FACULTAD DE INGENIERÍA INFORMÁTICA

PROCESADORES DE LENGUAJES

Práctica 3: Desarrollo de un compilador sencillo - Memoria

Realizado por:
Guillermo Tomás Fernández Martín, Luís González Naharro
{GuillermoTomas.Fernandez, Luis.Gonzalez9}@alu.uclm.es

Índice

Listings

1. Introducción

A lo largo de esta práctica, se ha realizado un compilador para un lenguaje tipo Pascal, del cual se nos proporciona la gramática que lo genera, así como ciertas pautas para la creación de este compilador. El objetivo es realizar un análisis completo hasta el punto de generación de código, el cual originalmente era orientado a la máquina virtual ROSSI. Sin embargo, debido a la falta de máquina virtual, la práctica finalmente se detiene en la creación del Árbol de Sintaxis Abstracta, sin que haya generación de código.

Esta memoria detallará las distintas partes que componen al compilador en su estado actual, de manera independiente, y explicando las partes del código que sean necesarias. Además, el proyecto con todo el código fuente, escrito en Python, se entregará de manera conjunta a esta memoria.

2. Descripción formal del lenguaje

2.1. Categorías léxicas del lenguaje

En nuestro compilador, se ha optado por representar cada categoría léxica con una clase. Por ello, al explicar cada categoría se hará referencia a la clase que la representa en el código. Todas estas clases se encuentran en el fichero *componentes.py*

- **Operador de asignación (*OpAsigna*):** el operador de asignación se representa con dos puntos y un símbolo de igual (**:=**)
- **Llave de Apertura (*LlaveAp*):** La llave de apertura es el símbolo *{*
- **Llave de Cierre (*Llaveci*):** La Llave de cierre es el símbolo *}*
- **Paréntesis de Apertura (*ParentAp*):** El paréntesis de apertura es el símbolo *(*
- **Paréntesis de Cierre (*ParentCi*):** El paréntesis de cierre es el símbolo *)*
- **Corchete de Apertura (*CorAp*):** El corchete de apertura es el símbolo *[*
- **Corchete de Cierre (*CorCi*):** El corchete de cierre es el símbolo *]*

- **Punto (*Punto*):** El punto es el símbolo *.*
- **Coma (*Coma*):** La coma es el símbolo *,*
- **Punto y Coma (*PtoComa*):** El punto y coma es el símbolo *;*
- **Dos Puntos (*DosPtos*):** Los dos puntos es el símbolo *:*
- **Fin de fichero (*EOF*):** Esta categoría léxica, que no se representa con ningún caracter, nos permite saber cuando el fichero de entrada ha terminado de ser leído.
- **Operador de Suma (*OpAdd*):** Este operador puede tomar dos valores aritméticos (*+,* -).
- **Operador de Multiplicación (*OpMult*):** Este operador puede tomar dos valores aritméticos (**, /*)
- **Operador Relacional (*OpRel*):** Este operador puede tomar los siguientes valores: $=, <>, <, <=, >=, >$
- **Identificadores (*Identif*):** Esta categoría léxica representa los identificadores que pueden aparecer en nuestro sistema. Éstos se generan a partir de la siguiente gramática:

$$\begin{aligned} letra &\rightarrow [a-zA-Z] \\ digito &\rightarrow [0-9] \\ id &\rightarrow letra(letra|digito)^* \end{aligned}$$

- **Número (*Número*):** Esta categoría léxica representa los números que pueden aparecer en nuestro sistema, ya sean enteros o reales. Éstos se generan a partir de la siguiente gramática:

$$\begin{aligned} digitos &\rightarrow digito\ digito^* \\ fraccion\_opt &\rightarrow .\ digitos \mid \lambda \\ num &\rightarrow digitos\ fraccion\_opt \end{aligned}$$

- **Palabra Reservada (*PR*):** Esta categoría léxica representa el conjunto finito de palabras reservadas de nuestro sistema. Dichas palabras reservadas son: *PROGRAMA, VAR, VECTOR, ENTERO, REAL, BOOLEANO, INICIO, FIN, SI, ENTONCES, SINO, MIENTRAS, HACER, LEE, ESCRIBE, Y, O, NO, CIERTO* y *FALSO*. Todas ellas deberán ir en mayúsculas.

2.2. Máquina Discriminadora Determinista

En esta sección se explicará cómo funciona la MDD en la implementación de nuestro compilador. Para ello, vamos a clasificar nuestras categorías léxicas en tres grandes grupos: Símbolos, Palabras y Números.

2.2.1. Símbolos

Dentro de este grupo, consideraremos todos los símbolos unitarios, los operadores y el fin de fichero. En este grupo, la mayoría de los elementos son únicos, y por lo tanto la MDD puede detectar directamente a que categoría pertenecen. En el caso de los símbolos compuestos de dos caracteres (operador de asignación u operador relacional), leemos el siguiente caracter. En caso de que sea un espacio, no existe ambiguedad (en el primer caso, sería la categoría **Dos puntos**, por ejemplo). Si no, el segundo caracter nos elimina también la ambiguedad del valor que puede tomar el operador.

