13068. Teoria d'Autòmats i Llenguatges Formals

1. Introducció

Francesc J. Ferri
Dept. d'Informàtica. Universitat de València

4 d'octubre de 2002

Definicions preliminars

símbol (etiqueta, caràcter): unitat lògica indivisible.

alfabet (vocabulari): conjunt finit de símbols x

$\Sigma_1 = \{a, b, c\}$, $\Sigma_2 = \{0, 1, 2, \ldots, 9\}$, $\Sigma_3 = \{$`vertader, fals, indefinit`$\}$

cadena (sentència, paraula, frase): successió finita de símbols d'un determinat alfabet.

$x_1 = aaabbba$, $x_2 = 1024$, $x_3 =$ `vertader·vertader·indefinit`

[Definicions preliminars]

longitud (talla): nombre de símbols d'una cadena

$|x_1| = 7, |x_2| = 4, |x_3| = 3$

Notació: $|x|_A$= { nombre de símbols de $A \subseteq \Sigma$ en x}

$|x_1|_a = |x_1|_{\{a\}} = 4,\ |x_1|_c = 0,\ |x_1|_{\{b,c\}} = 3$

concatenació: juxtaposició de dues cadenes. Operació interna dins del conjunt de totes les possibles paraules sobre un determinat alfabet.

$x \cdot y = xy, \quad aaa \cdot ab = aaaab$

cadena buida (nul·la,ε): cadena que no conté cap símbol. Element neutre de la concatenació.

[Definicions preliminars]

prefix: y és prefix de x si $\exists z$ tal que $x = yz$.

sufix: y és sufix de x si $\exists z$ tal que $x = zy$.

factor (subcadena, infix): y és factor de x si $\exists z, t$ tals que $x = zyt$. Els prefixs i sufixs en són casos particulars.

factor (prefix, sufix) propi de x: si és diferent de ε i de x.

factorització de x: descomposició d'una cadena en un nombre finit de factors, $x = x_1 \cdot x_2 \cdots x_k$.

[Definicions preliminars]

llenguatge: qualsevol conjunt de cadenes.

talla (grandària, cardinalitat) d'un llenguatge: nombre de cadenes que conté.

Llenguatges "especials" $\{\varepsilon\}$, i $\emptyset$

Exemples de llenguatges $\{a, aa, aaa, \ldots\}$, $\{0, 10, 20\}$ i $\{$<fals><fals>$, \varepsilon\}$, les talles dels quals són ∞, 3 i 2, respectivament

Cadenes sobre un alfabet

$P(\Sigma)$ és el conjunt format per *totes* les cadenes sobre un alfabet.

$$P(\{0,1\}) = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}$$

Aquesta ordenació de les cadenes s'anomena **lexicogràfica**.

El conjunt $P(\Sigma)$ té estructura de *monoide* respecte de la concatenació (p. associativa, elem. neutre).

$P(\Sigma)$: monoide lliure

$P(\Sigma) - \{\varepsilon\}$: semigrup lliure

Operacions sobre cadenes

Potència: $x^i = \underbrace{x \cdots x}_{i \text{ vegades}}$.

- $x^0 = \varepsilon$
- $x^i x^j = x^{i+j}$
- $x^{i+1} = x^i \cdot x = x \cdot x^i$
- $|x^i| = i \cdot |x|$

Inversió (reflexió): $x^{-1} = a_n \cdots a_2 a_1$ si $x = a_1 \cdots a_n$ i $a_i \in \Sigma$

$(\text{pepito})^{-1} = \text{otipep}$

Operacions sobre llenguatges

S'hi poden aplicar qualssevol operacions definides sobre conjunts

$$L_1 \cup L_2 \qquad L_1 \cap L_2 \qquad \overline{L_1} \qquad L_1 - L_2$$

concatenació: $L_1 \cdot L_2 = \{z | z = x \cdot y, x \in L_1, y \in L_2\}$

$\mathcal{L}(\Sigma)$: conjunt de tots els llenguatges sobre Σ (també $2^{P(\Sigma)}$), té estructura de monoide respecte de $\cdot$ i $\cup$.

