Ciencias de la Computación II

Alfabetos, Cadenas y Lenguajes

Eduardo Contrera Schneider

Universidad de la Frontera

17 de agosto de 2016

Definiciones

Pensemos primero en el alfabeto inglés. Éste está compuesto por una cantidad finita de símbolos que se unen para formar secuencias finitas de símbolos (palabras). También no debemos olvidar que además de las letras se usan otro tipo de símbolos, tales como el guión y el apóstrofe, para formar las palabras que pertenecen al lenguaje. También podemos considerar otros tipos de ejemplos.

Ejemplos

- La representación de números enteros son secuencias de carácteres del conjunto de los dígitos $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
- Los programas escritos en C/C++ se hace en base a la disposición secuencial de símbolos, donde el conjunto de símbolos es una colección de identificadores legales, palabras clave, palabras reservadas y símbolos especiales.
- Entre otros.

Lo anterior motiva la siguiente definición:

Definición: Alfabeto

Se conoce como **Alfabeto** a un conjunto **no vacío** y **finito** de símbolos. Se utiliza la letra griega Σ para denotar un alfabeto.

Si Σ es un alfabeto, $\alpha \in \Sigma$ denota que α es un símbolo de Σ. Como ejemplo, si $\Sigma = \{1, 2, 3, 4\}$ podemos decir entonces que $2 \in \Sigma$. Además, como un alfabeto es matemáticamente un conjunto finito no vacío, dados Σ_1 y Σ_2 alfabetos, se tiene que $\Sigma_1 \cup \Sigma_2$ también lo es. No es difícil ver que si $\Sigma_1 \cap \Sigma_2, \Sigma_1 \setminus \Sigma_2$ y $\Sigma_2 \setminus \Sigma_1$ son conjuntos no vacíos, también son alfabetos.

Definición: Palabra

Se conoce como **Palabra** sobre un alfabeto Σ a una secuencia finita de símbolos de Σ

- Como el término palabra la asociamos a palabras de algún lenguaje natural, a menudo se usa el término **cadena** para hacer referencia a lo mismo.
- En base a nuestra definición, **printer, place** y **table** son cadenas sobre el alfabeto inglés, así como también **kndysc** y **hdjajs** lo son.
- Cada símbolo de un alfabeto constituye una palabra sobre tal alfabeto.
- La *cadena vacía*, denotada por ϵ, es una palabra sobre cualquier alfabeto.

Definición: Lenguaje

Se conoce como **Lenguaje** a un conjunto de palabras.

Por ejemplo, el conjunto $\{1, 12, 123\}$ es un lenguaje sobre el alfabeto compuesto por dígitos. Podemos observar también lo siguiente:

- Si Σ es un alfabeto, entonces también es un lenguaje por sí mismo (aquel formado por todas las cadenas con único símbolo).
- Los lenguajes pueden ser infinitos.
- El lenguaje compuesto por ninguna cadena es el *lenguaje vacío* denotado por $\emptyset$. Este lenguaje no es el mismo que aquel compuesto por $\{\epsilon\}$.
- El lenguaje compuesto por todas las cadenas sobre el alfabeto Σ, se conoce como cerradura de Σ o lenguaje universal sobre Σ y se denota por Σ^*.

Operaciones con Cadenas

Si α es una cadena sobre cualquier alfabeto, su longitud se denota mediante el símbolo $|\alpha|$. La longitudes es el número de símbolos que tiene la cadena.

Concatenación

Si w y z son cadenas, la concatenación de w con z es la cadena que se obtiene al añadir la cadena w a la palabra z. La concatenación de las palabras w y z se denota como wz o $w \cdot z$.

Podemos observar que

$$|wz| = |w| + |z|$$

La identidad para la operación de concatenación es ϵ.

Potencia de una cadena

Sea w una cadena; para $n \in \mathbb{N}$ se define

$$w^n = \begin{cases} \epsilon, & n = 0 \\ ww^{n-1}, & n \geq 0 \end{cases}$$

Se dice que w^i es la potencia i-ésima de w.

Igualdad de cadenas

Si w y z son palabras, se dice que w es igual a z, si tienen la misma longitud y los mismos símbolos en la misma posición. Se denota mediante $w = z$.

Las nociones de sufijo y prefijo de cadenas sobre un alfabeto son análogas a las que se usan habitualmente.

Prefijo

Si w y x son palabras, se dice que x es prefijo de w, si existe alguna cadena y tal que $w = xy$.

Sufijo

Si w y y son palabras, se dice que y es sufijo de w, si existe alguna cadena x tal que $w = xy$.