2.2.2. Palabras

En este grupo consideramos los identificadores y las palabras reservadas, ya que se leen igual. La MDD lee el elemento mientras cumpla con la gramática establecida anteriormente, y una vez finaliza busca en la tabla de Palabras Reservadas si coincide con algún elemento. En caso contrario, sabemos que es un identificador.

2.2.3. Números

Si el primer caracter que se lee es un dígito, la categoría léxica va a ser **Número**, ya que un identificador no puede comenzar por un dígito. La MDD lee en este caso, comprobando según la gramática especificada anteriormente si el número es Entero o Real.

2.3. Comprobación de la gramática (A rellenar)

3. Detalles de implementación

A la hora de implementar el sistema, se nos proporcionó una serie de ficheros con el esqueleto del código que tendríamos que completar. Dicho esqueleto contenía parte del analizador léxico, además de una implementación anterior del Árbol de Sintaxis Abstracta que se realizó el año anterior, y que nos ha servido de orientación para realizar nuestra propia implementación. Los ficheros proporcionados son:

- **componentes.py:** Es en este fichero donde se encuentran todas las clases que representan las distintas categorías léxicas. Sin embargo, las clases venían vacías, y se debía añadir el código necesario para almacenar la información necesaria (como el valor, en el caso de un Número) en cada clase.
- **flujo.py:** Contiene una clase que nos permite leer el flujo de entrada de un programa. Ya venía completo, así que no había que modificar este fichero
- **Prueba1.eje:** Este fichero contiene un pequeño programa de prueba para que se pueda testear el código que se va generando.
- **analex.py:** El esqueleto del analizador léxico. Contenía el constructor de la clase, una función main para testear el programa que se le pasara por parámetro, y un ejemplo de como analizar la entrada.

A partir de estos ficheros, se ha generado un compilador capaz de crear un Árbol de Sintaxis Abstracta del código que se especifica en la descripción de la práctica.

3.1. Analizador léxico

Esta parte de nuestro compilador es la que implementa la MDD que se ha especificado anteriormente. Gracias al archivo **flujo.py**, contamos con una función `siguiente()` que nos devuelve el siguiente caracter del fichero y una función `devuelve()` que nos permite devolver el caracter leído al flujo, para volver a analizarlo (en caso de que hayamos entrado en un camino erróneo)

Hemos creado la función `Analiza()`, que nos devuelve el componente léxico que viene a continuación. Para ello, lee el primer caracter para detectar la categoría léxica a la que pertenece con un `if-else`, ya que Python no permite la instrucción `switch`. Si el primer caracter es un espacio, o una llave de apertura (que nos marca el comienzo de un comentario) omite dicho caracter y sigue leyendo caracteres hasta que encuentra algo distinto del espacio (Listing 1) o la llave de cierre (Listing 2)

Listing 1: Omisión de espacios

```
1  if ch==" ":
2   while ch == " ":
3     ch = self.flujo.siguiente()
4   self.flujo.devuelve(ch)
5   return self.Analiza()
```

Listing 2: Tratamiento de comentarios

```
elif ch == "{":
 while ch != "}":
   ch = self.flujo.siguiente()
 return self.Analiza()
```

Cuando el caracter que lee la función es un caracter alfabético, sigue leyendo todos los caracteres alfanuméricos que le siguen hasta encontrar un espacio. Para ello, se hace uso de las funciones predefinidas de Python, que detectan los dígitos, caracteres, etc. Al completar la palabra, comprueba si se encuentra en la tabla de palabras reservadas, y devuelve la categoría léxica apropiada. (Listing 3)

Listing 3: Detección de palabras

```
elif ch.isalpha():
 word = ""
 while ((ch).isalnum()):
   word += ch
   ch = self.flujo.siguiente()
 self.flujo.devuelve(ch)
 if word in self.PR:
   return componentes.PR(word, self.nlinea)
 else:
   return componentes.Identif(word, self.nlinea)
```

Cuando el caracter que se lee es un dígito, quiere decir que la categoría va a ser un Número. La función lee y guarda todos los dígitos hasta que se encuentra un espacio o un punto. En el primer caso, la función almacena que se trata de un Entero y devuelve la categoría correspondiente. En el segundo caso, la función comprueba que existe como mínimo un dígito después del punto, para almacenar que se trata de un Real; o devuelve un error. (Listing 4)

Listing 4: Detección de números

```
1       elif ch.isdigit():
2       num = ""
3       num+=ch
4       ch=self.flujo.siguiente()
5       while(ch.isdigit()):
6          num+=ch
7          ch=self.flujo.siguiente()
8       if (ch != '.'):
9          self.flujo.devuelve(ch)
10         return componentes.Numero(num, self.nlinea, 'ENTERO')
11      else:
12         newCh = self.flujo.siguiente()
13         if not (newCh.isdigit()):
14            self.flujo.devuelve(newCh)
15            self.flujo.devuelve(ch)
16            print "ERROR: NUMERO REAL MAL ESPECIFICADO" # tenemos un
                  comentario no abierto
17            return self.Analiza()
18         num+=ch
19         num+=newCh
20         ch=self.flujo.siguiente()
21         while((ch).isdigit()):
22            num+=ch
23            ch=self.flujo.siguiente()
24         self.flujo.devuelve(ch)
25         return componentes.Numero(num, self.nlinea, 'REAL')
```

Por último, los caracteres no visibles como los saltos de línea o retornos de carro se tratan con su correspondiente codificación numérica, al no tener otra herramienta para tratarlos (Listing 5). Cualquier otro caracter no nulo hace que la función devuelva un error, mientras que al encontrar el primer caracter nulo se considera que se ha llegado al fin de fichero y se devuelve el componente correspondiente (Listing 6).

