Las limitaciones de los modelos finitos de gramáticas para la definición de lenguajes

Esteban di Tada•

Resumen

El uso de reglas de reescritura para el estudio de lenguajes fue introducido por Axel Thue[1] y posteriormente profundizado por Avram Noam Chomsky[2] que creó el concepto de gramática generativa y estableció las cuatro jerarquías conocidas como *Jerarquías de Chomsky* en base a su capacidad de generación de lenguajes. El problema que planteamos aquí es el de demostrar que el modelo de gramática finita solo puede representar una cantidad muy reducida de lenguajes

Palabras clave: Lenguajes formales, gramáticas formales, Jerarquias de Chomsky

Abstract

The use of rewritten rules in the study of languages was used for the first time by Axel Thue and then formalized by Noam Chomsky who introduced the concept of generative grammars. He defined also the 4 hierarchies known as Chomsky Hierarchies based on the power to generate languages. The problem we will study here is to answer the question if with finite models defined over a given alphabet it is possible to represent all the possible languages that exist on this alphabet

Keywords: Formal Languages, formal grammars, Chomsky Hierarchies

Fecha de Recepción: octubre 2012 | Fecha de aceptación: noviembre 2012

• Universidad de Palermo, Decano de la Facultad de Ingeniería

1. Axel Thue, Matemático noruego nacido en 1863 y fallecido en 1922.

2. Noam Chomsky, lingüista norteamericano perfeccionador de las ideas originales de Thue, nacido en Filadelfia, EEUU en 1928

1. Introducción

1.1. Definiciones previas

- Se definirá como alfabeto a un conjunto finito de símbolos $\Sigma=\{\boldsymbol{\sigma 1}, \boldsymbol{\sigma 2}...\boldsymbol{\sigma n}\}$
- Se definirá como cadena a una n-tupla ordenada de símbolos pertenecientes a un alfabeto y se lo designará por medio de letras minúsculas griegas. Si por ejemplo $\boldsymbol{\Sigma}=\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{1}, \boldsymbol{4}\}$, $\boldsymbol{\delta}=\boldsymbol{ab4a}$ será una cadena definida sobre el alfabeto $\boldsymbol{\Sigma}$
- Se definirá la operación concatenación de dos cadena α, β a una nueva cadena γ que se forma agregando a continuación de los símbolos de α los símbolos de β. Si α=cd5r y β=r5a entonces será γ= cd5rr5a. Se empleara la notación de la potenciación para representar la operación de multiplicar una cadena n veces por sí misma. Así $\mathbf{a^5=aaaaa}$
- Se definirá como longitud de una cadena a cantidad de símbolos por la que está compuesta. La longitud de la cadena δ del ejemplo del párrafo anterior será 4 y se la denotará $|\lambda|=4$
- Se definirá como Σ^n al conjunto de cadenas φ tal que $\varphi\in\Sigma^n\Leftrightarrow|\varphi|=n$. El conjunto Σ^0 contendrá un solo elemento que será la cadena λ que no contiene ningún símbolo y por definición será $|\lambda|=0$.
- El conjunto de todas las cadenas se definirá como $\boldsymbol{\Sigma^*=\cup\Sigma^n\ para\ n\geqq 0}$ y $\boldsymbol{\Sigma^+=\cup\Sigma^n\ para\ n\geqq 1}$
- Funciones sobre cadenas. Sean los conjuntos A y B y sea f: A⟶B. Sea α una cadena de símbolos pertenecientes a A. La función $f:A^+\dashrightarrow B^+$ se definirá recursivamente de la siguiente manera para x ∈ A y $\alpha\in A^*$

$$f(x\alpha)=\begin{cases} f(x) & si\ |\alpha|=0 \\ f(x)f(\alpha) & si\ |\alpha|>0 \end{cases}$$

- Dado el conjunto arbitrario A se definirá el conjunto $\boldsymbol{2^A}$ al conjunto de todos los conjuntos que se pueden formar con los elementos de A incluyendo al conjunto vació $\boldsymbol{\emptyset}$.
- Dado el alfabeto $\boldsymbol{\Sigma}$ se dirá que el subconjunto $\boldsymbol{\mathcal{L}\subseteq 2^{\Sigma^*}}$ es un lenguaje definido sobre $\boldsymbol{\Sigma}$. La cardinalidad de $\boldsymbol{\Sigma^*}$ denotada $\boldsymbol{|\Sigma^*|=\aleph_0}$ es decir que el conjunto es equipotente al de los naturales. Por el teorema de Cantor la cardinalidad de $\boldsymbol{2^{\Sigma^*}}$ será $\boldsymbol{\aleph_1}$ es decir la cardinalidad del continuo.

