Tema 4: Aplicación de los autómatas: Lenguajes formales

1. Lenguajes formales.

Como se ha indicado en la introducción del tema, el concepto de autómata surgió cuando se modelizó matemáticamente el sistema nueronal humano. Sin embargo, se pueden encontrar conexiones con otras ramas, como por ejemplo los lenguajes formales.

Definición. Un **alfabeto** A es un conjunto finito cuyos elementos se denominan **letras**.

Definición. Un **lenguaje** Ł sobre el alfabeto A es un subconjunto de Ω_A. A los elementos de Ł se les denominan **palabras**.

Definición. Un lenguaje Ł sobre el alfabeto A se dice que es finito si $|Ł|$ es finito.

Definición. Llamaremos **lenguaje vacío** sobre el alfabeto A a $Ł_\emptyset = \emptyset$.

Debemos distinguir entre $Ł_\emptyset$ y $Ł_\Lambda = \{\Lambda\}$, esto es el lenguaje que contiene sólo la palabra vacía.

Como ya sabemos en Ω_A tenemos definida la operación concatenación. Por ello, cuando nos den una palabra $x \in \Omega_A$, podemos verla como la concatenación de varias palabras, alguna de ellas posiblemente vacías. Así, diremos que y_2 es una **subpalabra** de x si $x = y_1y_2y_3$, donde $y_i \in \Omega_A$. Observamos que y_1 ó y_3 pueden ser la palabra vacía. Si $y_1 = \Lambda$, diremos que y_2 es una subpalabra inicial y si $y_3 = \Lambda$, diremos que y_2 es una subpalabra final.

Ejemplos. $Ł_1 = \{\Lambda, a_1a_2, a_1, a_2a_1\}$ es un lenguaje sobre el alfabeto $\{a_1, a_2\}$. $Ł_2 = \{a_1^i | i \in \mathbb{N}\}$ es un lenguaje sobre el alfabeto $\{a_1\}$. $Ł_1$ es un lenguaje finito y $Ł_2$ no.

En el conjunto $\mathbf{L} = \{Ł | Ł \text{ es lenguaje sobre el alfabeto A}\}$, podemos definir las siguientes operaciones:

1. **Suma de dos lenguajes**: Dados $Ł_1, Ł_2 \in \mathbf{L}$,

$$Ł_1 + Ł_2 = \{x \in \Omega_A | x \in Ł_1 \vee x \in Ł_2\}.$$

2. **Intersección de dos lenguajes**: Dados $Ł_1, Ł_2 \in \mathbf{L}$,

$$Ł_1 \cap Ł_2 = \{x \in \Omega_A | x \in Ł_1 \wedge x \in Ł_2\}.$$

3. **Complementario de un lenguaje**: Dado $Ł_1 \in \mathbf{L}$,

$$Ł_1^c = \{x \in \Omega_A | x \notin Ł_1\}.$$

4. **Diferencia de dos lenguajes**: Dados $Ł_1, Ł_2 \in \mathbf{L}$,

$$Ł_1 - Ł_2 = \{x \in \Omega_A | x \in Ł_1 \wedge x \notin Ł_2\}.$$

   Es claro que $Ł_1 - Ł_2 = Ł_1 \cap Ł_2^c$.

5. **Concatenación ó producto de dos lenguajes** Dados $Ł_1, Ł_2 \in \mathbf{L}$,

$$Ł_1 Ł_2 = \{x \in \Omega_A | x = y_1 y_2, \quad y_i \in Ł_i,\ i = 1, 2\}.$$

   Es evidente que la concatenación de lenguajes es asociativo por serlo la concatenación de palabras. Además, $Ł_\emptyset$ y $Ł_\Lambda$ son elementos cero e identidad para la concatenación de lenguajes sobre el mismo alfabeto.

