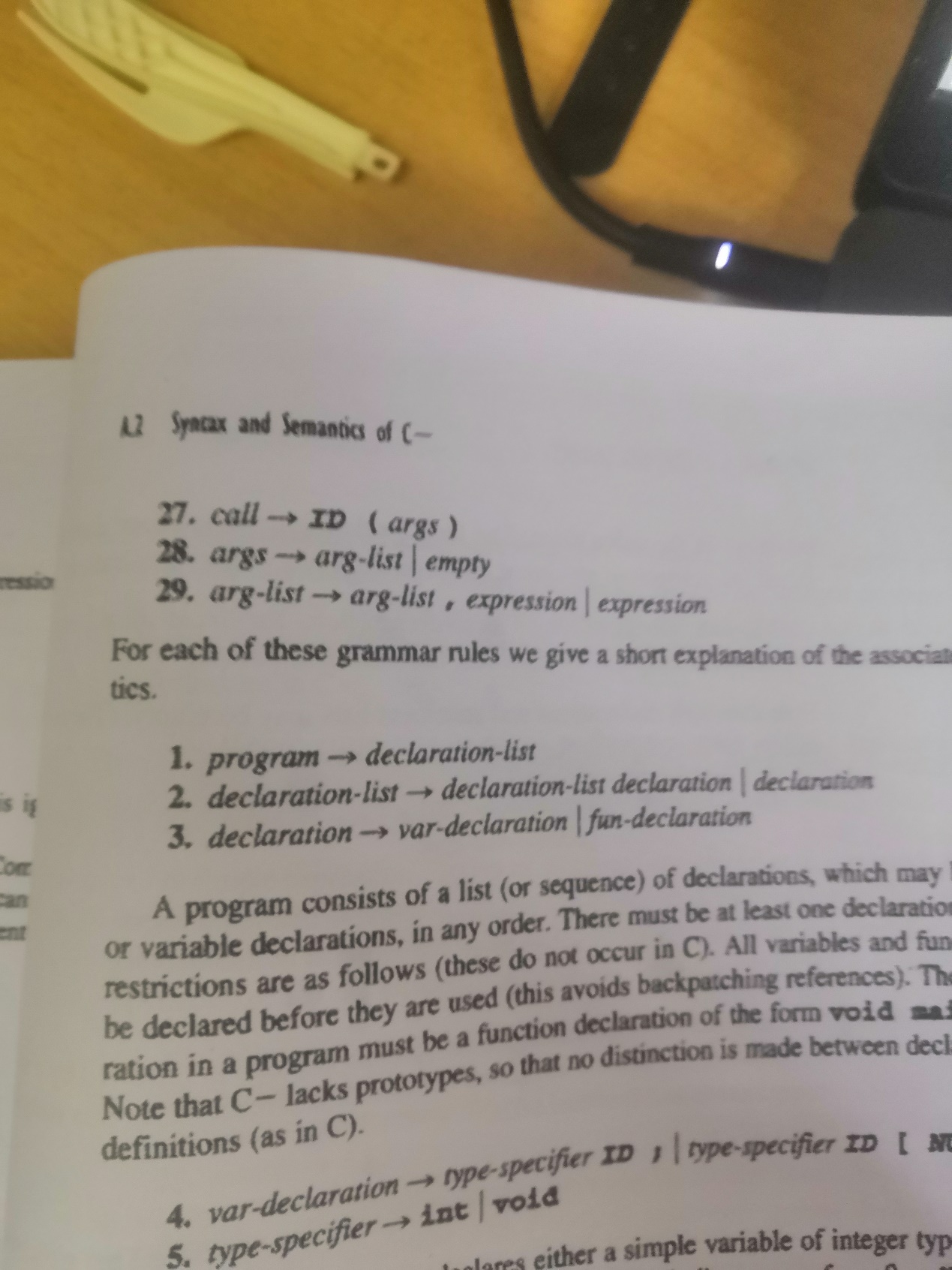
void j(int a[][5]);

5.待补充

## 2.2 上下文无关文法

我们使用的上下文无关文法主要基于课本附录中C-minus的文法，但是为了实现更多的功能，我们设法进行了一些优化。

课本的C-minus文法如下[[1]](#endnote-0)：

 图中斜体字表示非终继符，粗体正字表示终继符。

不难看出，这些语法规则可以作如下分类：

### 2.2.1面向过程

前三条规则反映了C语言面向过程的特性，我们把C语言的所有程序看做一系列的定义语句，包括函数定义（用户自定义的函数和main函数）和全局变量的定义。

1. program *→* declaration-list
2. declaration-list → declaration-list declaration | declaration
3. declaration → var-declaration | fun-declaration