Els elements neutres són $\{\varepsilon\}$ i $\emptyset$, resp.

Se l'anomena binoide lliure

Exemples (concatenació)

- $L \cdot \emptyset = \emptyset$
- $L \cdot \{\varepsilon\} = L$
- $\{a^n | n \geq 0\} \cdot \{b^n | n \geq 0\} = \{a^n b^m | n \geq 0, m \geq 0\}$
- $\{a^n | n \geq 0\} \cdot \{a^n | n \geq 0\} = \{a^n | n \geq 0\}$

$$L_1 = \{a, b\}^\star \qquad L_2 = a \cdot \{a, b\}^\star$$

$$L_1 \subset L_2?$$

$$|L_1| < |L_2|?$$

[Operacions sobre llenguatges]

Potència $L^i = \underbrace{L \cdot L \cdots L}_{i \text{ vegades}}$

$L^i = \{x_1 \cdot x_2 \cdots x_i | x_k \in L, k = 1, \ldots, i\}$

$L^0 = \{\varepsilon\}$

$L^{i+1} = LL^i = L^i L$

$L^{i+j} = L^i L^j = L^j L^i$

Inversió (reflexió): $L^{-1} = \{x^{-1} | x \in L\}$

[Operacions sobre llenguatges]

Tancament (clausura) de Kleene $L^\star = \bigcup_{i=0}^{\infty} L^i$

$\Sigma^\star = P(\Sigma)$

Tancament positiu: $L^+ = \bigcup_{i=1}^{\infty} L^i$

$L^\star = L^+ \cup \{\varepsilon\}$
$L^+ = L^\star - \{\varepsilon\}$ només si $\varepsilon \notin L$!
$L^+ = LL^\star = L^\star L$

Quocient: $L_1/L_2 = \{x \mid \exists y \in L_2 \ : \ xy \in L_1\}$

Prefixs de cadenes de L_1 el corresponent sufix del qual està en L_2.

Exemples (tancaments)

- $L^+L^+ = L^2L^\star = LL^\star L = L^\star L^2$
- $\emptyset^\star = \{\varepsilon\}$
- $\emptyset^+ = \emptyset$
- $\{a^n | n \geq 0\}^\star = \{a^n | n \geq 0\}$
- $\{a^n b^m | n \geq 0, m \geq 0\}^\star = \{a, b\}^\star$

Exemples (quocients)

Siga $L_a = \{a^n | n \geq 0\}$, $L_b = \{b^n | n \geq 0\}$, $L_{ab} = \{a^n b^m | n \geq 0, m \geq 0\}$

- $L/\emptyset = \emptyset$
- $L/\{\varepsilon\} = L$
- $L_a/L_b = L_a/\{\varepsilon\} = L_a$
- $L_{ab}/L_b = L_{ab}$
- $L_{ab}/\{ab\} = L_a$
- $L_{ab}/L_a = L_{ab}$

[Operacions sobre llenguatges]

Substitució: correspondència que associa un llenguatge a un símbol.

$f : \Sigma \longrightarrow 2^{\Gamma^\star}$

Podem definir $\begin{cases} f'(\varepsilon) = & \varepsilon \\ f'(ax) = & f(a)f'(x) \end{cases}$

$f' : \Sigma^\star \longrightarrow 2^{\Gamma^\star}$

També podem definir $f''(L) = \{f'(x) | x \in L\}$
$f'' : 2^{\Sigma^\star} \longrightarrow 2^{\Gamma^\star}$

No distingirem entre f, f' i f''

Exemple (substitució)

Siguen C i I dos llenguatges sobre l'alfabet $\Sigma = \{A-z, 0-9, +\}$.

$$C = \{1-9\} \cdot \{0-9\}^\star \qquad I = \{A-z\} \cdot \{A-z, 0-9\}^\star$$

Siga L un llenguatge sobre l'alfabet $\Gamma = \{c, i, +\}$.

$$L = (\{c\} \cup \{i\}) \cdot (\{+c\} \cup \{+i\})^\star$$

podem definir $f(c) = C$, $f(i) = I$, $f(+) = +$.