6. **Clausura de un lenguaje**: Dado $Ł_1 \in \mathbf{L}$,

$$Ł_1^* = \sum_{i=0}^{\infty} Ł_1^i,$$

   donde $Ł_1^i$ representa el producto de $Ł_1$ i veces y $Ł_1^0 = \{\Lambda\}$. Esto es, la clausura del lenguaje $Ł_1$ contiene todas las palabras de Ω_A que se pueden obtener como concatenación de palabras de $Ł_1$ más la palabra vacía.

2. Lenguajes regulares.

Introducimos ahora el concepto de expresión regular .

Definición. Una **expresión regular** sobre el alfabeto A es una palabra de $\Omega_{A \cup I}$, donde $I = \{+, *, \emptyset, (,)\}$, satisfaciendo las siguientes condiciones:

1. Cada letra de A y $\emptyset$ son expresiones regualares.

2. Si α y β son expresiones regulares sobre A, también lo son $(\alpha + \beta)$, $(\alpha\beta)$ y α^*.

3. Nada es expresión regular sobre A, salvo que se obtenga tras la aplicación de un número finito de veces de 1. y 2.

Ejemplo. $\emptyset^*$, $((a_1 + a_2)(a_1 + a_2))^*$ y $(a_1 + a_2^*)a_3$ son expresiones regulares sobre el alfabeto A, cuando $a_1, a_2, a_3 \in A$.

Definición. Sea α una expresión regular sobre el alfabeto A. se llama lenguaje asociado a α, y se denota por $|\alpha|$, al lenguaje de Ω_A que se obtiene de acuerdo a las siguientes condiciones:

1. $|\emptyset| = L_\emptyset$.

2. $\forall a \in A$, $|a| = \{a\}$.

3. $\forall \alpha, \beta$ expresiones regulares sobre A,

$$|(\alpha + \beta)| = |\alpha| + |\beta|, \qquad |(\alpha\beta)| = |\alpha||\beta|, \qquad |\alpha^*| = |\alpha|^*.$$

Ejemplo. $|\emptyset^*| = \{\Lambda\} = Ł_\Lambda$ es el lenguaje asociado a la expresión regular $\emptyset^*$. Por otro lado, el lenguaje asociado a la expresión regular a^*, donde $a \in A$, viene dado por $|a^*| = |a|^* = \{a^i | i \in \mathbb{N}\}$, entendiendo que $a^0 = \Lambda$.

Definición. Un lenguaje Ł sobre el alfabeto A se dice que es **regular** si existe una expresión regular α tal que $Ł = |\alpha|$.

Es fácil comprobar que todo lenguaje finito es regular. Puede suceder también que dos expresiones regulares distintas tengan asociado el mismo lenguaje regular. Por ejemplo, las expresiones regulares $(a_1 + a_2)^*$ y $(a_1^* a_2^*)^*$ tienen por lenguaje asociado a Ω_A, donde $A = \{a_1, a_2\}$.

3. Relacin entre los lenguajes regulares y los autmatas .

Definición. Sea S $= (S, E, \delta)$ un semiautómata con estado inicial e_1 y $E_1 \subseteq E$. Se llama **lenguaje representado por S respecto de** E_1 al conjunto

$$Ł(\mathrm{S}, E_1) = \{x \in \Omega_S | \hat{\delta}(e_1, x) \in E_1\}.$$

Es obvio que si $E_1 = \{e_{i1}, \ldots, e_{ir}\}$, entonces

$$Ł(\mathrm{S}, E_1) = \sum_{j=1}^{r} Ł(\mathrm{S}, e_{ij}).$$

Ejemplo. Consideramos el semiautómata S $= (S, E, \delta)$, donde $S = a_1, a_2$, $E = \{e_1, e_2, e_3\}$ y

$$\begin{array}{lccc} \delta : E \times S & \to & E & \\ & (e_1, a_1) & \mapsto & e_2 \\ & (e_1, a_2) & \mapsto & e_3 \\ & (e_2, a_1) & \mapsto & e_2 \\ & (e_2, a_3) & \mapsto & e_2 \\ & (e_3, a_1) & \mapsto & e_1 \\ & (e_3, a_2) & \mapsto & e_2 \end{array}$$

Entonces,

$$Ł(\mathrm{S}, \{e_1\}) = \{(a_2a_1)^i | i \in \mathbb{N} \cup \{0\}\}.$$

Observamos que el lenguaje del ejemplo anterior es un lenguaje regular, ya que es el lenguaje asociado a la expresión regular $(a_2a_1)^*$. En lo que sigue vamos a estudiar la relación existente entre los lenguajes regulares y los lenguajes asociados a semiautómatas.

