ICI 4242 - Autómatas y compiladores

Lenguajes y gramáticas formales

Rodrigo Olivares
Msc. en Ingeniería Informática
rodrigo.olivares@uv.cl

1er Semestre

Contenido

Introducción

Origen

Origen

La teoría de los lenguajes formales tienen su origen en el campo de la *lingüística*. En la década de los 50, los lingüistas elaboraron ideas informales acerca de la **gramática universal**.

Gramática Universal

La gramática universal caracteriza las propiedades generales del lenguaje humano (por ejemplo: oraciones y frases).

Introducción

Origen

Origen en la Informática

En el campo de la Informática, el concepto de *Gramática Formal* adquirió gran importancia para la **especificación de lenguajes de programación**; concretamente, se definió con sus teorías la sintaxis del lenguaje **ALGOL 60**, usándose una gramática libre de contexto. Ello condujo rápidamente al diseño riguroso de algoritmos de traducción y compilación.

Introducción

Origen

Origen en la Informática

Finalmente, y enlazando con el campo de la lingüistica, la **Teoría de Lenguajes Formales** es de gran utilidad para el trabajo en otros campos de la Informática, por ejemplo:

- → *Inteligencia Artificial*
- → *Procesamiento de Lenguajes Naturales* (comprensión, generación, y traducción)
- → *Reconocimiento del Habla*
- → Entre otros.

Introducción

Definiciones

Símbolo

- ✓ Es una entidad **abstracta** (no se define, *axioma*).
- ✓ Normalmente los símbolos son: *letras*, *dígitos* y otros caracteres.
- ✓ Los símbolos también pueden estar formados por varias letras o caracteres, así por ejemplo las **palabras reservadas de un lenguaje de programación** son símbolos de dicho lenguaje.

Ejemplo

- ✓ a, b, c, d, . . .
- ✓ 1, 2, 3, 4, . . .
- ✓ $+, *, \#, ?, \ldots$
- ✓ **if**, **else**, **switch**, **while**, . . .

Introducción

Definiciones

Vocabulario o Alfabeto

- ✓ Es un conjunto finito de símbolos, **no vacío**.
- ✓ Para definir que un símbolo a pertenece a un alfabeto Σ se utiliza la notación $a \in \Sigma$.
- ✓ Los alfabetos se definen por **enumeración de los símbolos que contienen**.

Ejemplo

- ✓ $\Sigma_1 = \{A, B, C, D, E, V, W, X, Y, Z\}$
- ✓ $\Sigma_2 = \{a, b, c, 0, 1, 2, 3, 4, *, \#, +\}$
- ✓ $\Sigma_3 = \{0, 1\}$
- ✓ $\Sigma_4 = \{if, else, switch, while, do, a, b, ;, (,), =, >\}$

Introducción

Definiciones

Cadena o Palabra

- ✓ Secuencia **finita** de símbolos de un determinado alfabeto.

Ejemplo: Se utilizan los alfabetos del ejemplo anterior

- ✓ *ABCD* es una cadena del alfabeto Σ_1.
- ✓ $a + 2 * b$ es una cadena del alfabeto Σ_2.
- ✓ 000111 es una cadena del alfabeto Σ_3.
- ✓ *if(a>b)b=a;* es una cadena del alfabeto Σ_4.

Introducción

Definiciones

Longitud de cadena

- ✓ La longitud de una cadena es el número de símbolos que contiene.

Ejemplo: Se utilizan las cadenas del ejemplo anterior

- ✓ $|ABCD| = 4$
- ✓ |a+2*b| $= 5$
- ✓ |000111| $= 6$
- ✓ $|if(a>b)b=a;| = 10$

Introducción

Definiciones

Cadena vacía

✓ Existe una cadena denominada **cadena vacía**, que no tiene símbolos y se denota con λ, entonces su longitud es :

$$|\lambda| \rightarrow 0$$

Introducción

Definiciones

Concatenación de cadenas

✓ Sean α y β dos cadenas cualesquiera, se denomina concatenación de α y β a una nueva cadena $\alpha\beta$ constituida por los símbolos de la cadena α **seguidos** por los de la cadena β.

