MoonBit 月兔编程

现代编程思想

案例：语法解析器

Hongbo Zhang

语法解析器

MoonBit 月兔编程

- 案例目标
  - 解析基于自然数的数学表达式：`"(1+ 5) * 7 / 2"`
  - 转化为单词列表

    ```
    LParen Value(1) Plus Value(5) Multiply Value(7) Divide Value(2)
    ```

  - 转化为抽象语法树

    ```
    Division(Multiply(Add(Value(1), Value(5)), Value(7)), Value(2))
    ```

  - 计算最终结果：21
- 语法分析
  - 对输入文本进行分析并确定其语法结构
  - 通常包含**词法分析**和**语法分析**
  - 本节课均利用**语法解析器组合子**（parser combinator）为例

词法分析

MoonBit 月兔编程

- 将输入分割为单词
  - 输入：字符串/字节块
  - 输出：单词流
  - 例如：`"12 +678"` -> `[ Value(12), Plus, Value(678) ]`
- 通常可以通过有限状态自动机完成
  - 一般用领域特定语言定义后，由软件自动生成程序
- 算术表达式的词法定义

```
1. Number     = %x30 / (%x31-39) *(%x30-39)
2. LParen     = "("
3. RParen     = ")"
4. Plus       = "+"
5. Minus      = "-"
6. Multiply   = "*"
7. Divide     = "/"
8. Whitespace = " "
```

MoonBit 月兔编程

词法分析

- 算术表达式的词法定义

```
1. Number = %x30 / (%x31-39) *(%x30-39)
2. Plus   = "+"
```

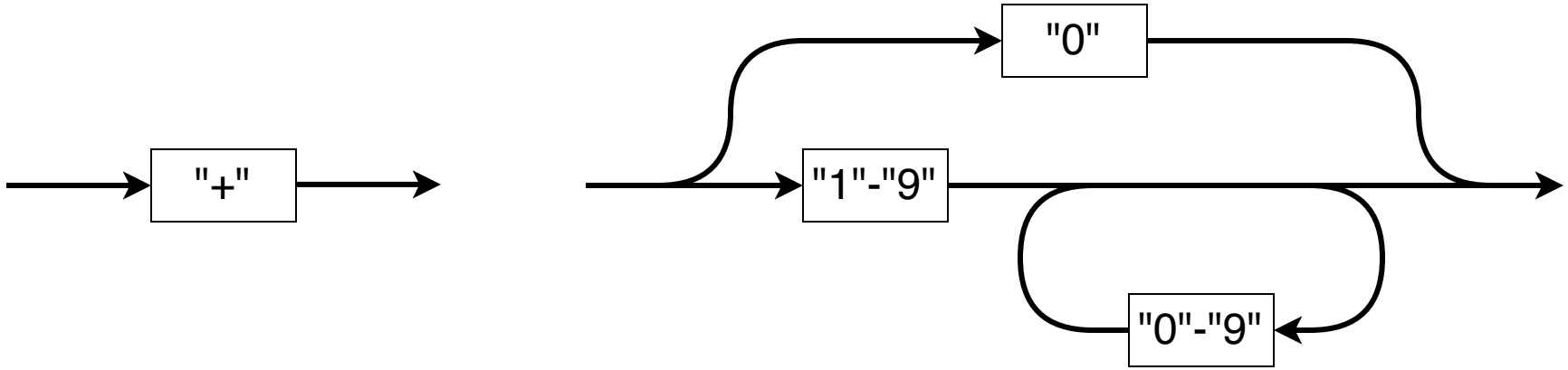
- 每一行对应一个匹配规则
  - `"xxx"`：匹配内容为xxx的字符串
  - `a b`：匹配规则a，成功后匹配规则b
  - `a / b`：匹配规则a，匹配失败则匹配规则b
  - `*a`：重复匹配规则a，可匹配0或多次
  - `%x30`：UTF编码十六进制表示为30的字符（`"0"`）

词法分析

MoonBit 月兔编程

- 算术表达式的词法定义

```
1. Number = %x30 / (%x31-39) *(%x30-39)
2. Plus   = "+"
```



- 单词定义

```
1. enum Token {
2.   Value(Int); LParen; RParen; Plus; Minus; Multiply; Divide
3. } derive(Debug)
```

MoonBit 月兔编程

解析器组合子

- 构造可组合的解析器

```
// V 代表解析成功后获得的值
// Lexer[V] == (String) -> Option[(V, String)]
type Lexer[V] (String) -> Option[(V, String)]

fn parse[V](self : Lexer[V], str : String) -> Option[(V, String)] {
  (self.0)(str)
}
```