### 2.2.2变量定义

第4第5条规则规定了变量的定义的语法：

1. var-declaration → type-specifier **ID ;** | type-specifier **ID [ NUM ] ;**
2. type-specifier → **int** | **void**

这两条规则实际上限定了C-只能定义整形和一维整形数组的变量，这不是我们想要的特性，我们对他们进行一系列扩充。

1. var-declaration → base-type id-list **;** | base-type **ID** array-post **;** | base-type **ID** array-post **=** **{** array-const-list **}** **;**
2. id-list → id-list **,** **ID** | id-list **,** **ID** **=** expression | **ID** | **ID** **=** expression
3. array-post → **[** **NUM** **]** | \*array-post **[** **NUM** **]**
4. array-const-list → array-const-list **,** single | single
5. base-type → **int** | **double** | **float** | **char** | **bool**

首先，通过引入了非终继符base-type，实现了整形、双精度浮点数、浮点数、字符和布尔类型的声明，大大提高了语言的灵活性。

其次，通过定义id-list 实现了基本类型的多变量定义，而且每个变量都可以单独初始化，语句int a,b,c;int a=2,b,c;int a=1,b=2,c=3;均合法，但是如前所述，不支持int a,b,c[20];因为我们不认为此时的c和a,b是同种类型

再次，为了实现多维数组的功能，我们需要非终继符array-post, 使其能够产生一对对括号。

最后，我们为了实现以为数组的初始化，设计了array-const-list，用于产生由逗号分割的数字列表，我们之所以不实现多维数组的初始化将在后文讲述。

### 2.2.3 函数定义

课本第6-9条规范了函数的定义：

1. fun-declaration → type-specifier **ID** **(** params **)** compound-stmt
2. params → param-list | **void**
3. param-list → param-list **,** param | param
4. param → type-specifier **ID** | type-specifier **ID** **[ ]**

通过params和param-list可以生成仅含有void，或者有包括一维数组在内的参数列表。

我们将其改进为：

1. fun-declaration → base-type **ID** **(** params **)** compound-stmt | **void** **ID** **(** params **)** compound-stmt | **extern** base-type **ID** **(** params **)** | **extern** **void** **ID** **(** params **)**
2. params → param-list | **void**
3. param-list → param-list **,** param | param
4. param → base-type **ID** | base-type **ID** array-post-param
5. array-post-param → **[** **]** | **[** **NUM** **]** | array-post-param **[** **NUM** **]**

这样我们在定义是可以指明外部函数，参数列表可以包含各个维度的数组，和标准C一样，除标识符最右侧的方括号内可以为空之外，其余括号均需指明数值。

### 2.2.3 复合语句结构

课本的语句实现为：

1. compount-stmt → **{** local-declarations statement-list **}**
2. local-declarations → local-declarations var-declaration | ***ϵ***
3. statement-list → statement-list statement | ***ϵ***

我们将其改进为：

1. *compound-stmt* → **{** *local-declarations statement-list* **}** | **{** *local-declarations* **}** | **{** *statement-list* **}** | **{** **}**
2. *local-declarations* → *local-declarations* *var-declaration* | *var-declaration*
3. *statement-list* → *statement-list* *statement* | *statement*

我们继承了定义语句严格先于其他语句这点，不过，比起课本的书写方式，我们有意回避在表达式中出现空白符，这和yacc的特性有关，我们将在下文详述，这里只需注意两套文化并在识别的语言上没有本质区别即可。

### 2.2.4 语句

课本的定义如下：

1. *statement* → *expression-stmt* | *compound-stmt* | *selection-stmt* | *iteration-stmt* | *return-stmt*
2. *expression-stmt* → *expression* **;**  | **;**
3. *selection-stmt* → **if** **(** *expression* **)** *statement* | **if** **(** *expression* **)** *statement* **else** *statement*
4. *iteration-stmt* →  **while** **(** *expression* **)** *statement*
5. *return-stmt* → **return ;**  | **return** *expression* **;**