Definición. Sea S $= (S, E, \delta)$ un semiautómata y $e_i, e_j \in E$. Diremos que de e_i se **pasa** a e_j si existe $x \in \Omega_S$ tal que $\hat{\delta}(e_i, x) = e_j$. Si $x = s_1 \dots s_r$, entonces a $\delta(e_i, s_1)$, $\hat{\delta}(e_i, s_1s_2)$, ..., $\hat{\delta}(e_i, s_1s_2s_{r-1})$ se les llama **estados intermedios** del paso de e_i a e_j mediante x.

Definición. Sea S $= (S, E, \delta)$ un semiautómata tal que $E = \{e_1, \dots, e_n\}$, y sean $i, j \in \{1, \dots, n\}$ y $k \in \{0, 1, \dots, n\}$. Se llama

$$Ł_{ij}^k = \{x \in \Omega_S | x = s_{i1} \dots s_{ir}, \hat{\delta}(e_i, x) = e_j \text{ y } \hat{\delta}(e_i, s_{i1}s_{it}) \notin E - \{e_1, \dots, e_k\}, t \in \{1, \dots, r-1\}\}$$

entendiendo que si $k = 0$ la última condición significa que no hay estados intermedios en el paso de e_i a e_j. Esto es,

$$Ł_{ij}^0 = \{x \in \Omega_S | \delta(e_i, x) = e_j \text{ y no hay estados intermedios}\}$$

y

$$Ł_{ij}^n = \{x \in \Omega_S | \delta(e_i, x) = e_j\}.$$

Un caso especial es cuando consideramos L_{1j}^n que coincide con $Ł(\mathrm{S}, e_j)$.

En lo que sigue vamos a probar que $Ł_{ij}^k$ es un lenguaje regular. En primer lugar relacionamos $Ł_{ij}^k$ con $Ł_{rs}^{k-1}$ para determinados índices r y s.

Lema 3.1. *Sea $S = (S, E, \delta)$ un semiautómata tal que $E = \{e_1, \dots, e_n\}$, y sean $i, j, k \in \{1, \dots, n\}$. Entonces,*

$$L_{ij}^k = L_{ij}^{k-1} + L_{ik}^{k-1}(L_{kk}^{k-1})^* L_{kj}^{k-1}.$$

Demostración. $L_{ij}^k \subseteq Ł_{ij}^{k-1} + Ł_{ik}^{k-1}(Ł_{kk}^{k-1})^* Ł_{kj}^{k-1}$. Sea $x \in L_{ij}^k$. Entonces, $\hat{\delta}(e_i, x) = e_j$ y los estados intermedios que aparecen son pertenecen al conjunto $\{e_1, \ldots, e_k\}$. Si e_k no es estado intermedio, entonces $x \in L_{ij}^{k-1}$. Si e_k es un estado intermedio, entonces podemos escribir $x = y_1 y_2 y_3$, donde $y_i \in \Omega_S$ y tales que

$$\begin{aligned} &\hat{\delta}(e_i, y_1) = e_k && \text{y } e_k \text{ no es estado intermedio} \\ &\hat{\delta}(e_k, y_2) = e_k && \text{y los estados intermedios pertenecen a } \{e_1, \ldots e_k\} \\ &\hat{\delta}(e_k, y_3) = e_j && \text{y } e_k \text{ no es estado intermedio.} \end{aligned}$$

Por tanto, $y_1 \in Ł_{ik}^{k-1}$ y $y_3 \in Ł_{kj}^{k-1}$. A su vez, descomponemos $y_2 = z_1 \ldots z_t$ de forma que

$$\hat{\delta}(e_k, z_r) = e_k \qquad \text{y } e_k \text{ no es estado intermedio para } r = 1, \ldots, t.$$

Entonces, $y_2 \in (Ł_{kk}^{k-1})^*$.