  - 我们简化处理报错信息以及错误位置（可以使用 `Result[A, B]` ）

解析器组合子

MoonBit 月兔编程

- 最简单的解析器
  - 判断下一个待读取的字符是否符合条件，符合则读取并前进

```
fn pchar(predicate : (Char) -> Bool) -> Lexer[Char] {
  Lexer(fn(input) {
    if input.length() > 0 && predicate(input[0]) {
      Some(input[0], input.to_bytes().sub_string(1, input.length() - 1))
    } else {
      None
} }) }
```

- 例如

```
fn init {
  debug(pchar(fn{ ch => ch == 'a' }).parse("asdf")) // Some(('a', "sdf"))
  debug(pchar(fn{
    'a' => true
    _   => false
  }).parse("sdf")) // None
}
```

词法分析

- 单词定义：数字或左右括号或加减乘除

```
enum Token {
  Value(Int)
  LParen; RParen; Plus; Minus; Multiply; Divide
} derive(Debug)
```

- 分析运算符、括号、空白字符等

```
let symbol: Lexer[Char] = pchar(fn{
  '+' | '-' | '*' | '/' | '(' | ')' => true
  _ => false
})
let whitespace : Lexer[Char] = pchar(fn{ ch => ch == ' ' })
```

解析器组合子

MoonBit 月兔编程

- 如果解析成功，对解析结果进行转化

```
fn map[I, O](self : Lexer[I], f : (I) -> O) -> Lexer[O] {
  Lexer(fn(input) {
    // 非空为Some(v)中的v，空值直接返回
    let (value, rest) = self.parse(input)?
    Some(f(value), rest)
}) }
```

- 分析运算符、括号并映射为对应的枚举值

```
let symbol: Lexer[Token] = pchar(fn{
    '+' | '-' | '*' | '/' | '(' | ')' => true
    _ => false
}).map(fn{
    '+' => Plus;     '-' => Minus
    '*' => Multiply; '/' => Divide
    '(' => LParen;   ')' => RParen
})
```

解析器组合子

MoonBit 月兔编程

- 解析 `a`，如果成功再解析 `b`，并返回 `(a, b)`

```
fn and[V1, V2](self : Lexer[V1], parser2 : Lexer[V2]) -> Lexer[(V1, V2)] {
  Lexer(fn(input) {
    let (value, rest) = self.parse(input)?
    let (value2, rest2) = parser2.parse(rest)?
    Some((value, value2), rest2)
}) }
```

- 解析 `a`，如果失败则解析 `b`

```
fn or[Value](self : Lexer[Value], parser2 : Lexer[Value]) -> Lexer[Value] {
  Lexer(fn (input) {
    match self.parse(input) {
      None => parser2.parse(input)
      Some(_) as result => result
} }) }
```

解析器组合子

- 重复解析 `a`，零或多次，直到失败为止

```
fn many[Value](self: Lexer[Value]) -> Lexer[List[Value]] {
  Lexer(fn(input) {
      let mut rest = input
      let mut cumul = List::Nil
      while true {
        match self.parse(rest) {
          None => break
          Some(value, new_rest) => {
            rest = new_rest
            cumul = Cons(value, cumul) // 解析成功添加内容
      } } }
      Some(reverse_list(cumul), rest) // ⚠️List是栈，需要翻转以获得正确顺序
}) }
```

MoonBit 月兔编程

词法分析

- 整数分析

```
// 通过字符编码将字符转化为数字
let zero: Lexer[Int] =
  pchar(fn{ ch => ch == '0' }).map(fn{ _ => 0 })
let one_to_nine: Lexer[Int] =
  pchar(fn{ ch => ch.to_int() >= 0x31 && ch.to_int() <= 0x39 }).map(fn { ch => ch.to_int() - 0x30 })
let zero_to_nine: Lexer[Int] =
  pchar(fn{ ch => ch.to_int() >= 0x30 && ch.to_int() <= 0x39 }).map(fn { ch => ch.to_int() - 0x30 })

// number = %x30 / (%x31-39) *(%x30-39)
let value: Lexer[Token] =
  zero.or(
    one_to_nine.and(zero_to_nine.many()) // (Int, List[Int])
      .map(fn{ // 1 2 3 -> 1 * 100 + 2 * 10 + 3
        (i, ls) => fold_left_list(ls, fn {i, j => i * 10 + j }, i)
      })
  ).map(Token::Value)
```