这些语法指定了单独的语句可以成为的语句类型，比如赋值、分支、迭代和返回，但是注意这里将所有迭代语句都作为while语句，可以进行改进：

1. *statement-list* → *statement-list* *statement* | *statement*
2. *statement* → *expression-stmt* | *compound-stmt* | *selection-stmt* | *iteration-stmt* | *return-stmt*
3. *expression-stmt* → *expression* **;** | **;**
4. *selection-stmt* → **if** **(** *expression* **)** *statement* | **if** **(** *expression* **)** *statement* **else** *statement*
5. *iteration-stmt* → *while-stmt* | *for-stmt*
6. *while-stmt* → **while** **(** *expression* **)** *statement*
7. *for-stmt* → **for** **(** *expression* **;** *expression* **;** *expression* **)** *statement*
8. *return-stmt* → **return ;** | **return** *expression* **;**

### 2.2.4 表达式

课本的表达式定义：

1. *expression* → *var* **=** *expression* | *simple-expression*
2. *var* → **ID** | **ID** **[** *expression* **]**
3. *simple-expression* → *additive-expression* *relop* *additive-expression* | *additive-expression*
4. *relop* → **<=** | **<** | **>** | **>=** | **==** | **!=**
5. *additive-expression* → *additive-expression* *addop* *term* | *term*
6. *addop* → **+**|**-**
7. *term* → *term* *mulop* *factor* | *factor*
8. *mulop* → **\***| **/**
9. *factor* → **(** *expression* **)** | *var* | *call* | **NUM**

其对于运算符的支持比较少，我们需要支持更多的运算符，但是考虑到更多的运算符如果将优先级写入规则的话，语法的数量将会非常多，于是我们将优先级/结合性和语法分离，语法仅写作如下：

1. *expression* → *var* *assop* *expression* | *operand*
2. *assop* → **=** | **+=** | **-=** | **\*=** | **/=**| **%=**|**^=** | **&=** | **|=** | **<<=** | **>>=**
3. *var* → **ID** | **ID** *array-post*
4. *operand* → *operand* **||** *operand* | *operand* **&&** *operand* | *operand* **|** *operand* |*operand* **^** *operand* | *operand* **&** *operand* | *operand* **=** *operand* | *operand* **!=** *operand*| *operand* **<** *operand* | *operand* **>** *operand* | *operand* **<=** *operand* | *operand* **>=** *operand* | *operand* **<<** *operand* | *operand* **>>** *operand* | *operand* **+** *operand* | *operand* **-** *operand* | *operand* **\*** *operand* | *operand* **/** *operand* | *operand* **%** *operand* | *prefix* **(** *operand* **)** | **(** *operand* **)** | *prefix* *single* | *single*
5. *prefix* → **!** | **~** | **-**
6. *single* → *var* | *call* | **NUM**| **DOUBLE** | **CHAR** | **TURE** | **FALSE**

对于优先级的实现见下文。

### 2.2.5 函数调用

1. *call* → **ID** ( *args* )
2. *args* → *arg-list* | ***ϵ***
3. *arg-list* → *arg-list* **,** *expression* | *expression*

