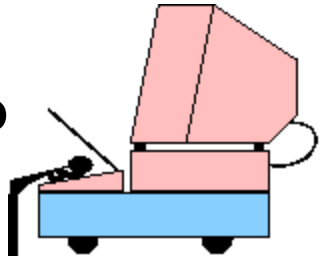
Práctica 1: análisis léxico y sintáctico



Objeto de la sesión

El objetivo de esta sesión es desarrollar un analizador léxico y un analizador sintáctico de acuerdo con la especificación que se presenta más adelante. Dichos analizadores recibirán como entrada un fichero de texto que contiene el programa fuente y en el caso de que no haya errores léxicos ni sintácticos producirán como resultado un objeto Java que represente el programa fuente en forma de árbol. Si hay errores léxicos o sintácticos se levantará una excepción.

En este documento se presenta una descripción general del programa fuente que se utilizará en la práctica. Se debe notar que ciertas partes de dicha descripción no se van a utilizar en la práctica 1. En particular, las comprobaciones relativas a los tipos de datos se realizarán en la práctica 2.

Tutoriales

Antes de empezar a realizar la práctica se recomienda consultar los manuales de CUP y JLex disponibles a través de Aula Global o bien estos pequeños resúmenes de CUP y JLex

Definición del lenguaje de programación a traducir

Descripción general del lenguaje de programación fuente PF2024

Empecemos con un pequeño programa de ejemplo:

```
pf2024 Ejem1

decl

intset a, b, c, d;
int a1,b1,c1,d1;

begin
a:= {0, 1, 3, 5, 7, 9};
b:= {0, 2, 4, 6, 8};
c:= a union b;
d:= a intersec b;
a1:= card(a);
printInt(a1);
b1:= card(b);
printInt(b1);
c1:= card(c);
printInt(c1);
d1:= card(d);
printInt(d1);
end;
```

Un programa en el lenguaje de programación `PF2024` consta de las siguientes partes:

1. Palabra clave `pf2024` que indica el inicio del programa.

2. Un identificador (el nombre del programa).

3. Opcionalmente, pueden declararse variables que serán utilizadas dentro del programa. Las declaraciones de variables están formadas por un tipo de datos seguido de una lista de identificadores (nombres de variables). Los tipos de datos existentes en este lenguaje de programación son `int`, `bool` e `intset`.

   Cuando se declara una variable se supondrá que inicialmente tomará el valor `false` si la variable es de tipo `bool`, 0 si la variable es de tipo `int` y conjunto vacío si la variable es de tipo `intset`.

4. Cuerpo del programa, que consta de una secuencia de sentencias. Comienza con la palabra clave `begin` y termina con `end;`.

   Las sentencias de un programa pueden ser de los siguientes tipos:

   - Sentencia de asignación: almacena en la variable indicada a la izquierda del `:=` el valor que se obtiene como resultado de evaluar la expresión a la derecha del `:=`.

   - Sentencia condicional `if <Exp> then <StamentList> end`.

   - Bucle `while <Exp> do <StamentList> end`.

   - Sentencia `printInt`. Esta sentencia recibe como argumento una variable de tipo `int`. El resultado de ejecutar esta sentencia es que se imprime por pantalla el siguiente mensaje:

     ```
     El valor de la variable <variable> es: <x>
     ```

     Donde `<variable>` es el nombre de la variable y `<x>` es el valor almacenado en dicha variable en el momento en que se ejecuta la sentencia `printInt`.

     Nota: la comprobación de que la variable que se pasa como argumento a la sentencia `printInt` es de tipo `int` se realizará en la práctica 2.

   - Sentencia `printBool`. Esta sentencia recibe como argumento una variable de tipo `bool`. El resultado de ejecutar esta sentencia es que se imprime por pantalla el siguiente mensaje:

     ```
     El valor de la variable <variable> es: <x>
     ```

     Donde `<variable>` es el nombre de la variable y `<x>` es el valor almacenado en dicha variable en el momento en que se ejecuta la sentencia `printBool`.

