Conceptos de lenguajes de programación y jerarquía de Chomsky

Pedro O. Pérez M., PhD.

Implementación de métodos computacionales
Tecnológico de Monterrey

pperezm@tec.mx

02-2023

¿Qué es un lenguaje?

Alfabeto

- Un alfabeto es un conjunto finito, no vacío de símbolos. Por convención, usaremos el símbolo Σ para representar este conjunto.
- Debido a que los elementos del alfabeto son indivisible, generalmente los denotaremos con caracteres únicos. Los letras a, b, c, d, e, con o sin subíndices, se usan para representar los elementos de un cierto alfabeto.
- Entre los ejemplos más comunes de alfabetos podemos encontrar:
  1. $\Sigma = 0, 1$, alfabeto binario.
  2. $\Sigma = a, b, c, ..., z$, el conjunto de todas las letras minúsculas.
  3. El conjunto de todos los caracteres ASCII, o el conjunto de todos los caracteres imprimibles ASCII.

Cadenas (strings)

- Una cadena (o palabras) es una secuencia finita de símbolos elegidos de algún alfabeto. Por ejemplo, 01101 es una cadena del alfabeto binarios $\Sigma = 0, 1$.
- Por convención, para representar sobre un alfabeto emplearemos letras que aparecen cerca del final del alfabeto. En particular, p, q, u, v, w, x, y, z se utilizan para indicar cadenas.
- La cadena vacía, ε, es una cadena con cero ocurrencia de símbolos y, por lo mismo, puede ser elegido de cualquier alfabeto.
- La longitud de una cadena es el número de posiciones para símbolos dentro de la misma. Por ejemplo, 01101 tiene una longitud de 5. La notación estándar para la longitud de una cadena w es $|w|$. Por ejemplo, $|011| = 3$ y $|\varepsilon| = 0|$.

Definición 1

Sea Σ un alfabeto. Σ^*, el conjunto de cadenas sobre Σ, se define recursivamente de la siguiente manera:

1. Base: $\varepsilon \in \Sigma^*$.
2. Recursivo: Si $w \in \Sigma^*$, entonces $wa \in \Sigma^*$.
3. Cerradura: $w \in \Sigma^*$ si y solo sí puede ser obtenido a partir de ε mediante un número finito de aplicaciones del paso recursivo.

Por ejemplo, sea $\Sigma = a, b, c$. Los elementos de Σ^* incluyen:

- Longitud 0: ε
- Longitud 1: a, b, c
- Longitud 2: aa, ab, ac, ba, bb, bc, ca, cb, cc.

Definición 2
Un lenguaje sobre un alfabeto Σ es un subconjunto de Σ^*.

Definición 3
Sea u, $v \in \Sigma$. La concatenación de u y v, escrita uv, es una operación binaria definida sobre Σ^* como sigue:

(i) Base: Si la $length(v) = 0$, entonces $v = \varepsilon$ y $uv = u$.

(ii) Recursivo: Sea v una cadena con $length(v) = \mathrm{n} > 0$. Entonces, $v = wa$, para alguna cadena w con longitud $n - 1$ y $a \in Sigma$, $uv = (uw)a$.

Por ejemplo, $u = ab$, $v = ca$, y $w = bb$. Entonces,

- $uv = abca$
- $vw = cabb$
- $(uv)w = abcabb$, $u(vw) = abcabb$

Definición 4
Sea u, v, $w \in \Sigma$. Entonces, $(uv)w = u(vw)$.

Los exponentes son usados para abreviar la concatenación de una cadena consigo misma. Así, uu se puede abreviar u^2, uuu puede escribirse u^3 y así sucesivamente. Para completar, u^0, representa la concatenación de u consigo mismo cero veces, lo que es ε. La operación de concatenación no es conmutativa. Por ejemplo, $u = ab$ y $v = ba$, $uv = abba$ y $vu = baab$.