我们和课本采用完全一致的语法。

## 2.3 语法树结点定义

我们使用llvm

粘贴相关代码/.drawio贴图

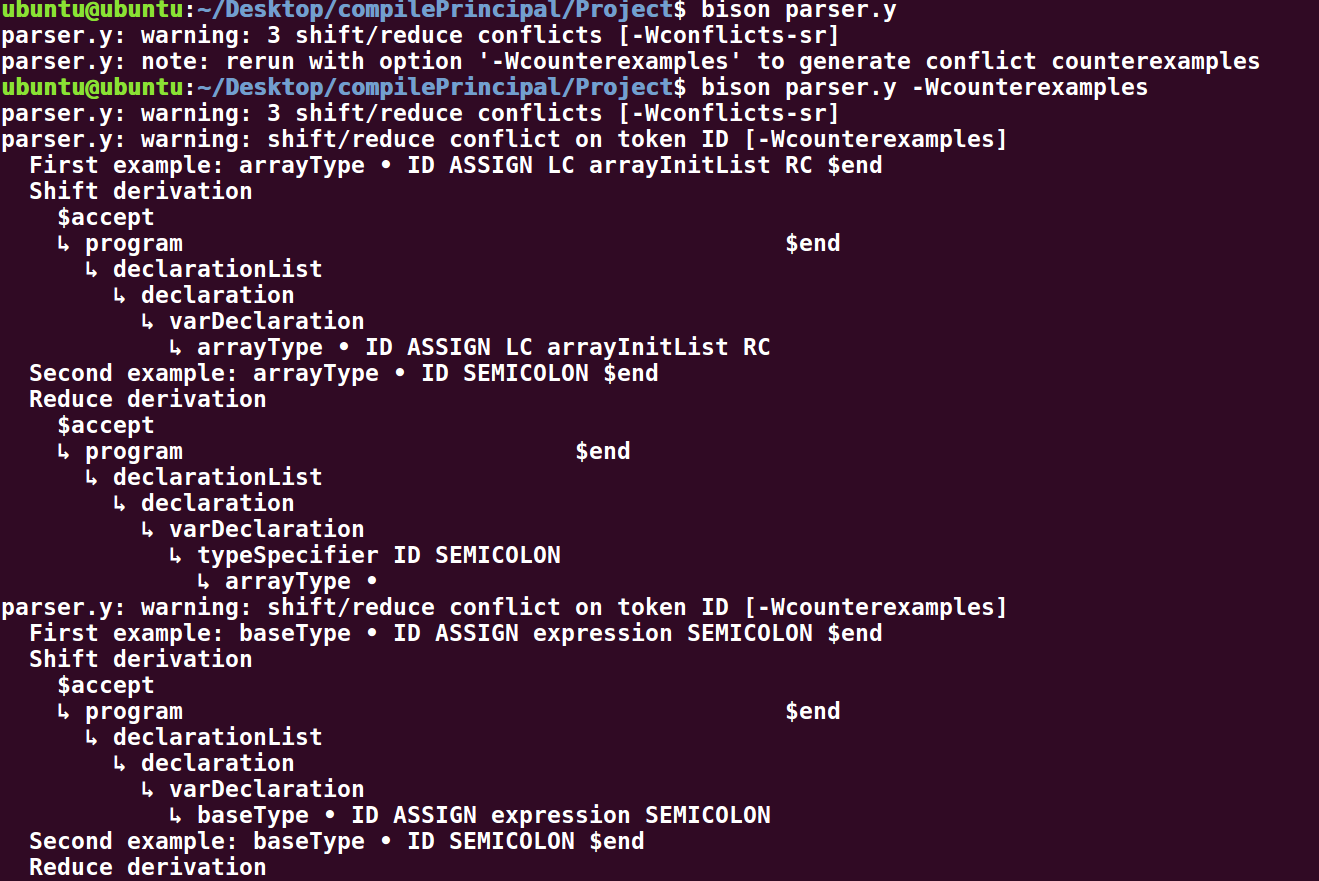
## 2.4 Yacc实现原理和方法

粘贴相关代码（设计例子即可）

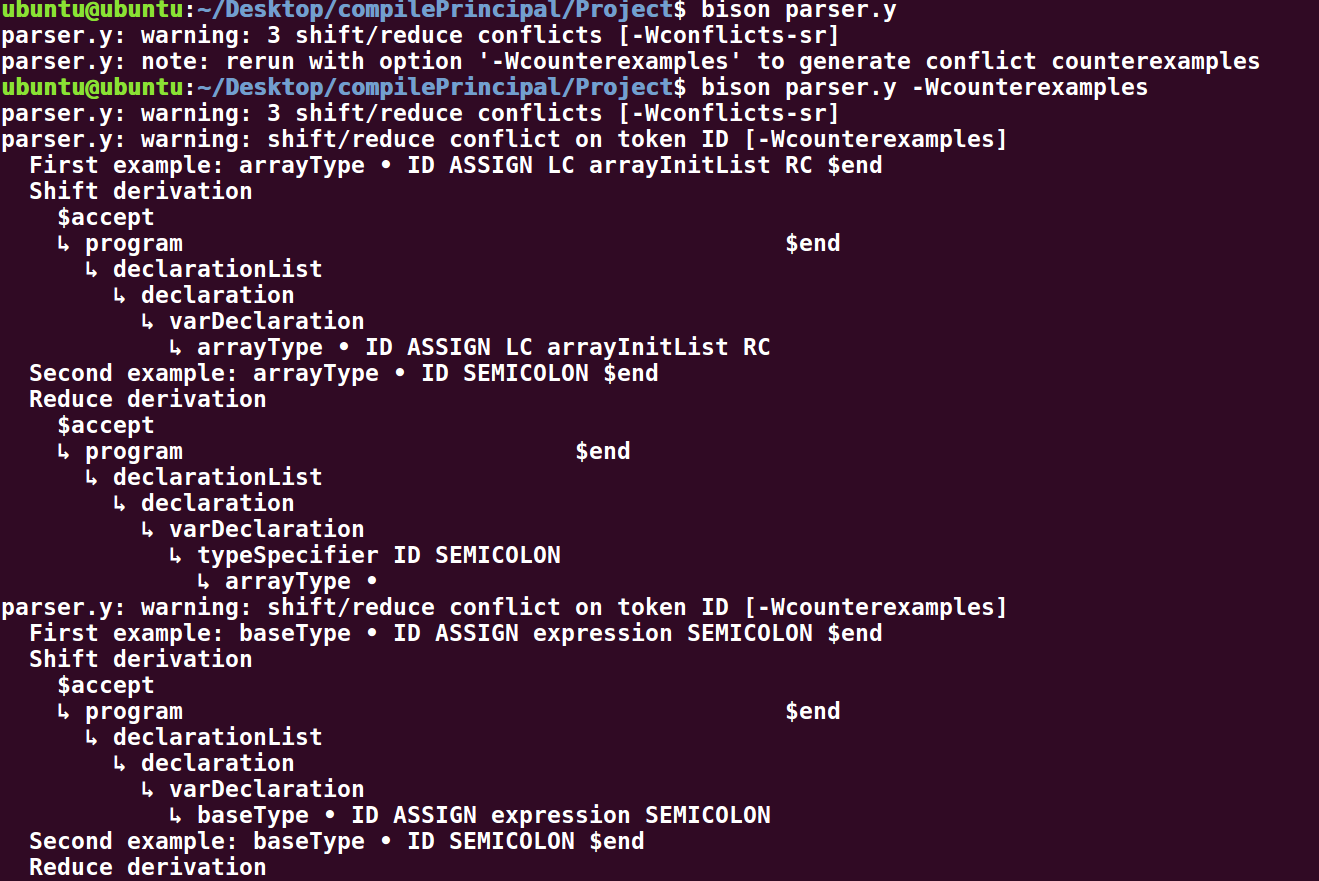
## 2.5 语法树的遍历和打印

## 2.6 测试

### 2.6.1简易测试和冲突处理

 我们在完成parser.y之后可以首先尝试使用bison编译出parser.tab.h和parser.tab.c文件，在修改掉所有语法错误之后，我们得到了几个warning，基本上都是shift/reduce conflicts[移进/规约冲突]，在较低版本的bison中，我们可能需要自己发现冲突所在的规则，但是在3.7.6及以上的版本中，bison增加了自动显示冲突例子的内容，而且会在提示用户存在移进/规约冲突之后使用相关的命令显示例。如下：

按照提示，我们使用命令-Wcounterexamples重新编译，可以得到具体的冲突实例：

我们注意，第一次出现语句warning：shift/reduce conflict on token ID [-Wcounterexample]是指，第一个可能出现冲突的实例出现在遇到标识符ID的情况下，其中Fisrst example[例一]和second example[例二]都显示了栈中和输入的情况，通过间隔号·分隔，左边为栈中，右边为等待输入，栈中只含有arrayType，假如读取的标记是ID，则会出现冲突，例一是指，首先移进ID ASSIGN LC arrayInitLit RC，之后在将这四者连同arrayType规约为varDeclaration的过程，是移进优先，而例二是指，首先将arrayType规约为typeSpecifier，之后再和移进ID SEMICOLON进行规约，是规约优先的选择，二者构成了冲突。

结合语义，我们成功通过修改规则消灭了除dangling-else之外的所有移进/规约问题，在语法设计移进没有大问题了。

### 2.6.2一个粗暴的手段：更换Yacc的匹配算法

Bison 默认的LR 算法是LALR，如果说发现设计的语法部分元素需要LR(1) 算法进行识别，那么LALR 算法是不太够的。在.y 文件开头设置开启LR(1) 算法，可以解决部分冲突。