1.2. Interpretación intuitiva

Dado que la cardinalidad del conjunto de lenguajes definidos sobre el alfabeto $\boldsymbol{\Sigma}$ es $\aleph_1$, es decir la cardinalidad del continuo, ningún conjunto de modelos enumerable podrá contemplar la totalidad de los lenguajes posibles sobre Σ ya que si así lo fuera ambos conjuntos serian equipotentes por lo que existiría una función biyectiva h entre ambos. Esta función biyectiva mapearía el conjunto de los naturales en los reales lo que es una contradicción. Por lo tanto hay muchos más lenguajes que modelos enumerables. El objetivo del presente trabajo es demostrar formalmente esto para una clase general de gramáticas.

2. Definición de gramática

Se definirá una gramática como $\boldsymbol{G=\langle \Sigma, N, P, S\rangle}$ en donde

- Σ es un conjunto finito de símbolos terminales $\boldsymbol{\sigma_k\ k=1,...,n}$
- N es un conjunto de símbolos no terminales $\boldsymbol{\eta_j\ j=1,...,m}$
- P es una relación llamada de producciones y definida de la siguiente manera: $\boldsymbol{P\subseteq(\Sigma\cup N)^*N(\Sigma\cup N)^*\times(\Sigma\cup N)^*}$. Nótese que la parte de la izquierda de la relación debe contener obligatoriamente al menos un símbolo no terminal en tanto que la parte derecha puede ser la palabra vacía $\boldsymbol{\gamma}$. Si $\boldsymbol{\alpha\in(\Sigma\cup N)^*N(\Sigma\cup N)^*}$ y $\boldsymbol{\beta\in(\Sigma\cup N)^*}$ y $\boldsymbol{\alpha, \beta\in P}$ esto se denotara $\boldsymbol{\alpha\xrightarrow{G}\beta}$ que se puede expresar enunciando que en la gramática G la subcadena $\boldsymbol{\alpha}$ puede ser reemplazada por la cadena $\boldsymbol{\beta}$. Cuando no pueda haber confusión se puede omitir la referencia a la gramática ($\boldsymbol{\alpha\rightarrow\beta}$). En lenguaje corriente la formalización anterior podría expresarse diciendo que una producción es una regla de reescritura que permite reemplazar una cadena formada por símbolos terminales y no terminales pero que al menos contiene un no terminal por una cadena formada por símbolos terminales y no terminales (que puede ser la palabra vacía λ).
- $\boldsymbol{S\in N}$ es el símbolo inicial.

3. Comportamiento de la Gramática

Se dirá que dadas dos cadenas $\boldsymbol{\propto,\beta\in(\Sigma\cup N)^*}$ $\boldsymbol{\beta}$ es derivable de $\boldsymbol{\alpha}$ y que se denotará $\alpha\longrightarrow\beta$si y sólo si

- $\boldsymbol{\alpha=\gamma\zeta\omega}$ donde $\boldsymbol{\gamma, \omega\in(\Sigma\cup N)^*}$ y $\boldsymbol{\zeta\in(\Sigma\cup N)^*N(\Sigma\cup N)^*}$ y
- $\boldsymbol{\beta=\gamma\eta\omega}$ y
- $\boldsymbol{\zeta\rightarrow\eta\in N}$

Esto se denota $\alpha \overset{G}{\Rightarrow} \beta$. La referencia a la gramática puede omitirse cuando ello no pueda generar una confusión.

Se definirá la relación $\alpha \overset{n}{\Rightarrow} \beta$ si y solo si existen $\gamma_0, \gamma_1, \ldots \gamma_n$ tal que $\gamma_{j-1} \rightarrow \gamma_j \; para \; j=1,..n$.