Aleshores $f(L)$ representa sumes arbitràries de identificadors i constants enteres.

[Operacions sobre llenguatges]

Homomorfisme: cas particular en què el llenguatge associat a un símbol consta d'una única cadena.

$h : \Sigma \longrightarrow \Gamma^\star$
$h' : \Sigma^\star \longrightarrow \Gamma^\star$
$h'' : 2^{\Sigma^\star} \longrightarrow 2^{\Gamma^\star}$

Homomorfisme invers:

$h^{-1} : \Gamma^\star \longrightarrow 2^{\Sigma^\star}$!!

$$h^{-1}(x) = \{w \in \Sigma^\star | h(w) = x\}$$

De la mateixa manera:

$$h^{-1}(L) = \{w \in \Sigma^\star | h(w) \in L\}$$

Exemples (homomorfismes)

Siga $h(a) = 0$ i $h(b) = 00$

$h^{-1}(x) = \{w \in \Sigma^\star | h(w) = x\}$
$h^{-1}(L) = \{w \in \Sigma^\star | h(w) \in L\}$

- $h(ab) = 000$
- $h^{-1}(000) = \{ab, ba, aaa\}$
- $h^{-1}(1) = \emptyset$
- $h^{-1}(\varepsilon) = \{\varepsilon\}$

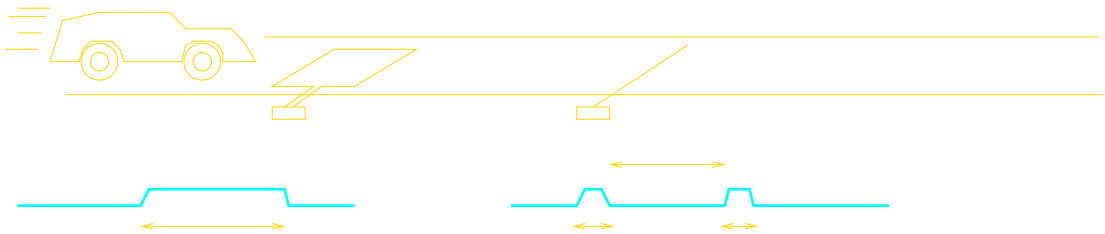
- $h(L_{ab}) = \{0^n | n \geq 0\}$
- $h(L_b) = \{0^{2n} | n \geq 0\}$
- $h^{-1}(\{0^{2n} | n \geq 0\}) =$ *exercici!*

Exemples de llenguatges

- $L_1 = \{a^n | n \geq 0\}$
- $L_2 = \{a^n b^m | n \geq 0, m \geq 0\}$
- $L_3 = \{w \in \{a, b\}^\star : |w|_a = |w|_b\}$
- $L_4 = \{w \in \{a, b\}^\star : |w| \leq 2\} = \{\varepsilon, a, b, aa, ab, ba, bb\}$
- $L_5 = \{w \in \{a, b\}^\star : w = w^{-1}\}$ (palíndromes)
- $L_6 = \{w \in \{a, b\}^\star : w = x \cdot x^{-1}\}$ (palíndromes parelles)

EXERCICIS: ${L_4}^\star$, $L_2 \cap L_3$, L_3/L_5, L_5/L_6, L_5/L_1, ...

Més exercicis



Siga $20 \leq T \leq 25$, $2 \leq R \leq 3$, $15 \leq E \leq 20$.

Descriure formalment els llenguatges L_1 i L_2 que representen cadenes de 0 i 1 corresponent a que ha passat un cotxe donades les mesures anteriors.