Elemento neutro

✓ El elemento neutro de la concatenación es λ :

$$\alpha\lambda = \lambda\alpha = \alpha$$

Introducción

Definiciones

Universo del discurso o Clausura

- ✓ El conjunto de **todas las cadenas** que se pueden formar con los símbolos de un alfabeto Σ se denomina universo del discurso (o clausura) de Σ y se representa por $W(\Sigma)$ ó Σ^*.
- ✓ Evidentemente Σ^* es un **conjunto infinito**.
- ✓ La cadena vacía **pertenece** a Σ^*.

Ejemplo

- ✓ Sea un alfabeto con un único símbolo, $\Sigma = \{a\}$, entonces el universo del discurso Σ^* es:

$$\Sigma^* = \{\lambda, a, aa, aaa, aaaa, \ldots\}$$
$$\{a^n \mid n \geq 0\}$$

Introducción

Definiciones

Lenguaje

- ✓ Se denomina lenguaje sobre un alfabeto Σ a un **subconjunto del universo del discurso**. También se puede definir como un conjunto de palabras de un determinado alfabeto.
- ✓ Habitualmente un lenguaje tiene infinitas cadenas, por lo que definirlo por enumeración es **ineficiente** y a veces **imposible**.
- ✓ Así los lenguajes se defienen por las **propiedades que cumplen** las cadenas del lenguaje.

Introducción

Definiciones

Ejemplo

- ✓ El conjunto de palíndromos (cadenas que se leen igual hacia adelante, que hacia atrás) sobre el alfabeto Σ_3. Evidentemente este lenguaje tiene infinitas cadenas.

$$\begin{gathered}\lambda\\0\\11\\010\\10101\\00000\\1111111\end{gathered}$$

Introducción

Definiciones

Lenguaje vacío

- ✓ Conjunto vacío y que se denota por $\emptyset$.
- ✓ El lenguaje vacío no debe confundirse con un lenguaje que contenga una sola cadena, y que ésta sea la cadena vacía, es decir $\{\lambda\}$, ya que el número de elementos (cardinalidad) de estos dos **conjuntos es diferente**.

$$Cardinal(\emptyset) = 0$$

$$Cardinal(\{\lambda\}) = 1$$

Introducción

Definición formal de gramática

Gramática

- ✓ N-tupla que permite especificar, de una manera finita, el conjunto de cadenas de símbolos que constituyen un lenguaje.

Introducción

Definición formal de gramática

Cuádrupla

$$G = (\Sigma, N, S, P)$$

donde:

- ✓ $\Sigma = \{$conjunto finito de símbolos terminales$\}$.
- ✓ $N = \{$conjunto finito de símbolos no terminales$\}$.
- ✓ S es el símbolo inicial y pertenece a N.
- ✓ $P = \{$conjunto de producciones o de reglas de derivación$\}$.

Introducción

Definición formal de gramática

Definición Σ

Todas las cadenas del lenguaje definido por la gramática están formados con símbolos del **alfabeto terminal** Σ. El alfabeto terminal se define por enumeración de los símbolos terminales.

Cadena vacía

- ✓ En ocasiones es importante distinguir si un determinado alfabeto incluye o no la cadena vacía, indicándose respectivamente con superíndice $^{+}$, o superíndice *, tal como se muestra a continuación :

$$\Sigma^{+} = \Sigma - \{\lambda\}$$
$$\Sigma^{*} = \Sigma + \{\lambda\} \quad \leftarrow \textit{universo del discurso}$$

Introducción

Definición formal de gramática

Generalización

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \cdots$$

donde $\Sigma^1 = \Sigma$ y $\Sigma^k = \Sigma \times \Sigma^{k-1}$ denota, el conjunto de todas las secuencias finitas de símbolos de Σ. El conjunto Σ^0 es especial, tiene un sólo elemento llamado λ, que corresponde a la cadena vacía.

