Parole e linguaggi

Arturo Carpi

Dipartimento di Matematica e Informatica
Università di Perugia

Corso di Linguaggi Formali e Compilatori - a.a. 2021/22

Finito e infinito

- Per descrivere qualsiasi cosa usiamo delle parole (e.g., La Divina Commedia, il codice genetico umano, . . .)
- un linguaggio è un insieme di parole,
- se è finito, posso elencarne gli elementi: $L = \{a, ab, abb, abbb\}$
- come descrivere un linguaggio infinito?

Teorema di Cantor

Non esiste una funzione iniettiva di $\wp(\Sigma^*)$ in Σ^*.
Quindi non è possibile associare a ciascun linguaggio una parola (la sua 'descrizione') che lo caratterizzi univocamente.

Esempio

È comunque possibile dare descrizioni finite di alcuni linguaggi infiniti:

- Le parole sull'alfabeto $\{a, b\}$ di lunghezza dispari,
- $\{a^n b^n \mid n \geq 0\}$,
- la chiusura di Kleene di un linguaggio L è il linguaggio L^* costituito dalla parola vuota, dalle parole di L e da tutte quelle che si ottengono concatenando due o più parole di L. Se L è un linguaggio finito, il linguaggio L^* è ben definito e, in generale, infinito.

Un esempio

Le frasi del tipo “Aldo, Bianca e Carlo” in cui:

1. i nomi sono separati da virgola, tranne gli ultimi due separati da ’e’,
2. sono ammesse ripetizioni,
3. nessun limite alla lunghezza della lista.

Quindi sono ammesse “Aldo, Bianca, Aldo, Carlo e Bianca” oppure “Bianca” ma non “Aldo, Bianca”

Le seguenti regole definiscono formalmente le nostre frasi:

1. Aldo, Bianca e Carlo sono nomi;
2. un nome è una frase;
3. un nome seguito da una virgola e una frase è anch’esso una frase;
4. prima di terminare, se sono presenti virgole, l’ultima va sostituita con la congiunzione e.

Un esempio

Osservazione

Possiamo riguardare i termini nome e frase come segnaposto e eseguire sostituzioni secondo regole predefinite.

1. nome può essere sostituito da Aldo
   nome può essere sostituito da Bianca
   nome può essere sostituito da Carlo
2. frase può essere sostituito da nome
   frase può essere sostituito da nome, frase
3. , nome a fine frase si deve sostituire con e nome prima che nome sia sostituito a sua volta
4. si inizia con frase
5. ci si arresta solo quando frase e nome non compaiono più nella frase che stiamo costruendo;

Osservazione

La regola 3 è diversa dalle precedenti. Può essere sostituita da altre regole del tipo precedente.

Un esempio

1. nome può essere sostituito da Aldo
   nome può essere sostituito da Bianca
   nome può essere sostituito da Carlo
2. frase può essere sostituito da nome
   frase può essere sostituito da listaNomi fineLista
3. listaNomi può essere sostituito da nome
   listaNomi può essere sostituito da nome, listaNomi
4. , nome fineLista può essere sostituito da e nome
5. si inizia con frase
6. ci si arresta solo quando frase, listaNomi, fineLista e nome non compaiono più nella frase che stiamo costruendo;

Un esempio

1. nome $\rightarrow$ Aldo
   nome $\rightarrow$ Bianca
   nome $\rightarrow$ Carlo
2. frase $\rightarrow$ nome
   frase $\rightarrow$ listaNomi fineLista
3. listaNomi $\rightarrow$ nome
   listaNomi $\rightarrow$ nome, listaNomi
4. , nome fineLista $\rightarrow$ e nome
5. si inizia con frase
6. ci si arresta solo quando frase, listaNomi, fineLista e nome non compaiono più nella frase che stiamo costruendo;

Parole

Definizione

Chiameremo alfabeto un insieme finito non vuoto Σ di simboli. I suoi elementi sono detti lettere.

Esempi

$$\Sigma_0 = \{a, b\}, \quad \Sigma_1 = \{0, 1\}, \quad \Sigma_2 = \{a, b, c\},$$

$$\Sigma_3 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}.$$

Definizione

Ogni sequenza finita di lettere di Σ si dice parola sull'alfabeto Σ. L'insieme delle parole sull'alfabeto Σ sarà denotato con Σ^*.

Esempi

a, abb, $ababbabb$ sono parole sull'alfabeto Σ_0,
01001010, 0110, 0000 sono parole sull'alfabeto Σ_1.
2020, $5E7$, $CD078B$ sono parole sull'alfabeto Σ_3.

Possiamo anche considerare la sequenza di zero lettere, che si denota con ε (oppure Λ) e si dice parola vuota.

Lunghezza

Una parola sull'alfabeto Σ è una sequenza

$$u = a_1 a_2 \cdots a_k$$

con $k \geq 0$, $a_1, a_2, \ldots, a_k \in \Sigma$.

Definizione

L'intero k si dice lunghezza della parola u e si denota con $|u|$, o anche $\ell(u)$.

Esempi

$|a| = 1, \quad |abb| = 3, \quad |ababbabb| = 8, \quad |\varepsilon| = 0.$

Concatenazione

Definizione

Si considerino le parole $u = a_1\,a_2 \cdots a_k$ e $v = b_1\,b_2 \cdots b_h$ ($k, h \geq 0$, $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_h \in \Sigma$). La concatenazione di u e v è la parola

$$uv = a_1\,a_2 \cdots a_k\,b_1\,b_2 \cdots b_h\,.$$

Esempi

La concatenazione delle parole abb e $aaab$ è la parola $abbaaab$.
La concatenazione delle parole $aaab$ e abb è la parola $aaababb$.
La concatenazione delle parole baa e ε è la parola baa.