%define lr.type canonical-lr

### 2.6.3 打印语法树的效果

粘贴结果

# 语义分析

（实现方法）

## 3.1 符号表设计

### 数据结构设计

我们采用了C++中自带的vector、map和pair三种数据结构来设计符号表。符号表的定义语句为：

static vector<map<string, pair<Value \*, vector<int> > > > varTable;

每次进入一个由花括号包裹的复合语句，程序都会向varTable中添加新的map成员，也即添加属于该作用域的符号表。如果所定义的变量为数组变量，则在map中插入键值对{变量名:<变量指针,数组形状>}，其中数组形状由vector<int>类型的对象来储存；而如果所定义的变量为非数组变量，则新建一个空vector填入与数组形状对应的参数。

### 符号表操作

进行中间代码生成时，每次读取到一个变量定义，就将其填入位于符号表vector尾部的map中，如果该变量是数组类型，则相应地填入该数组的形状。每次读取到一个变量名引用，就从符号表中查出其指针，而后通过llvm::CreateLoad方法返回其值。如果是数组类型的变量，则应额外根据下标以及其形状来计算出所需要读取的元素的地址，而后取值并进行引用。

### 作用域

每次进入一个复合语句，或者每次进入函数体，程序都会向符号表vector中插入一个空的map对象，该map对象即为该作用域下的符号表。随后，每次程序读取到一个变量定义，程序就会将该变量的变量名和相关信息记录到该map对象中。

## 3.2 语义分析（生成符号表和作用域）

似乎可以通过一个遍历方法把符号表得出

## 3.3 类型检查

计算属性。

## 3.4 测试

# 优化考虑

## 4.1 词法设计优化

如前所述，我们曾经设法将C语言的多行注释用正则表达式表达出来，但是其内容比较复杂，而且验证的复杂性比较高，因此，我们通过研究lex的语法设计了新的有限状态机状态<­C\_COMMENT>来实现对C语言注释的词法分析，大大降低了复杂度，同时也学习了lex通过新定义状态进行操作的功能。

## 4.2 语法设计优化

首先，bison默认使用LALR的方法解析语言，因此，我们可以在书写规则时偏好使用左递归，如课程所述，左递归可以避免长语句在分析时的爆栈。

其次，我们看之前提到的表达式的优先级和结合性的问题，如果将优先级写死在语法里面，那么我们的语法将会如下所述：

26. *simple-expression* → *simple-expression* *logop* *logic-expression* | *logic-expression*

1. *logop* → **||** | **&&**
2. *logic-expression* → *logic-expression* *bitop* *bit-expression* | *bit-expression*
3. *bitop* → **|** | **^** | **&**
4. *bit-expression* → *bit-expression* *relop* *shift-expression* | *shift-expression*
5. *relop* → **<=** | **<** | **>** | **>=** | **==** | **!=**
6. *shift-expression* → *shift-expression* *shiop* *additive-expression* | *additive-expression*
7. *shiop*→ **<<** | **>>**
8. *additive-expression* → *additive-expression* *addop* *term* | *term*
9. *addop* → **+**|**-**
10. *term* → *term* *mulop* *factor* | *factor*
11. *mulop* → **\***| **/** | **%**
12. *factor* → **!** *factor* | **~** *factor* | **-** *factor* | *incre*
13. *incre* → *incre***--** | *incre***++** | *terminal*
14. *terminal* → **(** *expression* **)** | *var* | *call* | **NUM**| **DOUBLE** | **CHAR** | **TURE** | **FALSE**

其内容将会非常复杂，我们可以直接将不含优先级的语法直接翻译成yacc的代码，通过排列或号之间各个子句的顺序，得到子句的优先级，这是yacc支持的，而结合性，则可以使用yacc内置的伪变量%left，%right指定。

%right ASSIGN PLUSASSIGN MINUSASSIGN MULTASSIGN DIVASSIGN MODASSIGN BORASSIGN BXORASSIGN BANDASSIGN SRASSIGN SLASSIGN

%left LOR

%left LAND

%left BOR

%left BXOR

%left BAND

%left EQ NEQ

%left LT GT LTE GTE

%left SL SR

%left PLUS MINUS

%left MULT DIV MOD

%right LNOT BNOT

......