Descriure'ls de la forma més compacta possible fent servir els mecanismes que s'han vist fins ara.

i més

Volem detectar pàgines web (cadenes sobre l'alfabet ASCII) que continguen paraules amb el prefix propi **sex-**.

Exemple: sextet, sexagesimal, (Atenció: la pàgina sencera és una cadena!)

Descriure formalment de forma compacta el llenguatge corresponent.

i més encara

Descriure formalment el llenguatge sobre l'alfabet $\{a, (,)\}$ format per cadenes ben parentitzades com

$$(((a)(a))(a))a$$

- $L = \{w : |w|_(= |w|_)\}$?
- $L = \{w : |w|_(= |w|_) \wedge |x|_(\geq |x|_),\ \forall x : w = xy\}$?
- $L = \{w : w = (x),\ x \in L \vee w = xy, x \in L, y \in L\}$?

Inducció i recursió

L'última definició no té sentit però dóna la idea d'una definició inductiva per a L

1. $a \in L$.
2. $xy \in L$ si $x \in L$, $y \in L$.
3. $(x) \in L$ si $x \in L$.

Aquesta definició es pot llegir recursivament:

> Les cadenes de L són: o 1) una a, o 2) la concatenació de dues cadenes (més petites) de L, o 3) una cadena (més petita) de L envoltada per parèntesis.

Gramàtiques

La definició recursiva anterior mijantçant regles:

1. <Cadena>$\rightarrow a$
2. <Cadena>$\rightarrow$ <Cadena><Cadena>
3. <Cadena>$\rightarrow$ (<Cadena>)

El símbol <Cadena> representa qualsevol cadena de L. És una variable. Les regles expressen com pot estar formada aquesta cadena en funció d'altres cadenes.

Es pot dir que <Cadena> es reescriu com a (<Cadena>) (r. 3)

que es reescriu com a (<Cadena><Cadena>) (r. 2)

que es reescriu com a (aa) (r. 1 dues vegades).

Gramàtiques. definició

$G = \langle V_T, V_N, S, P \rangle$

- $V_T = \Sigma$ és l'alfabet o vocabulari terminal.
- V_N és l'alfabet no terminal.
- P és un conjunt de **produccions**

  $\alpha \rightarrow \beta$: $\quad \beta \in V^\star$ i $\alpha \in V^\star V_N^+ V^\star$ $(V = V_T \cup V_N)$.
- $S \in V_N$ **axioma**.

[Gramàtiques. definició]

derivació directa $\boxed{x \underset{G}{\Rightarrow} y}$ $\exists u \to v \in P \quad \wedge \quad x = \alpha u \beta, \quad y = \alpha v \beta.$

derivació (tancament transitiu de $\underset{G}{\Rightarrow}$) $\boxed{x \overset{\star}{\underset{G}{\Rightarrow}} y}$

$$x = y \vee x \underset{G}{\Rightarrow} a_1 \underset{G}{\Rightarrow} a_2 \underset{G}{\Rightarrow} \cdots \underset{G}{\Rightarrow} y$$

formes sentencials $x \in V^\star : S \overset{\star}{\underset{G}{\Rightarrow}} x$

sentències formes sentencials de V_T

llenguatge generat per G $L(G) = \{w \in V_T^* \mid \ S \overset{+}{\underset{G}{\Rightarrow}} w\}$

Exemples. Gramàtiques

- $S \to SS|(S)|a$
  $S \Rightarrow S\underline{S} \Rightarrow S\underline{S}S \Rightarrow S(\underline{S})S \Rightarrow S(S\underline{S})S \Rightarrow S(S(S))S \quad \overset{\star}{\Rightarrow} \quad a(a(a))a$

- $S \to S + S|S * S|(S)|x$
  $\underline{S} \Rightarrow \underline{S} * S \Rightarrow (\underline{S}) * S \Rightarrow (\underline{S} + \underline{S}) * \underline{S} \quad \overset{+}{\Rightarrow} \quad (x + x) * x$