Se dirá que $\alpha \overset{*}{\Rightarrow} \beta$ si existe un m finito tal que $\alpha \overset{m}{\Rightarrow} \beta$

Ahora se está en condiciones de definir el lenguaje generado por una gramática. Dada la gramática

$$G = \langle \Sigma, N, P, S \rangle$$

el lenguaje L_G generado por G se definirá como el conjunto de cadenas α tal que

$$\alpha \in L_G \; si \; y \; solo \; si \; S \overset{*}{\Rightarrow} \alpha \; y \; \alpha \in \Sigma^*$$

Esta definición puede expresarse diciendo que

El lenguaje generado por una gramática está constituido por todas las cadenas de símbolos terminales que se derivan del símbolo inicial.

Como en los casos anteriores se puede obviar la mención de la gramática cuando esto no pueda traer indefiniciones.

Resulta conveniente hacer dos observaciones:

- Primero que la definición no garantiza que exista al menos una cadena perteneciente al lenguaje es decir que L=∅ y segundo
- Que siempre de acuerdo con la definición de la palabra nula pertenecerá al lenguaje y que esta podría ser su único elemento, es decir L={λ}. Para evitar esta posibilidad habría que modificar la definición anterior de lenguaje generado por G de la siguiente manera $\alpha \in L_G \; si \; y \; solo \; si \; S \overset{*}{\Rightarrow} \alpha \; y \; \alpha \in \Sigma^+$.

Resulta conveniente ilustrar las definiciones con un ejemplo. Sea la gramática $G = \langle \Sigma, N, P, S \rangle$ donde

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

$N = \{Q\}$

$P = \{(Q, aQa), (Q, b)\}$, $o \; en \; el \; otro \; formato \; P = \{Q \rightarrow aQa, Q \rightarrow b\}$

S=Q

Esta gramática genera todas las cadenas del tipo $a^n b \; a^n$ para n⩾0

4. Gramáticas isomorfas

Sean las gramática $\boldsymbol{G}=\langle \boldsymbol{\Sigma}, \boldsymbol{N},\boldsymbol{P},\boldsymbol{S}\rangle$ y los conjuntos T y U equipotentes con Σ y P respectivamente y las funciones biyectivas h:Σ↔T y g:P↔U. Si se define la función biyectiva f:Σ∪T↔$N \cup U$

$$f(x)=\begin{Bmatrix} x \in \Sigma: h(x) \\ x \in \mathrm{N}: g(x) \end{Bmatrix}$$

Cuya función inversa es (dado que h y g son biyectivas)

$$f^{-1}(x)=\begin{Bmatrix} x \in \mathrm{T}: h^{-1}(x) \\ x \in U: g^{-1}(x) \end{Bmatrix}$$

Se puede crear una nueva gramática $\boldsymbol{H}=\langle \mathbf{T}, \boldsymbol{U},\boldsymbol{Q},\boldsymbol{R}\rangle=\langle \boldsymbol{f}(\mathbf{T}),\boldsymbol{f}(\boldsymbol{U}),\boldsymbol{f}(\boldsymbol{Q}),\boldsymbol{f}(\boldsymbol{R})\rangle$

Por razones de simplificación se denotará H=f(G)

Hay una función, que será definida como canónica, que jugara un especial rol en las demostraciones que siguen. Sea la gramática $\boldsymbol{G}=\langle \boldsymbol{\Sigma}, \boldsymbol{N},\boldsymbol{P},\boldsymbol{S}\rangle$ en donde Σ= $\{\sigma_1, \sigma_2, ..,\sigma_n\}$ y N= $\{\eta_1, \eta_2, ..,\eta_m\}$ la función canónica será

$$c(\sigma j)=j$$
$$c(\eta_k)=n+k$$

Y la función inversa será par $x \in (\sum \cup N)$

$$\mathrm{c}^{-1}(i)=\begin{Bmatrix} \sigma_i \, si \; 0<i\leq n \\ \eta_{i-n} \, si \; n<i\leq n+m) \end{Bmatrix}$$

De esta manera los símbolos terminales y no terminales son reemplazados por números naturales lo que facilitara la aplicación de métodos similares a los empleados en la demostración del teorema de Göedel.

