La gerarchia di Chomsky

Arturo Carpi

Dipartimento di Matematica e Informatica
Università di Perugia

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

Grammatiche

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,
- $S \in N = V \setminus \Sigma$ è il simbolo iniziale o assioma,
- 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

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 $n \geq 0$, $\alpha_0, \alpha_1, \ldots, \alpha_n \in V^*$ 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.

Esempio

$$G = \langle V, \Sigma, P, S \rangle, \quad V = \{a, b, S\}, \quad \Sigma = \{a, b\}, \quad N = \{S\},$$
$$P: \qquad S \to ab\,, \qquad S \to aSb\,.$$

Si ha

$$bSa \Rightarrow baba, \quad aS \Rightarrow aaSb, \quad aaSb \Rightarrow aaaSbb, \quad aS \stackrel{*}{\Rightarrow} aaaSbb\,.$$

Ma qual'è il linguaggio generato?

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow \cdots \Rightarrow a^{n-1}Sb^{n-1} \Rightarrow a^n b^n\,.$$

Quindi

$$S(G) = \{a^n Sb^n \mid n > 0\} \cup \{a^n b^n \mid n > 0\}\,,$$
$$L(G) = \{a^n b^n \mid n > 0\}\,.$$

Esempio 2

Costruiamo una grammatica per il linguaggio

$$L = \{a^n b^n c^n \mid n > 0\}.$$

Prendiamo $G = \langle V, \Sigma, P, S\rangle$, con $\Sigma = \{a, b, c\}$, $N = \{S, B\}$, e produzioni

$$S \to aSBc\,, \quad S \to abc\,, \quad cB \to Bc\,, \quad bB \to bb\,.$$

- Verifichiamo che $L \subseteq L(G)$
  Per es., mostriamo che $a^3b^3c^3 \in L(G)$.

$$\begin{aligned} S &\Rightarrow aSBc \Rightarrow aaSBcBc \Rightarrow aaabcBcBc \\ &\Rightarrow aaabBccBc \Rightarrow aaabBcBcc \Rightarrow aaabBBccc \\ &\Rightarrow aaabbBccc \Rightarrow aaabbbccc\,. \end{aligned}$$

Quindi $S \Rightarrow^* a^3b^3c^3$, cioè, $a^3b^3c^3 \in L(G)$.

Esempio 2

$$S \to aSBc\,, \quad S \to abc\,, \quad cB \to Bc\,, \quad bB \to bb\,.$$

- Verifichiamo che $L(G) \subseteq L$.
  - Le forme sentenziali hanno una delle forme seguenti:
    1. $a^n S w$, ove w è una permutazione di $B^n c^n$;
    2. $a^n b^m w$, ove w è una permutazione di $B^{n-m} c^n$;

    (perchè S è del tipo 1 e le produzioni preservano la proprietà).
  - Una parola di $L(G)$ non può che essere del tipo 2 con $n = m$, cioè $a^n b^n c^n$.

Grammatiche equivalenti

Definizione

Due grammatiche si dicono equivalenti se generano lo stesso linguaggio.

Ricognizione e parsing

Problema di ricognizione

Input: una grammatica $G = \langle V, \Sigma, P, S \rangle$ e una parola $w \in \Sigma^*$;
Output: SI se $w \in L(G)$, NO altrimenti.

Problema di parsing

Input: una grammatica $G = \langle V, \Sigma, P, S \rangle$ e una parola $w \in L(G)$;
Output: una derivazione $S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \Rightarrow \alpha_n \Rightarrow w$ di w in G.

Non esiste un algoritmo che risolva questi problemi nel caso generale.

Occorre quindi restringersi a classi particolari di grammatiche bilanciando

- efficienza degli algoritmi
- espressività delle grammatiche

Linguaggi di tipo 0

Definizione

Le grammatiche a struttura di frase si dicono anche grammatiche di tipo 0.
I linguaggi generati da grammatiche di tipo 0 si dicono linguaggi di tipo 0 o anche linguaggi ricorsivamente enumerabili.

Esempio

La grammatica con le produzioni

$$
\begin{array}{lll}
S \to N & A \to N & N \to a \\
S \to N, AF & A \to N, A & N \to b \\
 & , NF \to \&N & N \to c
\end{array}
$$

genera, per es., le parole a $a, b\&c$ $a, c, a\&b$
È una grammatica di tipo 0.