$Ł_{ij}^{k-1} + Ł_{ik}^{k-1}(Ł_{kk}^{k-1})^* Ł_{kj}^{k-1} \subseteq Ł_{ij}^k$. Sea $x \in Ł_{ij}^{k-1} + Ł_{ik}^{k-1}(Ł_{kk}^{k-1})^* Ł_{kj}^{k-1}$. Si $x \in Ł_{ij}^{k-1}$, entonces $x \in Ł_{ij}^k$ ya que $Ł_{ij}^{k-1} \subseteq Ł_{ij}^k$. Si $x \in Ł_{ik}^{k-1}(Ł_{kk}^{k-1})^* Ł_{kj}^{k-1}$, entonces podemos descomponer $x = y_1 y_2 y_3$, donde $y_1 \in Ł_{ik}^{k-1}$, $y_2 \in (Ł_{kk}^{k-1})^*$ y $y_3 \in Ł_{kj}^k$. Pero entonces,

$$\hat{\delta}(e_i, x) = \hat{\delta}(\hat{\delta}(\hat{\delta}(e_i, y_1), y_2), y_3) = \hat{\delta}(\hat{\delta}(e_k, y_2), y_3) = \hat{\delta}(e_k, y_3) = e_j$$

y los estados intermedios que aparecen cuando calculamos $\hat{\delta}(e_i, y_1)$ pertenecen al conjunto $\{e_1, \ldots, e_{k-1}\}$, $\hat{\delta}(e_k, y_2)$ pertenecen al conjunto $\{e_1, \ldots, e_k\}$ y $\hat{\delta}(e_k, y_3)$ pertenecen al conjunto $\{e_1, \ldots, e_{k-1}\}$. Por tanto, $x \in Ł_{ij}^k$. □

Proposición 3.2. *Sea $S = (S, E, \delta)$ un semiautómata tal que $E = \{e_1, \ldots, e_n\}$, y sean $i, j \in \{1, \ldots, n\}$ y $k \in \{0, \ldots, n\}$. Entonces, $Ł_{ij}^k$ es un lenguaje regular.*

Demostración. (Por inducción sobre k) Si $k = 0$, entonces $Ł_{ij}^0 \subseteq S \cup \{\Lambda\}$ ó $Ł_{ij}^0 = \emptyset$ y en cualquier caso es un lenguaje regular.

Supongamos que el resultado es cierto para $k < t$ y veamoslo para $k = t$. Entonces, por el lema anterior sabemos que

$$Ł_{ij}^t = Ł_{ij}^{t-1} + Ł_{it}^{t-1}(Ł_{tt}^{t-1})^* Ł_{tj}^{t-1}$$

y empleando inducción, tenemos que $Ł_{ij}^{t-1}$, $Ł_{it}^{t-1}$, $Ł_{tt}^{t-1}$ y $Ł_{tj}^{t-1}$ son lenguajes regulares. Como la suma, concatenación y clausura de lenguajes regulares es regular, se sigue que $Ł_{ij}^t$ es regular. □

Como Corolario de esta proposición, tenemos

Corolario 3.3. *Sea $S = (S, E, \delta)$ un semiautómata tal que $E = \{e_1, \ldots, e_n\}$. Entonces, $L(S, e_j$ es un lenguaje regular para todo $j \in \{1, \ldots, n\}$.*

Corolario 3.4. *Sea $S = (S, E, \delta)$ un semiautómata tal que $E = \{e_1, \ldots, e_n\}$ y $E_1 \subseteq E$. Entonces, $L(S, E_1)$ es un lenguaje regular.*

Ahora nos falta ver que a partir de un lenguaje regular se puede construir un semiautómata y un subconjunto E_1 de forma que el lenguaje asociado al semiautómata respecto de E_1. En primer lugar, vamos a dar un ejemplo del que extraeremos la estrategia para probar el caso general.