- $S \to 0S|A|\varepsilon$
  $A \to 1A|\varepsilon$
  $S \Rightarrow 0S \Rightarrow 00S \Rightarrow 00A \Rightarrow 001A \Rightarrow 0011A \Rightarrow 00111A \Rightarrow 00111$

La jerarquia de Chomsky

tipus 3. Regulars

$A \to aB \quad o \quad A \to a \qquad a \in V_T,\ A, B \in V_N$

tipus 2. Incontextuals

$A \to \alpha \qquad A \in V_N,\ \alpha \in V^\star$

tipus 1. Contextuals

$xAy \to x\beta y \quad o \quad S \to \varepsilon \qquad x, y \in V^\star,\ A \in V_N,\ \beta \in V^+!$

tipus 0. No restringides

Jerarquia de Chomsky. Exemples

tipus 3: $S \to 0A|1A|\varepsilon$
$A \to 0S|1S$

tipus 2: $S \to 0S0|1S1|0|1|\varepsilon$

tipus 1: $S \to T|\varepsilon$
$T \to aTBD|abD \quad CD \to BD$
$DB \to CB \quad bB \to bb$
$CB \to CD \quad D \to c$

tipus 0: $S \to aSAc|abc|\varepsilon$
$cA \to Ac$
$bA \to bb$

Gramàtica de tipus 0?

$S \to aSAc \vert abc \vert \varepsilon$	$S \to aSAc \vert abc \vert \varepsilon$
$cA \to Ac$	$cA \to cX$ $\quad cX \to AX$
$bA \to bb$	$AX \to Ac$ $\quad bA \to bb$

$S \Rightarrow aSAc \Rightarrow aaSAcAc \overset{\star}{\Rightarrow} a^{n-1}S(Ac)^{n-1} \Rightarrow$

$\Rightarrow a^n bc(Ac)^{n-1} = a^n b(cA)^{n-1}c \overset{\star}{\Rightarrow} \boxed{a^n b(Ac)^{n-1}c}$

$= a^n bAc(Ac)^{n-2}c \Rightarrow a^n bbc(Ac)^{n-2}c =$

$a^n bb(cA)^{n-2}cc \overset{\star}{\Rightarrow} \boxed{a^n bb(Ac)^{n-2}cc}$

...

$\overset{\star}{\Rightarrow} a^n b^{n-1}(cA)^1 c^{n-1} \overset{\star}{\Rightarrow} a^n b^{n-1}(Ac)c^{n-1} \Rightarrow a^n b^n c^n$

Definicions alternatives

$\alpha \to \beta : |\alpha| \geq |\beta| \implies$ Contextual!

$xAy \to x\beta y \; x, y \in V^\star, \; A \in V_N, \; \boxed{\beta \in V^\star} \implies$ tipus 0!

$A \to \alpha B$
$A \to \alpha$ $\qquad A \in V_N, \; \alpha \in {V_T}^\star$ ho considerarem regular.

$A \to B\alpha$
$A \to \alpha$ $\qquad A \in V_N, \; \alpha \in {V_T}^\star$ regular per la *dreta*!

$S \to 0S \vert S1 \vert \varepsilon$ no és ni regular per la dreta ni per l'esquerra !!!