词法分析

MoonBit 月兔编程

- 对输入流进行分析
  - 在单词之间可能存在空格

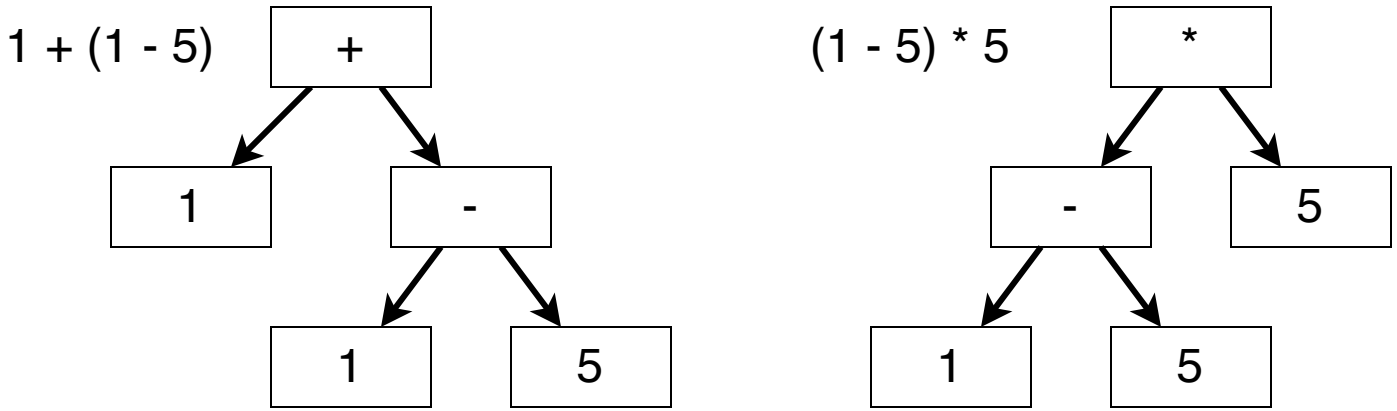
```
let tokens: Lexer[List[Token]] =
  number.or(symbol).and(whitespace.many())
    .map(fn { (symbols, _) => symbols }) // 忽略空格
    .many()

fn init {
  debug(tokens.parse("-10123-+-523 103    ( 5) )  "))
}
```

- 我们成功地分割了字符串
  - `-` `10123` `-` `+` `-` `523` `103` `(` `5` `)` `)`
  - 但这不符合数学表达式的语法

语法分析

- 对单词流进行分析，判断是否符合语法
  - 输入：单词流
  - 输出：抽象语法树

```
1. expression = Value / "(" expression ")"
2. expression =/ expression "+" expression / expression "-" expression
3. expression =/ expression "*" expression / expression "/" expression
```



语法分析

- 语法定义

```
1. expression = Value / "(" expression ")"
2. expression =/ expression "+" expression / expression "-" expression
3. expression =/ expression "*" expression / expression "/" expression
```

- 问题：运算符的优先级、结合性
  - 优先级：$\mathtt{a} + \mathtt{b} \times \mathtt{c} \rightarrow \mathtt{a} + (\mathtt{b} \times \mathtt{c})$
  - 结合性：$\mathtt{a} + \mathtt{b} + \mathtt{c} \rightarrow (\mathtt{a} + \mathtt{b}) + \mathtt{c}$
  - 当前语法具有二义性

语法分析

- 修改后的语法定义

```
1. atomic     = Value / "(" expression ")"
2. combine    = atomic  /    combine "*" atomic  /    combine "/" atomic
3. expression = combine / expression "+" combine / expression "-" combine
```

- 注意到除了简单的组合以外，出现了左递归
  - 左递归会导致我们的解析器进入循环
  - 解析器将尝试匹配运算符左侧的规则而不前进
  - 拓展：自底向上解析器可以处理左递归

语法分析

MoonBit
月兔编程

- 修改后的语法定义

```
atomic     = Value / "(" expression ")"
combine    = atomic  *( ("*" / "/") atomic)
expression = combine *( ("+" / "-") combine)
```

- 数据结构

```
enum Expression {
  Number(Int)
  Plus(Expression, Expression)
  Minus(Expression, Expression)
  Multiply(Expression, Expression)
  Divide(Expression, Expression)
}
```

MoonBit 月兔编程

语法解析

- 定义语法解析组合子

```
type Parser[V] (List[Token]) -> Option[(V, List[Token])]

fn parse[V](self : Parser[V], tokens : List[Token]) -> Option[(V, List[Token])] {
  (self.0)(tokens)
}
```

- 大部分组合子与 `Lexer[V]` 类似
- 递归组合：`atomic = Value / "(" expression ")"`
  - 延迟定义
  - 递归函数

MoonBit 月兔编程

递归定义

- 延迟定义
  - 利用引用定义 `Ref[Parser[V]]`：`struct Ref[V] { mut val : V }`
  - 在定义其他解析器后更新引用中内容

```
fn Parser::ref[Value](ref: Ref[Parser[Value]]) -> Parser[Value] {
  Parser(fn(input) {
    ref.val.parse(input)
  })
}
```