     Nota: la comprobación de que la variable que se pasa como argumento a la sentencia `printBool` es de tipo `bool` se realizará en la práctica 2.

Especificación del analizador léxico

El analizador léxico JLex a desarrollar debe reconocer los siguientes tokens:

- Palabras clave: `begin` (token `BEGIN`), `and` (token `AND`), `or` (token `OR`), `not` (token `NOT`), `pf2024` (token `PROG`), `decl` (token `DECL`), `while` (token `WHILE`), `do` (token `DO`), `printInt` (token `PRINT_I`), `printBool` (token `PRINT_B`), `in` (token `IN`), `union` (token `UNION`), `intersec` (token `INTERSEC`), `card` (token `CARD`), `lowest_elem` (token `LOWEST_ELEM`), `emptyset` (token `EMPTYSET`), `if` (token `IF`), `then` (token `THEN`) y `end` (token `END`).

- Las palabras clave para los tipos básicos: `int`, `bool` e `intset` comparten el mismo token: `TYPE`.
- Signos de puntuación y operadores: ";" (token `PC`), ":=" (token `ASIG`), "<" (token `MENOR`), ">" (token `MAYOR`), "+" (token `MAS`), "-" (token `MENOS`), "/" (token `DIV`), "*" (token `MUL`), "(" (token `PAREN`), ")" (token `TESIS`), "=" (token `IGUAL`), "," (token `COMA`), "{" (token `BRAC`) y "}" (token `KET`).
- Identificadores (token `IDENT`): comienzan por una letra (mayúscula o minúscula), seguida opcionalmente de cualquier cadena de letras y dígitos.
- Constantes, que pueden ser de dos tipos:
  - Números enteros no negativos (token `CINT`), es decir, cualquier cadena de dígitos. Ejemplos: `1`, `0`, `2010`
  - Las constantes lógicas `true` y `false` (token `CLOG`)

El lenguaje distingue mayúsculas y minúsculas; por lo tanto mientras `begin` es una palabra clave, `Begin` es un identificador y `BEGIN` es otro identificador distinto del anterior.

Recuerde que también tiene que especificar en el fichero JLex que se deben reconocer y consumir (sin generar token) los espacios en blanco, tabuladores, saltos de línea y retornos de carro.

Para desarrollar el analizador léxico pedido lo único que hay que hacer es completar este esqueleto de fichero JLex en el que lo único que falta es definir para cada token a reconocer, la expresión regular asociada y la acción asociada. A modo de ejemplo, se proporciona ya definido el token para la palabra clave `and`.

Como puede observarse en la definición de la regla para la palabra clave `and`, lo que hay que hacer es:

1. Para cada token definir la expresión regular asociada, según el formato requerido por JLex (véase la documentación que se proporciona).
2. A continuación de la expresión regular definir la acción que se toma cuando se reconoce dicho token. Para los tokens que no tienen ningún dato asociado, escribiremos `{return tok(sym.X, null); }`, donde `X` es el nombre que se le ha dado al token en cuestión en el fichero CUP.

   Para los tokens que tienen un dato asociado, escribiremos `{return tok(sym.X, obj); }`, donde `X` es el nombre que se le ha dado al token en cuestión en el fichero CUP y `obj` es el objeto Java asociado al token. Para construir el objeto Java asociado a un token puede utilizar el método `yytext()`, tal y como se explica en la documentación de JLex.

   Como se explica en el manual de CUP y se ve en clase, un fichero CUP declara una lista de tokens, donde cada token es representado por un idendificador (por ejemplo, `AND`). Cuando se genera el analizador sintáctico a partir del fichero CUP, se producen como resultado 2 ficheros Java. Uno de ellos, que lleva por nombre `sym.java` contiene una clase llamada `sym` en la que simplemente se asocia un número entero a cada uno de los tokens que se han declarado en el fichero CUP. Por lo tanto, `sym.X` es el entero que utiliza la clase `sym` para identificar al token `X`.
3. En el caso de los espacios en blanco, saltos de línea, etc., escribiremos como acción `{ }`, cuyo efecto es no devolver nada y seguir consumiendo caracteres de entrada para producir el siguiente token.