Utilizaremos también el término **prefijo (sufijo) propio** para denotar aquellas cadenas que son prefijos (sufijos) de una palabra pero no iguales a la misma. De lo anterior se deduce que la palabra vacía ϵ es prefijo y sufijo de cualquier palabra.

Por último, tenemos lo siguiente:

Subcadena

Una cadena w es una subcadena o subpalabra de otra cadena z si existen cadenas x e y tales que $z = xwy$.

Inversa de una cadena

La inversa o transpuesta de una palabra w se define como

$$w^I = \begin{cases} w, & w = \epsilon \\ z^I a, & w = az, a \in \Sigma \;\; y \;\; z \in \Sigma^* \end{cases}$$

En palabras simples, la inversa es la imagen refleja de una palabra.

- La inversa de la palabra **able** es **elba**.
- La inversa se deshace a sí misma $(x^I)^I = x$.

Operaciones con Lenguajes

Como un lenguaje está formado por palabras, es natural pensar en extender las operaciones anteriormente vistas a los lenguajes en sí mismo. Para hacerlo, debemos empezar de la siguiente manera. Sean A y B lenguajes sobre un alfabeto. Se define el lenguaje *concatenación* de A y B como

$$A \cdot B = \{w \cdot x | w \in A \quad y \quad x \in B\}$$

es decir, $A \cdot B$ está formado por todas las cadenas que se forman concatenando cada cadena A con todas las de B.

- El lenguaje concatenación $A \cdot B$ es un lenguaje sobre $\Sigma_1 \cup \Sigma_2$ y por tanto no es necesario que ambos lenguajes se formen sobre el mismo alfabeto.
- No es difícil ver que $A \cdot \{\epsilon\} = \{\epsilon\} \cdot A = A$. Es decir, el lenguaje cuyo único elemento es la palabra vacía, funciona como la identidad de la operación.
- La potencia de un lenguaje A se puede definir como

$$A^n = \begin{cases} \{\epsilon\}, & n = 0 \\ A \cdot A^{n-1}, & n \geq 1 \end{cases}$$

Puesto que un lenguaje es un conjunto de cadenas, se puede usar las operaciones de conjuntos para crear nuevos lenguajes. Si A y B son lenguajes sobre el alfabeto Σ, entonces la unión de A y B se denota mediante $A \cup B$ y está formada por todas las palabras que pertenecen al menos a uno de los dos lenguajes. Por tanto,

$$A \cup B = \{x | x \in A \quad o \quad x \in B\}$$

De la misma forma, la intersección de lenguajes se define como

$$A \cap B = \{x | x \in A \quad y \quad x \in B\}$$

es decir, está formado por las palabras que pertenecen simultáneamente a ambos lenguajes a la misma vez.

Sublenguajes

Si A y B son leguajes sobre un alfabeto Σ y si todas las cadenas de A son también cadenas de B entonces se dice que A es un sublenguaje de B. Se denota $A \subset B$.

De esto, podemos inferir que para toda lenguaje L tenemos que $L \subset \Sigma^*$. Al igual que en los conjuntos, la igualdad de dos lenguajes A y B se da si y solamente si $A \subset B$ y $B \subset A$.

Teorema: Distrubuitividad de la concatenación

Dados los lenguajes A, B y C sobre un alfabeto Σ, se cumple que:

$$A \cdot (B \cup C) = A \cdot B \cup A \cdot C$$

$$(B \cup C) \cdot A = B \cdot A \cup C \cdot A$$

Cerraduras

Si es un lenguaje sobre algún alfabeto Σ, se define la cerradura de Kleene o cerradura de estrella de un lenguaje A como

$$A^* = \bigcup_{n=0}^{\infty} A^n$$

Definiremos también la cerradura positiva de A como

$$A^+ = \bigcup_{n=1}^{\infty} A^n$$

Observaciones

- $A^n \subseteq \Sigma^* \quad \forall n \in \{0, 1, 2, ...\}$
- $A^+ \subseteq A^* \subseteq \Sigma^*$
- $\emptyset^0 = \{\epsilon\}$ y $\emptyset^n = \emptyset$ para todo $n \geq 1$.

Complemento

Definimos el **complemento** de un lenguaje A sobre el alfabeto Σ como

$$\bar{A} = \Sigma^* \setminus A$$

La concatenación y la diferencia no son compatibles debido a la intersección.

Teorema

$A^+ = A \cdot A^* = A^* \cdot A$

Además, no está demás nombrar que $(A^+)^+ = A^+$ y $(A^*)^+ = A^*$.