- Si Σ es un alfabeto, podemos expresar el conjunto de todas las cadenas de una cierta longitud usando la notación exponencial. Con esto, podemos definir Σ^k como el conjunto de cadenas de longitud k, seleccionados de los símbolos de Σ.
- Es importante hacer notar que $\Sigma^0 = \{\varepsilon\}$, sin importar a qué alfabeto nos referimos.
- De tal forma que si $\Sigma = \{0, 1\}$, entonces:
  - $\Sigma^1 = \{0, 1\}$,
  - $\Sigma^2 = \{00, 01, 10, 11\}$,
  - $\Sigma^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$

Las subcadenas se pueden definir mediante la operación de concatenación. Intuitivamente, u es una subcadena de v, si u "ocurre dentro de" v. Formalmente, u es una subcadena de v si hay cadenas x y y tales que $v = xuy$. Un prefijo de v es una subcadena u en la que x es la cadena vacía en la descomposición de v, Es decir, $v = uy$. De manera similar, u es sufijo de v si $u = xu$.

Definición 5
Sea u una cadena Σ^*. La inversión, denotada u^R, se define como sigue:

1. Base: Si la $length(u) = 0$, entonces $u = \varepsilon$ y $\varepsilon^R = \varepsilon$.
2. Recursivo: Si $length(u) = n > 0$, entonces $u = wa$ para alguna cadena w con longitud $n - 1$ y alguna $a \in \Sigma$, y $u^R = aw^R$.

Definición 6
Sean u, $v \in \Sigma^*$. Entonces, $(uv)^R = u^R v^R$.

- Un conjunto de cadenas, todas la cuales son seleccionadas de algún Σ^*, donde Σ es un alfabeto particular, es llamada lenguaje. Dicho de otra forma, si Σ es un alfabeto, y $L \subseteq \Sigma^*$, entonces L es un lenguaje sobre Σ. Por ejemplo,
  - El lenguaje Español, la colección legal de palabras españolas, es un conjunto de cadenas sobre el alfabeto que consiste de todas las letras.
  - El lenguaje C, o cualquier otro lenguaje de programación, es el conjunto de programas válidos que son un subconjunto de todas las posibles cadenas que pueden ser formadas sobre el conjunto de los caracteres ASCII.
- También podemos definir otros lenguajes:
  - El lenguaje de todas las cadenas consistentes de n 0's seguidos de n 1's para alguna $n \geq 0$: $\{\varepsilon, 01, 0011, 000111, ...\}$.

- También podemos definir otros lenguajes:
  - El lenguaje de todas las cadenas consistentes de n 0's seguidos de n 1's para alguna $n \geq 0$: $\{\varepsilon, 01, 0011, 000111, ...\}$.
  - El lenguaje de todas las cadenas que tiene un igual número de 0's y 1's: $\{\varepsilon, 01, 10, 0011, 1100, 0101, 1010, 1001, ...\}$.
  - El conjunto de números binarios cuyo valor es primo: $\{10, 11, 101, 111, 1011, ...\}$.
  - Σ^* es un lenguaje para cualquier alfabeto Σ.
  - $\emptyset$, el lenguaje vacío, es un lenguaje sobre cualquier alfabeto.
  - $\{\varepsilon\}$, lenguaje conformado por solo una cadena vacía, es, también, un lenguaje sobre cualquier alfabeto. Importante, $\emptyset \neq \{\varepsilon\}$.

