Context-Free languages (part III)

Prof. Dan A. Simovici

UMB

1. The Pumping Lemma for Context-Free Languages
2. Closure Properties of $\mathcal{L}_2$

A preliminary result

Theorem

Let $G = (A_N, A_T, S, P)$ *be a grammar in Chomsky normal form.*
If T *is a derivation tree for a derivation* $X \overset{*}{\underset{G}{\Rightarrow}} x$*, then for the height of the tree* T*,* height T *we have* $\text{height}(T) \geqslant \log_2 |x|$*.*

Proof

The proof is by strong induction on the length n of the derivation $X \overset{*}{\underset{G}{\Rightarrow}} x$.
Since G is a grammar in Chomsky normal form its productions have either the form $X \to YZ$ or $X \to a$.
For $n = 1$ the derivation is $X \underset{G}{\Rightarrow} x$, so $x = a$ and $1 \geqslant 0 = \log_2 |x|$.

(Proof cont'd)

If the first production applied in the derivation is $X \rightarrow YZ$, then

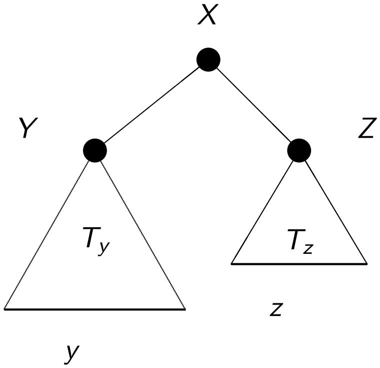
$$X \underset{G}{\Rightarrow} YZ \overset{*}{\underset{G}{\Rightarrow}} x,$$

and x can be factored as $x = yz$, where $Y \overset{*}{\underset{G}{\Rightarrow}} y$ and $Z \overset{*}{\underset{G}{\Rightarrow}} z$ are shorter derivations. Let T_y and T_z be the derivation trees that correspond to these derivations.

(Proof cont'd)

Note that height $T = 1 + \max\{\text{height}(T_y), \text{height}(T_z)\}$.

Derivation tree T



Note that for $a, b > 0$ we have

$$1 + \max\{log_2 a, \log_2 b\} \geqslant \log_2(a + b).$$

The equality takes place when $a = b$.
The inequality is equivalent to $\max\{log_2 a, \log_2 b\} \geqslant \log_2 \frac{a+b}{2}$, which is clearly true.
Therefore,

$$\begin{aligned} \text{height}(T) &= 1 + \max\{\text{height}(T_y), \text{height}(T_z)\} \\ &\geqslant 1 + \max\{\log_2 |y|, \log_2 |z|\} \\ &\quad \text{(by inductive hypothesis)} \\ &\geqslant \log_2(|y| + |z|) = \log_2 |x|, \end{aligned}$$

which concludes the proof.

An important observation

Let T be a derivation tree for $X \overset{*}{\underset{G}{\Rightarrow}} x$. If ϖ is a path of length ℓ in T that joins the root to a leaf, then there are ℓ nodes labeled by non-terminal symbols.

Theorem

Let G be a context-free grammar. There exists a number $n_G \in \mathbb{N}$ such that if $w \in L(G)$ and $|w| \geqslant n_G$, then we can write

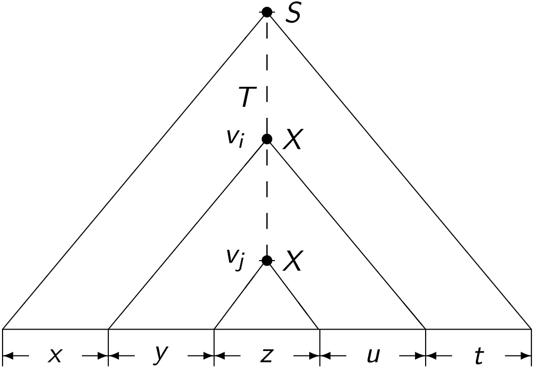
$$w = xyzut$$

such that $|y| \geqslant 1$ or $|u| \geqslant 1$, $|yzu| \leqslant n_G$ and $xy^n zu^n t \in L(G)$ for all $n \in \mathbb{N}$.