Grammatiche sensibili al contesto

Definizione

Una grammatica a struttura di frase si dice sensibile al contesto se le produzioni hanno la forma

$$\alpha_1 X \alpha_2 \to \alpha_1 \beta \alpha_2 \,, \qquad \text{con } X \in N, \quad \alpha_1, \alpha_2, \beta \in V^*, \quad \beta \neq \varepsilon \,.$$

I linguaggi generati da grammatiche di tipo 1 si dicono linguaggi di tipo 1 o anche sensibili al contesto.

Esempio

La grammatica con le produzioni

$$\begin{array}{llll} S \to NVS & VN \to, N & N \to a & U \to a \\ S \to U & VU \to \& U & N \to b & U \to b \\ & & N \to c & U \to c \end{array}$$

è una grammatica di tipo 1, equivalente a quella dell'esempio precedente.

Grammatiche monotòne

Definizione

Una grammatica si dice monotòna se tutte le produzioni hanno la forma

$$\alpha \to \beta \quad \text{con } |\alpha| \leq |\beta| \,.$$

- le grammatiche sensibili al contesto sono monotòne;
- non tutte le grammatiche monotòne sono sensibili al contesto;
- ogni grammatica monotòna è equivalente a una grammatica sensibile al contesto.

Esempio

La grammatica con le produzioni

$$S \to aSBc\,, \quad S \to abc\,, \quad cB \to Bc\,, \quad bB \to bb\,.$$

è monotòna ma non è sensibile al contesto. Il linguaggio generato $L = \{a^n b^n c^n \mid n > 0\}$ è un linguaggio di tipo 1.

Linguaggi di tipo 1

1. i linguaggi di tipo 1 costituiscono una sottoclasse propria dei linguaggi di tipo 0;
2. esistono algoritmi che risolvono il problema di ricognizione e il problema di parsing per grammatiche di tipo 1 (in generale, molto costosi)
3. invece esistono linguaggi di tipo 0 per cui un tale algoritmo non esiste.

Grammatiche non contestuali

Definizione

Una grammatica a struttura di frase si dice non contestuale o di tipo 2 se le produzioni hanno la forma

$$X \to \beta\,, \qquad \text{con } X \in N, \quad \beta \in V^*\,.$$

I linguaggi generati da grammatiche di tipo 2 si dicono linguaggi di tipo 2 o anche non contestuali.

Esempio

La grammatica con le produzioni

$$\begin{array}{lll} S \to N & M \to N \& N & N \to a \\ S \to M & M \to N, M & N \to b \\ & & N \to c \end{array}$$

è una grammatica di tipo 2, equivalente a quella di un esempio precedente.

Linguaggi di tipo 2

1. i linguaggi di tipo 2 costituiscono una sottoclasse propria dei linguaggi di tipo 1. Per esempio $L = \{a^n b^n c^n \mid n > 0\}$ è un linguaggio di tipo 1 ma non è di tipo 2;
2. gli algoritmi per ricognizione e parsing per grammatiche di tipo 2 saranno uno dei principali argomenti del corso. Essi hanno importanti applicazioni nell'analisi sintattica.

Grammatiche regolari

Definizione

Una grammatica a struttura di frase si dice regolare o di tipo 3 se le produzioni hanno la forma

$$X \to aY \quad \text{oppure} \quad X \to a\,, \qquad \text{con } X, Y \in N, \quad a \in \Sigma\,.$$

I linguaggi generati da grammatiche di tipo 3 si dicono linguaggi di tipo 3 o anche regolari.

Esempio

La grammatica con le produzioni

$S \to a$	$S \to aR$	$R \to \&N$	$M \to aR$	$N \to a$
$S \to b$	$S \to bR$	$R \to , M$	$M \to bR$	$N \to b$
$S \to c$	$S \to cR$		$M \to cR$	$N \to c$

è una grammatica di tipo 3, equivalente a quella dell'esempio precedente.

Linguaggi di tipo 3

1. i linguaggi di tipo 3 costituiscono una sottoclasse propria dei linguaggi di tipo 2. Per esempio $L = \{a^n b^n \mid n > 0\}$ è un linguaggio di tipo 2 ma non è di tipo 3;
2. gli algoritmi per ricognizione per linguaggi di tipo 3 saranno uno dei principali argomenti del corso. Essi hanno importanti applicazioni nell'analisi lessicale.