  - `ref.val` 将在使用时获取，此时已更新完毕

递归定义

MoonBit 月兔编程

- 延迟定义

```
fn parser() -> Parser[Expression] {
  // 首先定义空引用
  let expression_ref : Ref[Parser[Expression]] = { val : Parser(fn{ _ => None }) }

  // atomic = Value / "(" expression ")"
  let atomic =  // 利用引用定义
    (lparen.and(ref(expression_ref)).and(rparen).map(fn { ((_, expr), _) => expr}))
      .or(number)

  // combine = atomic *( ("*" / "/") atomic)
  let combine = atomic.and(multiply.or(divide).and(atomic).many()).map(fn {
    ...
  })

  // expression = combine *( ("+" / "-") combine)
  expression_ref.val = combine.and(plus.or(minus).and(combine).many()).map(fn {
    ...
  })

  expression_ref.val
}
```

递归定义

MoonBit 月兔编程

- 递归函数
  - 解析器本质上是一个函数
  - 定义互递归函数后，将函数装进结构体

```
fn recursive_parser() -> Parser[Expression] {
  // 定义互递归函数
  // atomic = Value / "(" expression ")"
  fn atomic(tokens: List[Token]) -> Option[(Expression, List[Token])] {
    lparen.and(
      Parser(expression) // 引用函数
    ).and(rparen).map(fn { ((_, expr), _) => expr})
      .or(number).parse(tokens)
  }
  fn combine(tokens: List[Token]) -> Option[(Expression, List[Token])] { ... }
  fn expression(tokens: List[Token]) -> Option[(Expression, List[Token])] { ... }

  // 返回函数代表的解析器
  Parser(expression)
}
```

MoonBit 月兔编程

语法树之外：Tagless Final

- 计算表达式，除了生成为抽象语法树再解析，我们还可以有其他的选择
- 我们通过“行为”来进行抽象

```
1. trait Expr {
2.   number(Int) -> Self
3.   op_add(Self, Self) -> Self
4.   op_sub(Self, Self) -> Self
5.   op_mul(Self, Self) -> Self
6.   op_div(Self, Self) -> Self
7. }
```

  - 接口的不同实现即是对行为的不同语义

语法树之外：Tagless Final

MoonBit 月兔编程

- 我们利用行为的抽象定义解析器

```
fn recursive_parser[E : Expr]() -> Parser[E] {
  let number : Parser[E] = ptoken(fn { Value(_) => true; _ => false})
    .map(fn { Value(i) => E::number(i) }) // 利用抽象的行为

  fn atomic(tokens: List[Token]) -> Option[(E, List[Token])] { ... }
  // 转化为 a * b * c * ... 和 a / b / c / ...
  fn combine(tokens: List[Token]) -> Option[(E, List[Token])] { ... }
  // 转化为 a + b + c + ... 和 a - b - c - ...
  fn expression(tokens: List[Token]) -> Option[(E, List[Token])] { ... }

  Parser(expression)
}

// 结合在一起
fn parse_string[E : Expr](str: String) -> Option[(E, String, List[Token])] {
  let (token_list, rest_string) = tokens.parse(str)?
  let (expr, rest_token) : (E, List[Token]) = recursive_parser().parse(token_list)?
  Some(expr, rest_string, rest_token)
}
```

MoonBit 月兔编程

语法树之外：Tagless Final

- 我们可以提供不同的实现，获得不同的诠释

```
enum Expression { ... } derive(Debug) // 语法树实现
type BoxedInt Int derive(Debug) // 整数实现
// 实现接口（此处省略其他方法）
fn BoxedInt::number(i: Int) -> BoxedInt { BoxedInt(i) }
fn Expression::number(i: Int) -> Expression { Number(i) }
// 解析
debug((parse_string_tagless_final("1 + 1 * (307 + 7) + 5 - (3 - 2)")
  : Option[(Expression, String, List[Token])])) // 获得语法树
debug((parse_string_tagless_final("1 + 1 * (307 + 7) + 5 - (3 - 2)")
  : Option[(BoxedInt, String, List[Token])])) // 获得计算结果
```

总结

MoonBit 月兔编程

- 本节课展示了一个语法解析器
  - 介绍了词法解析的概念
  - 介绍了语法解析的概念
  - 展示了语法解析组合子的定义与实现
  - Tagless Final的概念与实现
- 拓展阅读
  - 调度场算法
  - 斯坦福CS143 第1-8课 或
  - 《编译原理》前五章 或
  - 《现代编译原理》前三章
- 拓展练习
  - 实现兼容各类“流”的语法解析组合子