Proof

Suppose initially that $\lambda \notin L(G)$ and that G is a grammar in Chomsky normal form.
If $x \in L(G)$ and $|x| \geqslant 2^{1+|A_N|}$, the tree T has height at least $1 + |A_N|$. Let ϖ be a longest path in this tree; its length is at least $|A_N| + 1$, so the labels of non-final nodes on this path (which are non-terminals) cannot all be distinct.

Let X be the last occurrence of a symbol that is repeated on the path ϖ of a derivation tree T for $S \overset{*}{\underset{G}{\Rightarrow}} w$.



(Proof cont'd)

Since there is only one repeated symbol on the path that originates in the first X in this tree the length of this path is at most $|A_N| + 1$. Therefore, the word yzu for which $X \overset{*}{\underset{G}{\Rightarrow}} yzu$ cannot be longer than $2^{|A_N|+1} = n_L$.

The existence of the tree T implies the existence of the following derivations:

- $S \overset{*}{\underset{G}{\Rightarrow}} xXt$;
- $X \overset{*}{\underset{G}{\Rightarrow}} yXu$;
- $X \overset{*}{\underset{G}{\Rightarrow}} z$.

(Proof cont'd)

A repeated application of the second derivation yields:

$$S \overset{*}{\underset{G}{\Rightarrow}} xXt \quad \overset{*}{\underset{G}{\Rightarrow}} xyXut \overset{*}{\underset{G}{\Rightarrow}} xy^2Xu^2t \\ \cdots \quad \overset{*}{\underset{G}{\Rightarrow}} xy^nXu^nt \overset{*}{\underset{G}{\Rightarrow}} xy^nzu^nt,$$

hence $xy^nzu^nt \in L$ for $n \in \mathbb{N}$, which completes the proof.

Example

Let $A = \{a, b, c\}$ and let $L = \{a^n b^n c^n \mid n \in \mathbb{N}\}$. The language L is not context-free.

Suppose that L were a context-free language. Then, there is $n_G \in \mathbb{N}$ satisfying the properties stated in the Pumping Lemma.

Let $w = a^{n_G} b^{n_G} c^{n_G}$. Clearly, $|w| = 3n_G > n_G$, so $w = xyzut$ for some x, y, z, u, t such that $|y| \geqslant 1$ or $|u| \geqslant 1$, $|yzu| \leqslant n_G$ and $xy^n zu^n t \in L(G)$ for all $n \in \mathbb{N}$.

(Example cont'd)

Neither y nor u may contain more than one type of symbol; indeed, if y contained both as and bs, then we could write $y = y'a \cdots ab \cdots by''$. Since

$$xy^2zu^2t = xy'a \cdots ab \cdots by''y'a \cdots ab \cdots by''zu^2t$$

we obtain a contradiction, since no b symbol may precede an a symbol in any word of $L(G)$.

(Example cont'd)

A similar argument works for u. Thus, each y and u may contain only one kind of symbol.

Consequently, pumping y and u would violate the condition $n_a(w) = n_b(w) = n_c(w)$ satisfied by all words $w \in L(G)$. This shows that $L(G)$ does not satisfy the conditions of Pumping Lemma, hence $L(G)$ is not context-free.

We know that the class of context-free languages is closed with respect to union, product and Kleene closure.
Next we examine closure (and non-closure) properties of this class of languages with respect to other operations.

Theorem

The class $\mathcal{L}_2$ is not closed with respect to intersection.