Sean la gramática $\boldsymbol{G}=\langle \boldsymbol{\Sigma}, \boldsymbol{N},\boldsymbol{P},\boldsymbol{S}\rangle$ dos cadenas $\propto,\boldsymbol{\beta} \in (\boldsymbol{\Sigma} \cup \boldsymbol{N})^*$ tal que $\alpha \longrightarrow \beta$ luego se verificará por definición lo siguiente:

- $\alpha=\gamma\zeta\omega$ donde $\gamma, \omega \in (\boldsymbol{\Sigma} \cup \boldsymbol{N})^*$ y $\boldsymbol{\zeta} \in (\boldsymbol{\Sigma} \cup \boldsymbol{N})^* \boldsymbol{N} (\boldsymbol{\Sigma} \cup \boldsymbol{N})^*$ y
- $\boldsymbol{\beta}=\boldsymbol{\gamma\eta\omega}$ y
- $\zeta \longrightarrow \eta \in N$

Por lo tanto se verificará que $c(\alpha) \longrightarrow c(\beta)$ dado que:

- $\boldsymbol{c(\alpha)=c(\gamma)c(\zeta)c(\omega)}$ donde c($\boldsymbol{\gamma}$),($\boldsymbol{\omega}$)$\in\boldsymbol{c}((\boldsymbol{\Sigma}\cup\boldsymbol{N})^*)$ y $\boldsymbol{c(\zeta)}\in\boldsymbol{c}((\boldsymbol{\Sigma}\cup\boldsymbol{N})^*\boldsymbol{N}(\boldsymbol{\Sigma}\cup\boldsymbol{N})^*)$ y
- $\boldsymbol{c(\beta)=c(\gamma)c(\eta)c(\omega)}$ y
- $\boldsymbol{c(\zeta)\rightarrow c(\eta)}\in\boldsymbol{c(N)}$

Dado que c es una función biyectiva se puede demostrar entonces que $x\in L_G$ si y solo si $c(x)\in c(L_G)$.

Se definirá la gramática $c(L_G)$ como la forma canónica de la gramática G

En otras palabras los símbolos terminales y no terminales fueros reemplazados por números enteros mediante la función c y las cadenas generadas por esta nueva gramática (que serán cadenas de naturales) restituyen la cadena de G mediante la aplicación de c^{-1}. La cardinalidad de L_G y $L_{c(G)}$ es la misma.

5. Teorema de Schröder-Bernstein-Cantor

Un teorema fundamental para la teoría de los números transfinitos de Cantor es el demostrado por Schröder, Bernstein y Cantor que establece que dados dos conjuntos A y B y dos funciones sobreyectivas f:A⟶B y g:B⟶A entonces existe una función biyectiva h:A↔B. El teorema resulta bastante evidente para conjuntos finitos pero no lo es así para conjuntos no finitos.

No se incluirá aquí la demostración de este teorema pero para una mejor comprensión se presentará un ejemplo. Se desea demostrar que existe una función biyectiva h:[0,1]↔[0,1) – entre el intervalo cerrado 0≤ x ≤1 y el intervalo semiabierto 0≤ x <1. Es fácil construir las funciones sobreyectivas f y g de la siguiente manera:

$$f(x)=x/2\colon [0,1]\rightarrow[0,1) \text{ y } g(x)=x\colon [0,1)\rightarrow [0,1]$$

La construcción de la función biyectiva se hará en forma iterativa. Si se elige la función $f_0(x)=x$ se ve que no satisface las condiciones requeridas para ser una función ya que $f_0(1)=1$ no pertenece a [0,1) que, por definición, no contiene el punto 1. Para resolver este problema se modifica la función de la siguiente manera:

$$f_1(x)=\begin{cases} x & \Leftrightarrow x\neq 1 \\ 1/2 & \Leftrightarrow x=1 \end{cases}$$

Con esta solución queda resuelto el problema anterior pero esta función presenta ahora el problema que tanto para x =1 como para x=1/2 la función vale ½. Para solucionar este problema se modifica la definición de la siguiente manera

$$f_2(x) = \begin{Bmatrix} x & \Leftrightarrow x \notin \{1, 1/2\} \\ 1/2 & \Leftrightarrow x \in \{1\} \\ 1/4 & \Leftrightarrow x \in \{1/2\} \end{Bmatrix} = \begin{Bmatrix} x & \Leftrightarrow x \notin \{1, 1/2\} \\ x/2 & \Leftrightarrow x \in \{1, 1/2\} \end{Bmatrix}$$