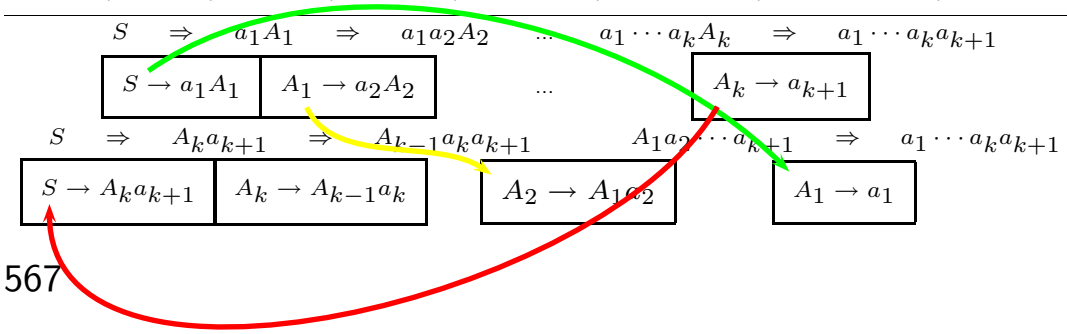
Exemple. Gramàtiques regulars

- $S \to S1|A|\varepsilon$ (per la dreta)
  $A \to A0|\varepsilon$

  $S \Rightarrow S1 \Rightarrow S11 \Rightarrow S111 \Rightarrow A111 \Rightarrow A0111 \Rightarrow A00111 \Rightarrow 00111$

- $S \to 0S|A|\varepsilon$ (per l'esquerra)
  $A \to 1A|\varepsilon$

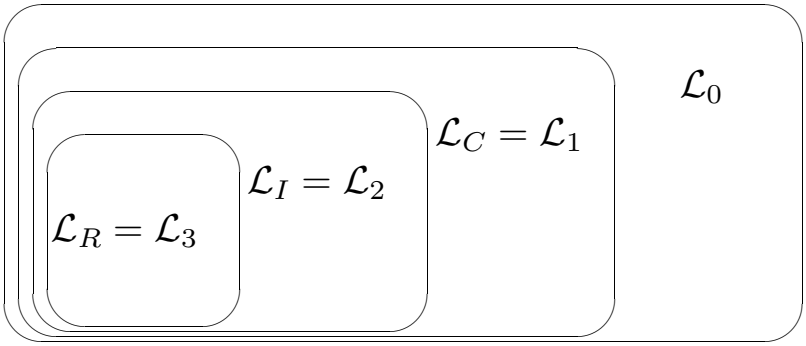
  $S \Rightarrow 0S \Rightarrow 00S \Rightarrow 00A \Rightarrow 001A \Rightarrow 0011A \Rightarrow 00111A \Rightarrow 00111$



567

Jerarquia de Chomsky. Llenguatges

Un llenguatge és de tipus N si existeix alguna gramàtica de tipus N que el genere.



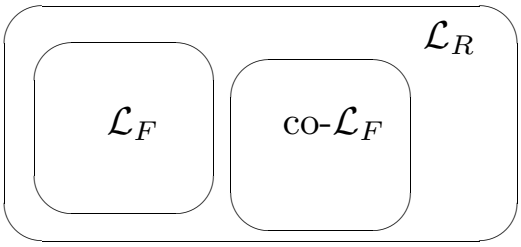
Exemples

$\{a^n b^n c^n | n \geq 0\} \in \mathcal{L}_C \subseteq \mathcal{L}_0$

$\{0^n 1^n | n \geq 0\} \in \mathcal{L}_I \subseteq \mathcal{L}_C \subseteq \mathcal{L}_0$

llenguatges finits: $|L| < \infty$ $\{a\}$

llenguatges cofinits: $|\overline{L}| < \infty$ $\{a, b\}^+$



Grans preguntes sense resposta ...

(...de moment)

- $\mathcal{L}_R \subseteq \mathcal{L}_I \subseteq \mathcal{L}_C \subseteq \mathcal{L}_0$ però són pròpies aquestes inclusions?
- Hi ha llenguatges que no siguen $\mathcal{L}_0$?
- On està la subclasse dels llenguatges coregulars?
- I els coincontextuals?
- Coincideix la noció intuïtiva de complexitat (estructural) d'un llenguatges amb els tipus de la jerarquia de Chomsky?
- Hi ha més classes dins de la jerarquia? I fora?