El analizador léxico deberá lanzar una excepción de tipo `LexerException`, que se proporciona, cuando se detecte un error léxico.

Definición de la sintaxis del lenguaje de programación fuente

A continuación vamos a dar la definición de la sintaxis del lenguaje fuente por medio de una gramática independiente del contexto (excepto para las expresiones, que se van a definir en lenguaje natural).

La gramática que define la sintaxis del lenguaje de programación fuente es la siguiente:

```
G=<ET, EN, S, P>
ET= {COMA, PC, PAREN, TESIS, BEGIN, END, ASIG, AND, OR, NOT,
```

```
    IF, THEN, PROG, DECL, WHILE, DO, PRINT_I, PRINT_B, BRAC, KET,
    IN, UNION, INTERSEC, CARD, LOWEST_ELEM, MAS, MENOS, MUL,
    DIV, MAYOR, MENOR, IGUAL, IDENT, CLOG, CINT, TYPE, EMPTYSET}

EN= {S, <IdentList>, <StatementList>, <Body>, <Statement>,
    <Exp>, <LDecl>, <Decl>}

S= S

P= {
S->  PROG IDENT DECL <LDecl> <Body>
   |  PROG IDENT <Body>

<LDecl>-> <Decl> PC
       | <Decl> PC <LDecl>

<Decl>-> TYPE <IdentList>

<IdentList>-> IDENT
            | IDENT COMA <IdentList>

<Body>-> BEGIN <StatementList> END PC

<StatementList>-> <Statement> PC
                | <Statement> PC <StatementList>

<Statement>-> IDENT ASIG <Exp>
            | PRINT_I PAREN IDENT TESIS
            | PRINT_B PAREN IDENT TESIS
            | IF <Exp> THEN <StatementList> END
            | WHILE <Exp> DO <StatementList> END
}
```

Expresiones en el lenguaje de programación fuente

Las expresiones del lenguaje de programación fuente (identificadas en la gramática por el símbolo no terminal `<Exp>`) deben de ajustarse a lo siguiente:

Expresiones sin tipo definido:

- Un identificador que sea el nombre de una variable. Será del tipo de la variable.
- Una expresión entre paréntesis. Será del tipo de la expresión contenida dentro de los paréntesis.

Expresiones de tipo `int`:

- Una constante expresada en decimal.
- Una suma, resta, multiplicación o división de expresiones de tipo `int` utilizando los operadores habituales "+", "-", "*", "/".
- El opuesto de una expresión de tipo `int`, utilizando el símbolo habitual "-" antes de la expresión.
- Una expresión que pertenezca al lenguaje definido por la expresión lingüística `card "(" <Exp> ")"`, donde la subexpresión entre paréntesis ha de ser de tipo `intset` (conjunto de enteros). El resultado de evaluar esta expresión será el número de elementos del conjunto que se obtendría al evaluar la subexpresión de tipo `intset`.
- Una expresión que pertenezca al lenguaje definido por la expresión lingüística `lowest_elem "(" <Exp> ")"`, donde la subexpresión entre paréntesis ha de ser de tipo `intset` (conjunto de enteros). El resultado

de evaluar esta expresión será el valor del menor elemento del conjunto que se obtendría al evaluar la subexpresión de tipo `intset`. **Nota**: el código generado en la práctica 3 debe lanzar una excepción cuando `lowest_elem` sea aplicado al conjunto vacío.

Expresiones de tipo `intset` (conjunto de enteros):

- Constante conjunto vacío (`{ }` o `emptyset`).
- Una expresión que comienza por `{`, termina por `}` y contiene una lista de subexpresiones de tipo `int`, separadas por comas. Representa un conjunto de enteros cuyos elementos son los enteros resultado de evaluar las subexpresiones.
- Unión, intersección y diferencia de conjuntos utilizando los operadores `union`, `intersec` y "`-`", respectivamente. Dados dos conjuntos `A` y `B`, la diferencia de conjuntos `A - B` es el conjunto que está formado por aquellos elementos de `A` que no pertenecen a `B`.