El problema ahora se trasladó al punto x = 1/4. Para resolverlo defínase la nueva función

$$f_3(x) = \begin{Bmatrix} x & \Leftrightarrow x \notin \{1, 1/2, 1/4\} \\ x/2 & \Leftrightarrow x \in \{1, 1/2, 1/4\} \end{Bmatrix}$$

El problema surge ahora en el punto 1/8 que corresponde a dos valores de x que son x = 1/4 y x = 1/8. Continuando con este proceso, puede comprobarse que la función biyectiva será

$$f(x) = \lim_{n\to\infty} f_n(x) = \begin{Bmatrix} x & \Leftrightarrow x \notin \{1, 1/2, 1/4, 1/8, \dots\} \\ x/2 & \Leftrightarrow x \in \{1, 1/2, 1/4, 1/8, \dots\} \end{Bmatrix}$$

Si se define el conjunto $X_n = \{1, 1/2, 1/4, / 1/8, \dots, 1/n\}$ y

$$X = \lim_{n\to\infty} X_n = \lim_{n\to\infty} \{1, 1/2, 1/4, 1/8, \dots, 1/n\}$$

La función puede redefinirse de la siguiente manera

$$f(x) = \begin{Bmatrix} x & \Leftrightarrow x \notin X \\ x/2 & \Leftrightarrow x \in X \end{Bmatrix}$$

Este teorema permite demostrar fácilmente que existe una función biyectiva entre los pares ordenados $(x,y) \in \mathbb{N}^2$ y $x \in \mathbb{N}$. Demostraremos que existe una función

sobreyectiva f: $\mathbb{N}^2 \rightarrow \mathbb{N}$. La función es sencilla $f(x,y)=2^x3^y$. Esta función está bien definida porque dado que el teorema fundamental de la aritmética garantiza que la descomposición en los factores primos de un natural es única salvo el orden de los factores. La función g se puede definir $g(x)=(x,1)$: $\mathbb{N} \rightarrow \mathbb{N}^2$. Por lo tanto existe una función biyectiva entre h: $\mathbb{N} \leftrightarrow \mathbb{N}^2$.

El conjunto de gramáticas $G=\langle \Sigma, N, P, S\rangle$ tiene la cardinalidad $\aleph_0$

Se demostrara que existe una función biyectiva entre el conjunto Ψ de todas las gramáticas de la forma $\boldsymbol{G=\langle \Sigma, N, P, S\rangle}$ y el de los números naturales $\mathbb{N}$ donde G está en la forma canónica.

Para ello se empleará el teorema de Schröder-Bernstein-Cantor. El primer paso será demostrar que existe una función sobreyectiva f(G): $\Psi \rightarrow \mathbb{N}$ que mapea cada gramática G en un numero natural

Sea la función f definida de la siguiente manera:

$$f(G)=2^{|\Sigma|}\,3^{|N|}\,5^{q}$$

- Donde $|\Sigma|$ es la cantidad de símbolos del alfabeto o cantidad de terminales y
- $|N|$ es la cantidad de símbolos no terminales y
- q es un producto de potencias de números primos $q=p_1r^1\,p_2r^2...p_{|N|}{}^{r|N|}$ donde
  - pi es el i_esimo número primo y
  - la producción $(\gamma_j,\varphi_j)\epsilon P$ es la j.esima producción de la gramática y
  - $r_j=2^{aj}3^{bj}$ donde $a_j=\prod_{i=1}^{|\gamma_j|}p_i^{\gamma_j^i}$ y b_j y $bj=\prod_{i=1}^{|\sigma_j|}p_i{}^{\sigma_j^i}$ siendo $|\gamma\_j|$ y $|\sigma j|$ la longitud de la parte de la izquierda y de la parte de la derecha de la j-esima producción respectivamente.

Se demostrará que la función f(G) está bien definida. Dado m=f(G) es posible reconstruir la gramática canónica. La demostración de basa en el teorema fundamental de la aritmética por el cual la descomposición de un número natural en sus factores primos es única. En efecto dado m de su descomposición en sus factores primos se pueden determinar $|\Sigma|$, $|N|$ y q. Continuando con un procedimiento análogo dado q se pueden determinar los r_j $j=1,..., |N|$. Dados los valores de r_j y dado que $r_j=2^{aj}3^{bj}$ se pueden determinar los valores de a_j y b_j.

