Azzolini Riccardo 2020-12-07

Alcune proprietà dei linguaggi context-free

1 Pumping lemma per i linguaggi context-free

Come per i linguaggi regolari, anche per quelli context-free esiste un **pumping lemma**, il quale può essere usato per dimostrare che determinati linguaggi non sono context-free.

Teorema: Sia L un linguaggio context-free. Esiste una costante n tale che, per ogni stringa $z \in L$ con $|z| \geq n$, esiste una scomposizione $z = uVwXy$ che soddisfa le seguenti proprietà:

1. $|VwX| \leq n$;
2. $VX \neq \epsilon$, cioè almeno una tra le stringhe V e X deve contenere almeno un simbolo;
3. per ogni $i \geq 0$, $uV^iwX^iy \in L$.

Siccome la scelta di n per un determinato linguaggio L non è univoca, per semplificare il discorso si chiamerà **costante di pumping** N o N_L la più piccola n per cui vale il lemma.

1.1 Esempio di applicazione

Dato il linguaggio

$$L = \{0^n1^n2^n \mid n \geq 1\} = \{012, 001122, 000111222, \dots\}$$

si dimostra che esso non è context-free usando il pumping lemma, seguendo lo stesso schema applicato nel caso dei linguaggi regolari.

Si suppone per assurdo che L sia context-free, e che N sia la relativa costante di pumping. Allora, si considera la stringa

$$z = 0^N1^N2^N = 0_1 \dots 0_N 1_1 \dots 1_N 2_1 \dots 2_N \in L$$

(qui le N occorrenze di ciascun simbolo $a \in \{0, 1, 2\}$ sono state indiciate da a_1 a a_N). Siccome $|z| = 3N > N$, per il pumping lemma esiste una scomposizione $z = uVwXy$ che dovrebbe verificare le proprietà (1)–(3) del lemma, ma assumendo le prime due si dimostra che la terza non può essere verificata.

Per la (1), $|VwX| \leq N$, la stringa VwX o non contiene 2

$$0_1 \dots\dots \overbrace{0_N 1_1 \dots\dots}^{VwX} 1_N 2_1 \dots\dots 2_N$$

oppure non contiene 0:

$$0_1 \dots\dots 0_N 1_1 \overbrace{\dots\dots 1_N 2_1 \dots\dots}^{VwX} 2_N$$

Ora, per la (3) nel caso in cui $i = 0$, dovrebbe essere $uV^0wX^0y = uvy \in L$, ma:

- se VwX non contiene 2, allora per la proprietà (2), $VX \neq \epsilon$, in V o X deve essere presente almeno uno 0 oppure un 1, che in uV^0wX^0y viene eliminato, lasciando complessivamente almeno uno 0 o un 1 in meno rispetto al numero di 2, che rimane invariato;
- analogamente, se VwX non contiene 0, per la (2) VX deve contenere almeno un 1 o un 2, quindi uV^0wX^0y ha come minimo un 1 o un 2 in meno rispetto al numero di 0.

In entrambi i casi, $uV^0wX^0y \notin L$: questo contraddice il pumping lemma, portando a dedurre che L non è un linguaggio context-free.

1.2 Esempio di applicazione: numeri primi

Si consideri il linguaggio dei numeri primi in rappresentazione unaria,

$$L_{pr} = \{w \in \{1\}^* \mid |w| \text{ è un numero primo}\} = \{1^p \mid p \text{ è un numero primo}\}$$

che si è già dimostrato essere non regolare. Adesso, si vuole dimostrare che esso non è nemmeno context-free.

Per iniziare, si suppone per assurdo che L_{pr} sia context-free, e che N sia la relativa costante di pumping. Si sceglie poi un numero primo $p \geq N + 2$ (che esiste sicuramente perché i numeri primi sono infiniti), e si considera la stringa $z = 1^p \in L_{pr}$. Siccome $|z| = p > N$, per il pumping lemma esiste una scomposizione $z = uVwXy$ che dovrebbe soddisfare le proprietà (1)–(3), ma invece, assumendo la (1) e la (2), si dimostra che la (3) non può essere verificata.