Expresiones de tipo `bool`:

- Las constantes `true` y `false`.
- La conjunción o disyunción de expresiones lógicas, utilizando los operadores `and` y `or` respectivamente.
- Una expresión lógica negada, utilizando el operador `not` antes de la expresión lógica.
- La igualdad entre 2 expresiones utilizando el operador "=". Las dos expresiones comparadas deberán de ser del mismo tipo. Se entiende que dos conjuntos son iguales cuando contienen exactamente los mismos elementos.
- La comparación entre 2 expresiones de tipo `int` utilizando los operadores "<" (menor que) o ">" (mayor que).
- Una expresión que pertenezca al lenguaje definido por la expresión lingüística `<Exp> in <Exp>`, donde la primera subexpresión ha de ser de tipo `int` y la segunda de tipo `intset`. Devuelve `true` si el entero resultado de evaluar la primera subexpresión pertenece al conjunto resultado de evaluar la segunda y `false` en caso contrario.

Nota: para definir las expresiones puede ser necesario definir algún símbolo no terminal adicional en la gramática.

Reglas de precedencia

La precedencia de los operadores es, de menor a mayor:

1. `union`
2. `intersec`
3. `in`
4. `or`
5. `and`
6. `not`
7. =
8. < y >
9. + y -
10. * y /
11. - (opuesto)

Todos los operadores asocian por la izquierda.

Se proporciona un esqueleto del fichero en formato CUP en el que solamente falta incluir las declaraciones de símbolos terminales y no terminales, las reglas de precedencia y las reglas de producción.

Requisitos que deben cumplir los ficheros JLex y CUP

Para la prueba del analizador léxico y el analizador sintáctico, se utilizara la clase Main que se proporciona. La ejecución de dicha clase se realizará de acuerdo con lo siguiente:

```
java Main <nombre_fichero>
```

Donde `<nombre_fichero>` es el nombre del fichero (con el path si es necesario) del programa fuente a analizar.

El programa deberá de levantar una excepción que no deberá de ser capturada en el caso de que el programa fuente tenga algún error de sintaxis. Los mensajes emitidos por excepciones producidas por errores en la fase de análisis sintáctico deberán indicar una línea del programa orientativa del lugar en el que se ha producido el error.

Asímismo, el programa deberá de levantar una excepción de tipo `LexerException`, que no deberá de ser capturada en el caso de que el programa fuente tenga algún error léxico.

Obligatoriamente las clases generadas por CUP deberán pertenecer a un paquete llamado Parser y la clase generada por JLex deberá pertenecer a un paquete llamado Lexer.

Orden de compilación

Antes de entregar la práctica deberá cerciorarse de que su práctica puede compilarse correctamente con `javac` siguiendo el siguiente orden:

1. Clases del paquete `Errors`.
2. Clases del paquete `AST` que usted debe desarrollar.
3. Clases `parser` y `sym` generadas por CUP.
4. Clase `Yylex` generada por JLex.
5. Clase `Main` que se proporciona.

Aquellas prácticas que no se puedan compilar en este orden serán calificadas con 0.

Ayudas y sugerencias

Se proporciona un juego de tests con el que probar la práctica. De entre ellos solamente deberían de producir un error de sintaxis los tests en carpetas cuyo nombre comienza por `ErrSint` y error léxico los tests en carpetas cuyo nombre comienza por `ErrLex`. Puede ocurrir que alguno de los ejemplos `ErrLex` de error de sintaxis (y no léxico).

Se advierte que se realizarán tests adicionales a los proporcionados a las prácticas recibidas, por lo que se recomienda a los alumnos que planifiquen tests complementarios a su práctica.

Ficheros a entregar

Se deberán entregar exclusivamente los siguientes ficheros:

- fichero `Yylex` en formato JLex para el analizador léxico

- fichero `parser` en formato CUP para el analizador sintáctico
- fichero `AST.zip`, que contiene exclusivamente un directorio de nombre `AST` que a su vez contiene exclusivamente los ficheros `.java` que usted ha desarrollado para modelar árboles de sintaxis abstracta.

Last Revision: 03/04/2024 14:25:10