Finalmente dado que $a_j = \prod_{i=1}^{|\gamma_j|} p_i^{\gamma_j^i}$ y $b_j = \prod_{i=1}^{|\sigma_j|} p_i^{\sigma_j^i}$ se pueden determinar los valores de γ_i^j y σ_i^j con lo que se reconstruye el conjunto de producciones.

La demostración que existe una función sobreyectiva g que va de los naturales al conjunto de gramáticas es sencillo ya que dado un numero natural n cualquiera se le puede asociar la gramática $\boldsymbol{G=\langle \Sigma, N, P, S\rangle}$ donde

$$\boldsymbol{\Sigma=\{a\}, N=\{T\}, P=\{(R,a^n)\}\ S=T.}$$

El lenguaje generado por esta gramática contiene una sola cadena que consiste en n símbolos a seguidos.

Por lo tanto existen las funciones sobreyectivas $f:\Psi\rightarrow\mathbb{N}$ y $g:\mathbb{N}\rightarrow\Psi$ y por lo tanto debe existir de acuerdo con el teorema de Schröder-Bernstein-Cantor una función biyectiva $h:\Psi\leftrightarrow\mathbb{N}$.

Con esto queda demostrado que la cardinalidad de Ψ es $\aleph_0$.

Sea nuevamente la gramática $\boldsymbol{G=\langle \Sigma, N, P, S\rangle}$ donde

$\boldsymbol{\Sigma=\{a, b\}}$
$\boldsymbol{N=\{Q\}}$
$\boldsymbol{P=\{(Q,aQa),(Q,b)\}}$, $\boldsymbol{o\ en\ el\ otro\ formato\ Q\rightarrow aAa, Q\rightarrow b}$
S=Q

La gramática canónica correspondiente seria $\boldsymbol{G'=\langle \Sigma', N', P', S'\rangle}$ donde

$\boldsymbol{\Sigma'=\{1, 2\}}$
$\boldsymbol{N'=\{3\}}$
$\boldsymbol{P'=\{(3,131), (3,2)\}}$
$\boldsymbol{S'=3}$
$\boldsymbol{a_1=2^3=8,\ b_1=2^1 3^3 5^1=270}$
$\boldsymbol{a_2=23=8,\ b_2=2^2=4}$
Por lo que será
$\boldsymbol{r1=2^8 3^{270}}$
$\boldsymbol{r2=2^8 3^4}$

El valor de q resulta

$$q = 2^{2^{8}3^{270}}3^{2^{8}3^{4}}$$

Y finalmente

$$fG = 2^{2}3^{1}5^{2^{2^{8}3^{270}}3^{2^{8}3^{4}}}$$

Este número, para esta gramática tan sencilla, es enormemente grande y casi imposible de poder calcular. Pero lo importante es que existe.

7. Conclusiones

De lo demostrado previamente se concluye que la cardinalidad del conjunto de gramáticas $G=\langle \Sigma, N, P, S\rangle$ es $\aleph_0$ en tanto que la cardinalidad del conjunto de gramáticas que existen sobre Σ^* es la de 2^{Σ^*} que por el teorema de Cantor tiene la cardinalidad $\aleph_1$ del continuo. Por lo tanto existen infinitas gramáticas que no pueden ser representadas con dicho modelo. Ello explica la gran dificultad que existe en la representación de lenguajes naturales por medio de modelos finitos.

Bibliografía

[1] K. R. Chowdhari: "Fundamental of Discrete Mathematical", Second Edition, Structures PHI Learning Private Limited, January 2012

[2] Philip Edward Bertrand Jourdain, "Selected essays on the History of Set Theory and Logics", CLUEB 1991

[3] Joseph Warren Dauben, "Georg Cantor His Mathematics and Philosophy of the infinity, Princeton University Press 1979

[4] John E. Hopcroft, "Introduction To Automata Theory, Languages, And Computation", Pearson Education, 2008

[5] Rebecca Goldstein, "Gödel Paradoja y vida" , Antoni Bosh Editor España, 2005

[6] Ernest Nagel, James R. Newman, "Gödel's Proof", New York University Press, 2001

[7] Georg Cantor, "Contribution to the Founding of the theory of transfinite Numbers, Cosimo Inc. 2007