Sia $m = |VX|$. Dalla (2), $VX \neq \epsilon$, si ha che $m \geq 1$, mentre dalla scelta di $p \geq N + 2$ e dalla (1), $|VwX| \geq N$, si deduce che $m \leq N \leq p - 2$; complessivamente:

$$1 \leq m \leq N \leq p - 2$$

Sapendo la lunghezza di VX, si ricava quella delle altre parti della stringa $z = uVwXy$,

$$|uwy| = |uVwXy| - |VX| = |z| - |VX| = p - m$$

e da $m \leq p - 2$ segue che
$$|uwy| = p - m \geq 2$$

Per la (3) del pumping lemma, con $i = p - m$, dovrebbe essere $uV^{p-m}wX^{p-m}y \in L_{pr}$. La lunghezza di questa stringa è

$$\begin{aligned}
|uV^{p-m}wX^{p-m}y| &= |uvw| + |V^{p-m}X^{p-m}| \\
&= |uvw| + |(VX)^{p-m}| \\
&= |uvw| + |VX|(p-m) \\
&= (p-m) + m(p-m) \\
&= (p-m)(m+1)
\end{aligned}$$

cioè un numero che è il prodotto di due fattori interi, entrambi maggiori di 1:

- si è visto prima che $p - m \geq 2 > 1$;
- si ha $m + 1 > 1$ perché $m \geq 1$.

Allora, $|uV^{p-m}wX^{p-m}y|$ è un numero composto, non primo, quindi per definizione $uV^{p-m}wX^{p-m}y \notin L_{pr}$, contrariamente al pumping lemma. Ciò è assurdo, dunque deve essere falsa l'assunzione iniziale che L_{pr} fosse un linguaggio context-free.

2 Proprietà di chiusura dei linguaggi context-free

Teorema: La classe dei linguaggi liberi dal contesto è chiusa rispetto alle seguenti operazioni:

- unione (se L_1 e L_2 sono CFL, anche $L_1 \cup L_2$ è un CFL);
- concatenazione (L_1L_2);
- chiusura di Kleene (L^*) e chiusura positiva (L^+);
- inversione ($L^R = \{w^R \mid w \in L\}$).

Invece, a differenza dei linguaggi regolari, i linguaggi context-free *non* sono chiusi rispetto all'operazione di intersezione: se L_1 e L_2 sono CFL, non è garantito che anche $L_1 \cap L_2$ sia un CFL. Ad esempio:

- Il linguaggio $L_1 = \{0^n1^n2^i \mid n \geq 1, i \geq 1\}$ comprende le stringhe formate da un certo numero di 0, seguiti dallo stesso numero di 1 e poi da un numero qualsiasi di 2. Esso è un CFL in quanto si può dimostrare che è generato da una grammatica con le regole di produzione
$$\begin{aligned}
&S \to AB \\
&A \to 0A1 \mid 01 \\
&B \to 2B \mid 2
\end{aligned}$$

dove S è il simbolo iniziale.

- Il linguaggio $L_2 = \{0^i 1^n 2^n \mid n \geq 1, i \geq 1\}$ comprende le stringhe formate da un numero qualsiasi di 0, seguiti da un determinato numero di 1 e dallo stesso numero di 2. In pratica, questo linguaggio è ottenuto da L_1 scambiando i ruoli di 0 e 2. Anch'esso è un CFL, in quanto generato dalla seguente grammatica:

$$\begin{aligned} &S \to AB \\ &A \to 0A \mid 0 \\ &B \to 1B2 \mid 12 \end{aligned}$$

L'intersezione di questi due linguaggi contiene le stringhe che hanno lo stesso numero di 0, 1 e 2,

$$L_1 \cap L_2 = \{0^n 1^n 2^n \mid n \geq 1\}$$

ma si è dimostrato prima che tale linguaggio non è context-free.