Si una cadena $x \in \Sigma^k$, entonces decimos que su largo es $\mid x \mid = k$ (por ello $\mid \lambda \mid = 0$).

Introducción

Definición formal de gramática

Definición N

El **alfabeto no terminal** N es el conjunto de símbolos introducidos como elementos auxiliares para la definición de la gramática, y que **no figuran en las sentencias del lenguaje**. El alfabeto no terminal se define por enumeración de los símbolos no terminales.

Introducción

Definición formal de gramática

Símbolo inicial S

✓ El símbolo inicial S es un símbolo **no terminal** a partir del cual se aplican las reglas de la gramática para obtener las distintas cadenas del lenguaje.

Introducción

Definición formal de gramática

Las producciones P

- ✓ Son las reglas que se aplican desde el símbolo inicial para obtener las cadenas del lenguaje.
- ✓ El conjunto de producciones P se define por medio de la enumeración de las distintas producciones, en forma de reglas o por medio de un *metalenguaje* por ejemplo *BNF* (Backus Naur Form) o *EBNF* (Extended Backus Naur Form).

Introducción

Definición formal de gramática

Ejemplo

✓ Sea la gramática : $G = (\Sigma, N, S, P)$ donde $\Sigma = \{a, b\}$, $N = \{S\}$, y el conjunto de producciones es :

1. $S \rightarrow ab$
2. $S \rightarrow aSb$

Introducción

Definición formal de gramática

Dada la gramática anterior, determinar si las cadenas *ab*, *aaabbb* y *aabbb* son reconocidas (**estructura formal de derivación**).

$S \Rightarrow ab$ ✓

$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaabbb$ ✓

La cadena *aabbb* no es reconocida por la gramática.

* La gramática reconoce las cadenas (son símbolos terminales) de longitud par y con la misma cantidad de símbolos *a* y **b** concatenados.

Introducción

Definición formal de gramática

Ejemplo

✓ Según el siguiente conjunto de producciones :

$$
\begin{array}{ll}
1. & S \rightarrow aA \\
2. & S \rightarrow bA \\
3. & A \rightarrow aB \\
4. & A \rightarrow bB \\
5. & A \rightarrow a \\
6. & B \rightarrow aA \\
7. & B \rightarrow bA
\end{array}
\quad \Leftrightarrow \quad
\begin{array}{ll}
1. & S \rightarrow aA \mid bA \\
2. & A \rightarrow aB \mid bB \mid a \\
3. & B \rightarrow aA \mid bA
\end{array}
$$

Introducción

Definición formal de gramática

Determinar la **estructura formal de derivación** para las siguientes cadenas y la **gramática** asociada

1. a
2. b
3. $aaaaaa$
4. $bbbbba$
5. $abbaaabbbba$

Introducción

Definición formal de gramática

Solución

- × a (Repeticiones únicas de a con cardinal par)
- × b (Por reglas de la gramática, las cadenas terminan con a)
- ✓ $aaaaaa$ (Repeticiones únicas de a con cardinal par)

  $S \Rightarrow aA \Rightarrow aaB \Rightarrow aaaA \Rightarrow aaaaB \Rightarrow aaaaaA \Rightarrow aaaaaa$
- ✓ $bbbbba$

  $S \Rightarrow bA \Rightarrow bbB \Rightarrow bbbA \Rightarrow bbbbB \Rightarrow bbbbbA \Rightarrow bbbbba$
- ✓ $abbaaabbbba$

  $S \Rightarrow aA \Rightarrow abB \Rightarrow abbA \Rightarrow abbaB \Rightarrow abbaaB \Rightarrow abbaaaA \Rightarrow abbaaabB \Rightarrow abbaaabbA \Rightarrow abbaaabbbB \Rightarrow abbaaabbbbA \Rightarrow abbaaabbbba$

Gramática: $G = (\Sigma, N, S, P)$ donde $\Sigma = \{a, b\}$, $N = \{S, A, B\}$, S es el símbolo inicial y P el conjunto de producciones.

Preguntas

Preguntas ?