Listing 5: Caracteres no visibles

```
1    elif (ch is not '' and ord(ch) == 13):
2       self.nlinea += 1
3       return self.Analiza()
4    elif (ch is not '' and ord(ch) == 10):
5       return self.Analiza()
```

Listing 6: Errores y fin de fichero

```
1   elif len(ch) is not 0:
2     print "ERROR: CARACTER NO DEFINIDO EN LA ESPECIFICACION DEL
          LENGUAJE"
3     return self.Analiza()
4   else:
5     return componentes.EOF()
```

3.2. Analizador semántico

Para la construcción del analizador semántico, no disponíamos de un esqueleto de código, pero si de unas instrucciones básicas en el material de clase de teoría. A partir de éstas, pudimos deducir la estructura del analizador semántico. En primer lugar, se han creado dos funciones que nos ayudarán con las comprobaciones. La primera es una función `avanza()` que nos devuelve el siguiente componente en nuestro programa, extraido del analizador léxico explicado en la sección anterior. La segunda función, `comprueba()`, nos realiza la comprobación de que un componente pertenece a una determinada categoría léxica, o devuelve un error.

La aproximación que se ha tomado a la hora de programar el analizador semántico es la creación de una función por cada no terminal. En estas funciones, se van comprobando uno a uno los elementos que deberían aparecer, según la especificación sintáctica proporcionada en el material de la práctica.

El primer paso es comprobar que el primer elemento leído pertenece a los primeros de la regla, que ya se han sacado y explicado anteriormente. Así, podemos detectar a qué regla ha entrado nuestro programa. Gracias a la función `comprueba()` definida anteriormente, se van comprobando todos los terminales que deberían aparecer. En caso de que el elemento siguiente sea un no terminal, se llama a la función de dicho no terminal. Si la regla de producción es una regla vacía, no se hace nada en ese paso. Se puede observar como se ha hecho la implementación del no terminal $\langle resto\_term \rangle$ (Listings 7)

$$
\begin{aligned}
\langle resto\_term \rangle &\rightarrow \textbf{opmult } \langle factor \rangle \langle resto\_exsimple \rangle \\
\langle resto\_term \rangle &\rightarrow \textbf{O } \langle factor \rangle \langle resto\_exsimple \rangle \\
\langle resto\_term \rangle &\rightarrow \lambda
\end{aligned}
$$

Listing 7: analizaRestoTerm()

```
1       def analizaRestoTerm(self):
2          if self.componente.cat == "OpMult":
3             self.avanza()
4             self.analizaFactor()
5             self.analizaRestoTerm()
6          elif self.componente.cat == "PR" and self.componente.valor
               == "Y":
7             self.avanza()
8             self.analizaFactor()
9             self.analizaRestoTerm()
10         elif (self.componente.cat == "PR" and self.componente.valor
               in ["ENTONCES", "HACER", "SINO", "O"]) or
               self.componente.cat == "ParentCi" or
               self.componente.cat == "CorCi" or self.componente.cat
               == "OpRel" or self.componente.cat == "PtoComa" or
               self.componente.cat == "OpAdd":
11            pass
12         else:
13            print "Error: SE ESPERABA Resto de Termino en linea " +
                  str(self.lexico.nlinea)
14            while not ((self.componente.cat == "PR" and
                  self.componente.valor in ["O", "ENTONCES", "HACER",
                  "SINO"]) or self.componente.cat == "OpRel" or
                  self.componente.cat == "OpAdd" or
                  self.componente.cat == "CorCi" or
                  self.componente.cat == "ParentCi" or
                  self.componente.cat == "PtoComa"):
15               self.avanza()
16            return
```

Para el tratamiento de errores, se ha optado por un tratamiento en modo pánico. Esto quiere decir que si el componente que encontramos no es el componente esperado, avanzamos y tiramos todos los componentes que vienen a continuación hasta llegar a uno de los siguientes del no terminal que ha dado el problema. A partir de ahí, podemos continuar el análisis de una manera normal, habiendo informado del problema. Ésto se observa muy bien en el código anterior (Listings 7), ya que la condición `else` muestra un bucle que va avanzando en el analizador léxico sin hacer nada con los componentes mientras el elemento no pertenezca a los siguientes de $\langle resto\_term \rangle$. Además, en el momento que encuentre un error, el compilador avisa de la línea de error y del tipo de error encontrado, mostrándolo por pantalla.