Proof

Consider the context-free grammars

$$\begin{aligned} G_0 &= (\{S, X, Y\}, \{a, b, c\}, S, \{S \to XY, X \to aXb, X \to \lambda, \\ &\quad Y \to cY, Y \to \lambda\}) \\ G_1 &= (\{S, X, Y\}, \{a, b, c\}, S, \{S \to XY, X \to aX, X \to \lambda, \\ &\quad Y \to bYc, Y \to \lambda\}). \end{aligned}$$

It is easy to see that

$$\begin{aligned} L(G_0) &= \{a^m b^m c^p \mid m, p \in \mathbb{N}\} \\ L(G_1) &= \{a^m b^p c^p \mid m, p \in \mathbb{N}\}. \end{aligned}$$

Therefore, both $L_0 = \{a^m b^m c^p \mid m, p \in \mathbb{N}\}$ and $L_1 = \{a^m b^p c^p \mid m, p \in \mathbb{N}\}$ are context-free languages. On the other hand, since $L_0 \cap L_1 = \{a^n b^n c^n \mid n \in \mathbb{N}\}$, $L_0 \cap L_1$ does not belong to $\mathcal{L}_2$.

Corollary

The class $\mathcal{L}_2$ is not closed with respect to complement.

Proof.

Indeed, suppose that $\mathcal{L}_2$ were closed with respect to complement. Since $\mathcal{L}_2$ is closed with respect to union, and intersection can be expressed through union and complement (using De Morgan's equalities) this would imply that $\mathcal{L}_2$ is closed with respect to intersection. □

Theorem

If L is a context-free language and R is a regular language, then $L \cap R$ is a context-free language.

Proof

Without loss of generality, we may assume that both L and R are languages over the same alphabet A_T. Initially, we also assume that neither L nor R contains the null word.

Suppose that $L = L(G)$, where $G = (A_N, A_T, S_0, P)$ is a λ-free context-free grammar and $R = L(\mathcal{M})$, where $\mathcal{M}$ is a deterministic finite automaton $\mathcal{M} = (A_T, Q, \delta, q_0, F)$.

(Proof cont'd)

Define the context-free grammar $G' = (A'_N, A_T, S', P')$ as follows. The nonterminal alphabet A'_N consists of the new initial symbol S' together with $(|A_N| + |A_T|)|Q|^2$ new symbols of the form $s^{qq'}$ for every symbol $s \in A_N \cup A_T$ and every pair of states (q, q') of the automaton $\mathcal{M}$. The set P' consists of the following productions:

- $S' \to S^{q_0 q}$ for every final state q of $\mathcal{M}$;
- $X^{qq'} \to s_0^{qq_1} s_1^{q_1 q_2} \cdots s_{n-1}^{q_{n-1}q'}$ for every production $X \to s_0 \ldots s_{n-1}$ in P and every sequence of states $(q_1, \ldots, q_{n-1})$;
- $a^{qq'} \to a$ for every terminal symbol a of G such that $\delta(q, a) = q'$ in $\mathcal{M}$.

(Proof cont'd)

We claim that if $s^{qq'} \overset{n}{\underset{G'}{\Rightarrow}} x$ for some $x \in A_T^*$, then $\delta^*(q, x) = q'$ in $\mathcal{M}$ and that, if $s \in A_N \cup A_T$, then $s \overset{*}{\underset{G}{\Rightarrow}} x$. The argument is by induction on $n \geqslant 1$. If $n = 1$ we have $s = x = a$ for some $a \in A_T$ and $\delta(q, a) = q'$ and the claim is clearly satisfied.

(Proof cont'd)

Suppose that the claim holds for derivations of length less than n, and let $s^{qq'} \overset{n}{\underset{G'}{\Rightarrow}} x$ be a derivation of length n. If we write the first step of this derivation explicitly, we obtain

$$s^{qq'} \underset{G'}{\Rightarrow} s_0^{qq_1} s_1^{q_1q_2} \cdots s_{n-1}^{q_{n-1}q'} \overset{*}{\underset{G'}{\Rightarrow}} x.$$

Therefore, we have the production $s \to s_0 \ldots s_{n-1}$ in P, and we can write x as $x = x_0 \cdots x_{n-1}$, such that we have the derivations

$$\begin{array}{ll} s_0^{qq_1} \overset{*}{\underset{G}{\Rightarrow}} x_0 & s_1^{q_1q_2} \overset{*}{\underset{G}{\Rightarrow}} x_1 \\ & \vdots \\ s_{i-1}^{q_{i-1}q_i} \overset{*}{\underset{G}{\Rightarrow}} x_{i-1} & \cdots s_{n-1}^{q_{n-1}q'} \overset{*}{\underset{G}{\Rightarrow}} x_{n-1} \end{array}$$

that are all shorter than n.

