Compiler Lab1 Lexer

201250104 苏致成

November 2022

1 概述

1.1 目标

使用 Antlr 对 SysY 语言生成词法分析器，对其进行解析。

1. 若各行均符合 SysY 语言规范，则输出标识的 token 。
2. 若有未识别的 token，表明此时存在错误，即不符合 SysY 语言的规范。

2 实现

2.1 使用工具

git、Antlr、Intellij idea、JDK11、Makefile

2.2 实现功能

向 main 方法中传递文件参数，程序可以对其进行词法分析，其余过程详见“目标”。

2.3 实现过程

1. 编写 g4 文件，需注意最上方需要 lexer grammar SysYLexer，以免生成 Parser；注意对于非解析元素加上 fragment 标记。
2. 利用 Antlr 生成对应的 SysYLexer.java 文件。
3. 在 Main 类中调用 SysLexer 进行 token 的识别和输出。
4. 编写 VerboseListener 并移除原有的 ErrorListeners 。
5. 在 main 方法中加入判断逻辑，即若已进入 ErrorListeners，则无需输出已分析的 token 。

3 bug

3.1 无法提交

3.1.1 问题描述

尝试提交压缩包，发现无法提交。即便手动提交也产生 400 的状态码。

3.1.2 解决方式

压缩包过大，助教将原先 1M 的上限更改为 10M。

3.2 冗余输出

3.2.1 问题描述

即便错误仍输出原 token。即原先认为捕获异常之后进行逻辑判断失效。

```
if(!listener.getEntered()){
    List<? extends Token> allTokens = sysYLexer.getAllTokens();
}
```

3.2.2 解决方式

意识到调用 getAllTokens 时 lexer 才开始进行识别工作，而并非在初始化的时候，因此上述方式中 getEntered() 返回值为 false，即仍会进入如下输出的逻辑中。需改成：

```
List<? extends Token> allTokens = sysYLexer.getAllTokens();
if(!listener.getEntered()){}
```

3.3 错误输出

3.3.1 问题描述

若遇到不可识别的 token 时，需要获取该 token 并输出，但面对如下函数签名时并不清楚如何获取 token。

```
public void syntaxError(Recognizer<?, ?> recognizer,Object
    offendingSymbol,int line, int charPositionInLine,String msg,
    RecognitionException e);
```

3.3.2 解决方式

1. 仔细阅读手册发现并不需要输出不可识别的 token 。
2. 可通过 offendingSymbol 获取，但尝试发现有误，后续将继续尝试。

3.4 make 导致删除文件

3.4.1 问题描述

若执行 make submit 时，若文件并未 add 或 commit，则会使得原有文件删除。

3.4.2 解决方式

在 make submit 的 shell 代码更改为 clean、compile。

3.5 进制转换

3.5.1 问题描述

十六进制和八进制进制转换。

3.5.2 解决方式

使用 parseInt() 函数进行进制转换。