Proprietà

La concatenazione è un'operazione binaria (totale) su Σ^*. Verifica le seguenti identità

- per ogni $u, v, w \in \Sigma^*$, $\quad (uv)w = u(vw)$ (proprietà associativa),
- per ogni $u \in \Sigma^*$, $\quad u\varepsilon = \varepsilon u = u$ (elemento neutro),
- se si ha $\quad uw = vw$ oppure $\quad wu = wv$ con $u, v, w \in \Sigma^*$, allora $u = v$ (cancellatività a destra e a sinistra).

Potenze e fattori

Definizione

Sia $n \geq 0$. La potenza n-esima di una parola w si ottiene concatenando n copie della parola w:

$$w^n = \underbrace{w\, w\, \cdots\, w}_{n \text{ volte}}$$

In particolare, $w^0 = \varepsilon$ e $w^1 = w$.

Esempio

Se $u = abb$, allora $u^0 = \varepsilon,\ u^1 = u = abb,\ u^2 = abbabb$, $u^3 = abbabbabb$.

Definizione

Diremo che una parola v è un fattore di una parola w se risulta $w = xvy$ per opportune parole x, y. Nel caso in cui $x = \varepsilon$ (risp., $y = \varepsilon$) il fattore v si dice prefisso (risp., suffisso) di w. Diremo che v è un fattore proprio se $v \neq w$.

Esempio

I fattori di abb sono ε, a, b, ab, bb e abb, i prefissi sono ε, a, ab e abb, e i suffissi sono ε, b, bb e abb.

Linguaggi formali

Definizione

Ogni sottoinsieme di Σ^* si dice linguaggio formale (o, brevemente, linguaggio) sull'alfabeto Σ.

Esempio

Sono linguaggi formali sull'alfabeto $\Sigma = \{a, b\}$:

$$L_0 = \{a, b\}, \quad L_1 = \{a, ab, abb\}, \quad L_2 = \{ab^n a \mid n \geq 0\},$$
$$L_3 = \emptyset, \quad L_4 = \Sigma^*.$$

Definizione

Una grammatica a struttura di frase è una quadrupla

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

ove

- V è un alfabeto finito, detto vocabolario totale,
- $\Sigma \subseteq V$ è l'alfabeto dei simboli terminali,
- P è un insieme finito di espressioni della forma

$$\alpha \to \beta$$

  con $\alpha \in V^* \setminus \Sigma^*$ e $\beta \in V^*$, detto insieme delle produzioni
- $S \in N = V \setminus \Sigma$ è il simbolo iniziale o assioma,

Le lettere di $N = V \setminus \Sigma$ si dicono variabili.

Convenzione

Si suole indicare

- le variabili con $A,\ B,\ C,\ X,\ Y,\ Z,\ \ldots$
- i simboli terminali con $a,\ b,\ c,\ \ldots$
- le parole sull'alfabeto V con $\alpha,\ \beta,\ \gamma,\ \ldots$
- le parole sull'alfabeto Σ dei terminali con $u, v, w, \ldots$

Adeguiamoci!

Il linguaggio generato

Siano $\alpha, \beta \in V^*$.

- Diremo che β è una conseguenza diretta di α (e scriveremo $\alpha \Rightarrow \beta$) se esistono parole $\gamma_1, \gamma_2 \in V^*$ e una produzione $\gamma \to \gamma'$ in P tali che

$$\alpha = \gamma_1 \gamma \gamma_2, \quad \beta = \gamma_1 \gamma' \gamma_2 \,.$$

- Diremo che β si deriva (o è una conseguenza) di α in G (e scriveremo $\alpha \Rightarrow^* \beta$) se esistono $n > 0$, $\alpha_0, \alpha_1, \ldots, \alpha_n \in V^*$ tali che

$$\alpha = \alpha_0 \Rightarrow \alpha_1 \Rightarrow \cdots \Rightarrow \alpha_n = \beta \,.$$

- Le conseguenze del simbolo iniziale S si dicono forme sentenziali.
- Il linguaggio generato da G è l'insieme delle forme sentenziali prive di variabili.

Il nostro esempio

Nel nostro esempio,

- le variabili sono

  $$\langle \text{frase} \rangle, \langle \text{listaNomi} \rangle, \langle \text{fineLista} \rangle, \langle \text{nome} \rangle,$$

- i terminali sono le lettere delle parole Aldo, Bianca, Carlo, e, la virgola, lo spazio
- le produzioni sono:

  $\langle \text{nome} \rangle \to$ Aldo
  $\langle \text{nome} \rangle \to$ Bianca
  $\langle \text{nome} \rangle \to$ Carlo
  $\langle \text{frase} \rangle \to \langle \text{nome} \rangle$

  $\langle \text{frase} \rangle \to \langle \text{listaNomi} \rangle \langle \text{fineLista} \rangle$
  $\langle \text{listaNomi} \rangle \to \langle \text{nome} \rangle$
  $\langle \text{listaNomi} \rangle \to \langle \text{nome} \rangle, \langle \text{listaNomi} \rangle$
  , $\langle \text{nome} \rangle \langle \text{fineLista} \rangle \to$ e $\langle \text{nome} \rangle$