%%

......

expression:         var assop expression { $$ = new ExpressionNode($2, $1, $3, nullptr); }

                    | operand { $$ = new ExpressionNode(nullptr, nullptr, nullptr, $1); }

                    ;

assop:              ASSIGN { $$ = new string("ASSIGN"); }

                    | PLUSASSIGN { $$ = new string("PLUSASSIGN"); }

                    | MINUSASSIGN { $$ = new string("MINUSASSIGN"); }

                    | MULTASSIGN { $$ = new string("MULTASSIGN"); }

                    | ......

                    ;

var:                ID { $$ = new VarNode(new string($1), nullptr); }

                    | ID arrayPost { $$ = new VarNode(new string($1), $2); }

                    ;

operand:            operand LOR operand { $$ = new OperandNode(new string("LOR"), $1, $3, nullptr); }

                    | operand LAND operand { $$ = new OperandNode(new string("LAND"), $1, $3, nullptr); }

                    | operand BOR operand { $$ = new OperandNode(new string("BOR"), $1, $3, nullptr); }

                    | operand BXOR operand { $$ = new OperandNode(new string("BXOR"), $1, $3, nullptr); }

                    | operand BAND operand { $$ = new OperandNode(new string("BAND"), $1, $3, nullptr); }

                    | operand EQ operand { $$ = new OperandNode(new string("EQ"), $1, $3, nullptr); }

                    |

 ......

再次，我们考虑尽可能减少语法中出现的empty语句，以免shift-reduce冲突，实践中，我们通过将含有空字符串的语法重新设计去除了很多移进/规约冲突的警告。

还有，孤悬-否则问题（dangling-else problem）如前所述，不需要我们解决，yacc在面临shift-reduce冲突的过程中，总会选择优先shift，这和我们设计if-else语句时，偏好将else和最靠近的无配对if配对是一致的，于是我们将允许这样的warning存在。

最后，我们解释一下为何我们不再支持使用上下文无关语法实现多维数组的初始化，因为我们发现，如若我们使用语句：

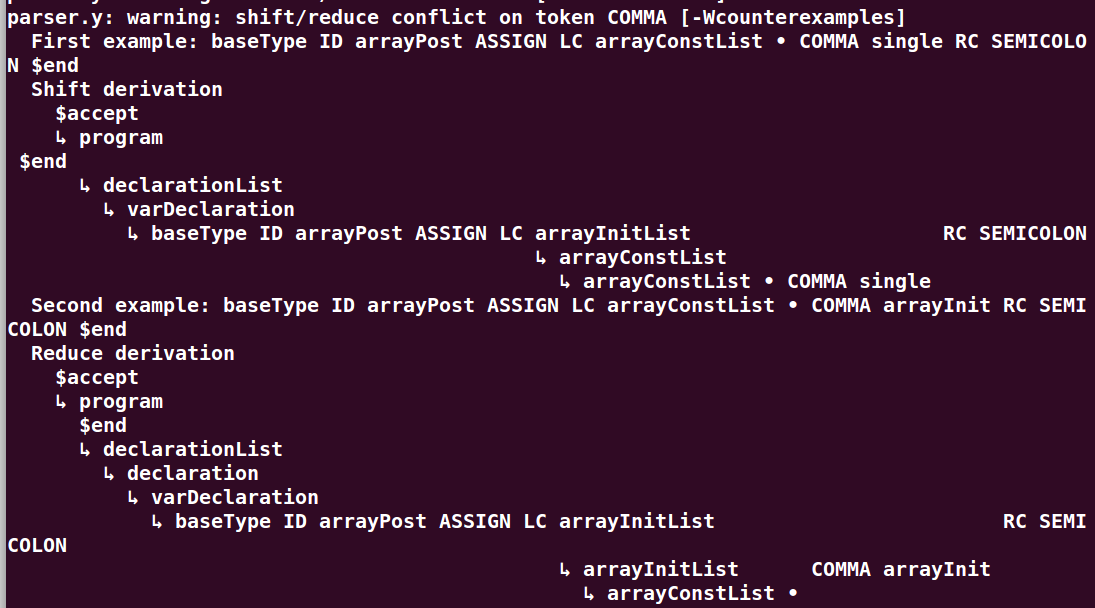
arrayInitList:      arrayInitList COMMA arrayInit {}

                    | arrayInit {}

| arrayConstList {}

;

arrayInit:          LC arrayInitList RC {};