(Proof cont'd)

By the inductive hypothesis we have

$$\delta^*(q, x_0) = q_1, \delta^*(q_1, x_1) = q_2, \ldots, \delta^*(q_{n-1}, x_{n-1}) = q',$$

so $\delta^*(q, x_0 \cdots x_{n-1}) = \delta^*(q, x) = q'$. Also, if s_i is a nonterminal, then $s_i \overset{*}{\underset{G}{\Rightarrow}} x_i$; otherwise, that is, if $s_i \in A_T$, we have $s_i = x_i$, so $s_i \overset{*}{\underset{G}{\Rightarrow}} x_i$ for $0 \leqslant i \leqslant n-1$. This allows us to construct the derivation

$$s \underset{G}{\Rightarrow} s_0 \ldots s_{n-1} \overset{*}{\underset{G}{\Rightarrow}} x_0 \ldots x_{n-1},$$

which justifies our claim.

(Proof cont'd)

We prove the theorem by showing that $L(G') = L \cap R$. Suppose that $x \in L(G')$. We have the derivation $S' \overset{*}{\underset{G'}{\Rightarrow}} x$, that is, $S' \underset{G'}{\Rightarrow} S^{q_0 q} \overset{*}{\underset{G'}{\Rightarrow}} x$. By the previous claim this implies both $S \overset{*}{\underset{G}{\Rightarrow}} x$ and $\delta^*(q_0, x) = q$. Thus, $x \in L \cap R$.

(Proof cont'd)

Conversely, suppose that $x = a_0 \cdots a_{n-1} \in L \cap R$. We have the derivation $S \underset{G}{\overset{*}{\Rightarrow}} x$, so in G' we can write

$$S' \underset{G}{\Rightarrow} S^{q_0 q} \underset{G}{\overset{*}{\Rightarrow}} a_0^{q_0 q_1} a_1^{q_1 q_2} \cdots a_{n-1}^{q_{n-1} q},$$

for some final state q and any states $q_1, \ldots, q_{n-1}$. We can select these intermediate states such that $\delta(q_i, a_i) = q_{i+1}$ for $0 \leqslant i \leqslant n-2$ and $\delta(q_{n-1}, a_{n-1}) = q'$. Therefore, we have the productions in P':

$$a_0^{q_0 q_1} \to a_0, a_1^{q_1 q_2} \to a_1, \ldots, a_{n-1}^{q_{n-1} q'} \to a_{n-1}$$

(Proof cont'd)

This implies the existence of the derivation $S' \overset{*}{\underset{G}{\Rightarrow}} a_0 a_1 \cdots a_{n-1}$, so $x \in L(G')$.
If $\lambda \in R$ or $\lambda \in L$ we consider the regular language $R' = R - \{\lambda\}$ and the context-free language $L' = L - \{\lambda\}$. By the previous argument we can construct a λ-free context-free grammar G' such that $L(G') = L' \cap R'$. If $\lambda \notin L \cap R$, then $L \cap R = L' \cap R'$ and this shows that $L \cap R$ is context-free.
If $\lambda \in L \cap R$, we have $L \cap R = (L' \cap R') \cup \{\lambda\}$. Then, starting from $G' = (A'_N, A_T, S', P')$ we construct the context-free grammar

$$G'' = (A'_N \cup \{S''\}, A_T, S'', P' \cup \{S'' \to S', S'' \to \lambda\})$$

and we have $L \cap R = L(G'')$.