Ejemplo. Sea L el lenguaje regular cuya expresión regular viene dada por $a(a^* + b)^* + abb^*ab^*$, esto es,

$$L = |a(a^* + b)^* + abb^*ab^*|.$$

Paso 1. Transformamos la expresión regular que define L en otra en la que las letras que lo forman aparezcan subindicadas por la posición que ocupan

$$a(a^* + b)^* + abb^*ab^* \Rightarrow a_1(a_2^* + b_3)^* + a_4b_5b_6^*a_7b_8^*$$

A c_i le llamaremos **descendiente** de c, siendo c una de las letras que parece en la expresión regular. Al lenguaje que tiene por expresión regular a la nueva lo denotamos por L_1, esto es,

$$L_1 = |a_1(a_2^* + b_3)^* + a_4b_5b_6^*a_7b_8^*|$$

Paso 2. Localizamos

(i) Pares de letras que aparezcan consecutivas en palabras de L_1. En nuestro caso,

$$\begin{aligned} P = &\{(a_1, a_2), (a_1, b_3), (a_2, a_2), (a_2, b_3), (b_3, a_2), (b_3, b_3), \\ &(b_4, a_5), (a_5, b_6), (b_6, b_6), (a_5, a_7), (b_6, a_7), (a_7, b_8), (b_8, b_8)\}. \end{aligned}$$

(ii) Letras de inicio de palabra. En nuestro caso,

$$I = \{a_1, b_4\}.$$

(iii) Letras de final de palabra. En nuestro caso,

$$F = \{a_1, a_2, b_3, a_7, b_8\}.$$

Paso 3. Se define el semiautómata S $= (S, E, \delta)$, donde S es el albabeto del lenguaje L y la función δ y E se definen como sigue: Dado $c \in S$,

$$\delta(e_1, c) = \{\text{descendientes de } c \text{ que satisfacen 2(ii) }\} \qquad \forall c \in S$$

$$\delta(e_i, c) = \{\text{descendientes de } c \text{ que siguen a alguna letra que aparece en } e_i\} \quad \forall i > 1.$$

En nuestro caso,

$$S = \{a, b\},$$

$$\begin{array}{ll}
\delta(e_1, a) = \{a_1\} = e_2 & \delta(e_1, b) = \{b_4\} = e_3 \\
\delta(e_2, a) = \{a_2\} = e_4 & \delta(e_2, b) = \{b_3\} = e_5 \\
\delta(e_3, a) = \{a_5\} = e_6 & \delta(e_3, b) = \emptyset = e_7 \\
\delta(e_4, a) = \{a_2\} = e_4 & \delta(e_4, b) = \{b_3\} = e_5 \\
\delta(e_5, a) = \{a_2\} = e_4 & \delta(e_5, b) = \{b_3\} = e_5 \\
\delta(e_6, a) = \{a_7\} = e_8 & \delta(e_6, b) = \{b_6\} = e_9 \\
\delta(e_7, a) = \emptyset = e_7 & \delta(e_7, b) = \emptyset = e_7 \\
\delta(e_8, a) = \emptyset = e_7 & \delta(e_7, b) = \{b_8\} = e_{10} \\
\delta(e_9, a) = \{a_7\} = e_8 & \delta(e_9, b) = \{b_6\} = e_9 \\
\delta(e_{10}, a) = \emptyset = e_7 & \delta(e_{10}, b) = \{b_8\} = e_{10}
\end{array}$$

y

$$E = \{e_1, \ldots, e_{10}\}.$$

Paso 4. Se toma como E_1 al subconjunto de estados de E tales que en su definición contienen al menos una letra final de palabra de L_1. En nuestro caso,

$$E_1 = \{e_2, e_4, e_5, e_8, e_{10}\}.$$

Siguiendo estos mismos pasos, se prueba

Proposición 3.5. *Sea L un lenguaje regular. Entonces, existe $S = (S, E, \delta)$ y $E_1 \subseteq E$ tal que*

$$L = L(S, E_1).$$