Sean X un conjunto. Entonces,

$$X^* = \bigcup_{i=0}^{\infty} X^i$$

$$X^+ = \bigcup_{i=1}^{\infty} X^i$$

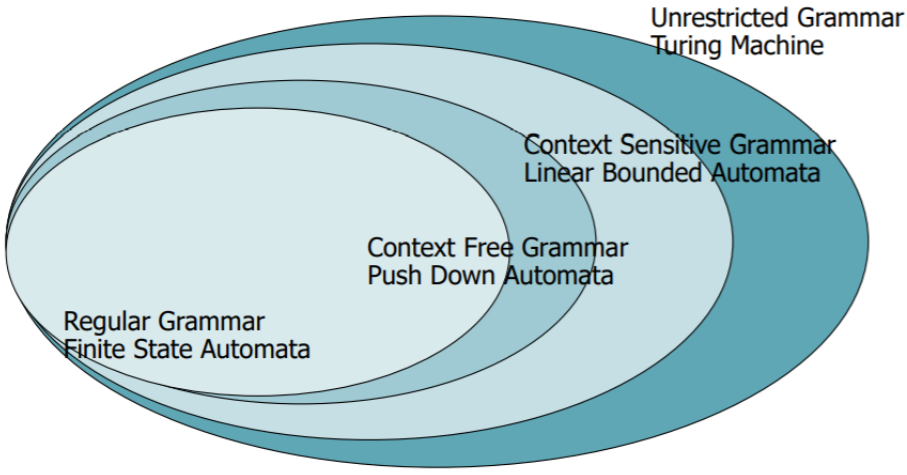
Ejemplo

- Define el lenguaje L sobre el alfabeto $\Sigma = a, b$ en el que cada cadena comience con una a y tengan longitud par.

- Define el lenguaje L sobre el alfabeto $\Sigma = a, b$ en el que cada cadena comience con una a y tengan longitud par.

Jerarquía de Chomsky-Schützenberger

- La jerarquía de Chomsky-Schützenberger es la base formal para describir un lenguaje (natural o artificial). “How Complex is Natural Language? The Chomsky Hierarchy”.


Unrestricted Grammar
Turing Machine
Context Sensitive Grammar
Linear Bounded Automata
Context Free Grammar
Push Down Automata
Regular Grammar
Finite State Automata

Definición de gramática

Una grámatica formal es un cuadrupla $G = (N, \Sigma, P, S)$ donde

- N es un conjunto finito de No Terminales.
- Σ es un conjunto finito de terminales y es disjunto de N.
- P es un conjunto finito de reglas de producción de la forma $w \epsilon (N \cup \Sigma)* \rightarrow w \epsilon (N \cup \Sigma)*$.
- $S \epsilon N$ es el símbolo inicial.

- Los lenguajes definidos por gramáticas Tipo-3 son aceptadas por Autómatas de Estados Finitos.
- La mayoría de la sintaxis de dialogo hablado informal.
- Las reglas son de la forma: $A \rightarrow \alpha$, $A \rightarrow \varepsilon$, $A \rightarrow \alpha B$ donde A, $B \epsilon N$, $\alpha \epsilon \Sigma$.

- Los lenguajes definidos por gramáticas Tipo-2 son aceptadas por Autómatas Push-Down.
- El lenguaje natural es casi enteramente definible por árboles sintácticos de Tipo-2.
- Las reglas son de la forma: $A \rightarrow \alpha$ donde $A \epsilon N$, $\alpha \epsilon (N \cup \Sigma)*$.

Tipo-1, Sensibles al contexto

- Los lenguajes definidos por gramáticas Tipo-1 son aceptadas por Autómatas Delimitados Linealmente.
- Sintaxis de algunos lenguajes naturales (Alemán).
- Las reglas son de la forma: $\alpha A \beta \rightarrow \alpha B \beta$, $S \rightarrow \varepsilon$ donde A y $S \epsilon N$, α, β, $B \epsilon (N \cup \Sigma)*$ y $B \neq \varepsilon$.

Tipo-0, Sin restricciones

- Los lenguajes definidos por gramáticas Tipo-0 son aceptadas por Maquinas de Turing.
- Las reglas son de la forma: $\alpha \rightarrow \beta$, donde α y β son cadenas arbitrarias sobre N y $\alpha \neq \varepsilon$.
- Ejemplos de producciones: Sab $\rightarrow$ ba, A $\rightarrow$ S.

Si quieres conocer un poco más, revisa estos dos vídeos:

- "TYPES OF GRAMMAR- Type 0, Type 1, Type 2, Type 3 (CHOMSKY HIERARCHY)|| THEORY OF COMPUTATION".
- "Noam Chomsky, Fundamental Issues in Linguistics (April 2019 at MIT) - Lecture 1".