 那么我们将会产生shift-reduce冲突，如下：

arrayConstList是一系列用逗号分割的初始值的列表，现在的问题是，当我们读入了这样一个列表之后，如果遇到逗号，我们无法分辨这个逗号是需要在这个初始值列表之后新加入一个初始值元素，还是已经表示逗号之后需要新加入一个包含大括号的，新的一列初始值列表，这是LR(1)无法解决的问题，可能需要2位前瞻才可以做到。也许重新设计语法会消除这个问题，但是我们最终因为时间有限放弃了多维数组的初始化。

## 4.3 llvm优化

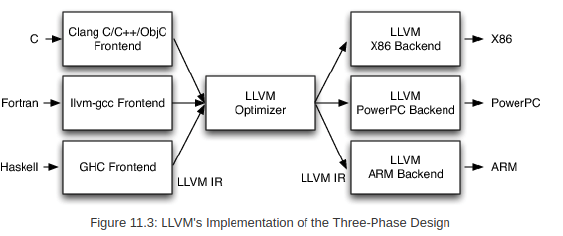
llvm自己的常量运算优化；

llvm中的控制流优化。

# 代码生成

（所有语句的代码生成的处理）

## 5.1 llvm模块和函数



调用<llvm/IR/……>中的文件内容，我们可以使用llvm现成的模块，有利于中间代码的生成。

### 5.1.1 Module

llvm::Module模块将会包括所有的函数和全局变量，而作为顶层结构，该模块也会拥有所有中间代码的信息。

### 5.1.2 Function

llvm::Function模块将会通过llvm::FunctionType类型的子对象来记录函数返回值类型、函数参数类型、链接方法等函数原型信息。而得到正确的llvm::FunctionType对象后，就可以通过Function::Create方法来创建函数，并将相关信息注册到llvm::Module模块中。

### 5.1.3 BasicBlock

BasicBlock 是代码块写入和管理的基本单位，诸如跳转、函数定义等功能均需要使用BasicBlock对象来进行分块与规划，而IRBuilder所写入的代码位置也常常需要通过特定的BasicBlock来指定。

### 5.1.4 IRBuilder

IRBuilder是代码生成和插入的基本管理模块，基本的数学运算、变量生成和流程控制对应的中间代码均需要通过IRBuilder的对应方法来进行生成。

### 5.1.5 Value

Value是基本的数据管理类，局部变量、全局变量、常量、运算结果、代码量等数据量均可以直接或者简介地转化成Value类型。同时，Value类型也包含有变量类型、变量名等大量信息。

## 5.2 环境维护

### 5.2.1 变量信息的维护

变量信息主要通过符号表来进行维护。本程序中，符号表用一个模板类型为map<string, Value \*>的vector对象来进行维护。每次进入函数体或者进入由花括号包围的复合语句时，程序都会将一个空的map对象添加进入vector，从而创建一个新的定义域。读取到变量定义时，程序会将其添加进vector对象当前的最后一个map中，也即添加到当前定义域的符号表中。读到变量引用时，程序会倒序读取vector中的map，一旦发现与当前引用变量同名的变量，就将其Value指针和维度信息返回，进而进行进一步调用。

### 5.2.2函数信息的维护

我们没有专门使用一个单独的符号表来维护函数信息，因为我们需要的函数信息都已在llvm::Module模块中得到储存。

## 5.3 llvm 变量生成

### 5.3.1 类型生成

### 5.3.2 类型转换？

### 5.3.3常量信息

## 5.4 llvm中间代码生成

### 5.4.1表达式生成

常规的和短路的

### 5.4.2语句生成

#### 5.4.2.1 变量声明语句

#### 5.4.2.2控制流语句（if/for/while）

### 5.4.3 函数生成

#### 5.4.3.1 用户定义函数

#### 5.4.3.2 系统函数

### 5.4.4 效果测试：

表达式+变量

控制流

函数定义和调用；

## 5.5 llvm 目标代码

### 5.5.1 .asm结构

命令

### 5.5.2 .obj结构

命令

### 5.5.3 .out, Linux可执行文件

命令

# 测试案例

（每个语句成分的测试案例， 至少两个复合语句组合后的测试案例）

## 6.1 构建工程

shell+makefile脚本

## 6.2 编译器运行测试

## 6.3 各语句测试

表达式+变量

控制流

函数定义和调用；

## 6.4 复合语句测试

表达式+变量

控制流

函数定义和调用；

22组合

1. compiler construction principles and practice P492 [↑](#endnote-ref-0)