Context-Free languages (part II)

Prof. Dan A. Simovici

UMB

Definition

A context-free grammar $G = (A_N, A_T, S, P)$ is in *Chomsky normal form* if all productions are either of the form $X \to YZ$ or of the form $X \to a$, where $X, Y, Z \in A_N$ and $a \in A_T$.

If G is in Chomsky normal form, then G is λ-free, so $\lambda \notin L(G)$.

Theorem

For every context-free grammar G such that $\lambda \notin L(G)$ there is an equivalent grammar in Chomsky normal form.

Proof.

We can assume that G is a λ-free grammar, G has no chain productions and that every production that contains a terminal symbol is of the form $X \to a$.

Thus, the productions of G have either the form $X \to a$ or the form $X \to X_{i_0} \cdots X_{i_{k-1}}$ with $k \geq 2$. □

(Proof cont'd)

Productions of the form $X \rightarrow a$ or $X \rightarrow X_{i_0}X_{i_1}$ already conform to Chomsky normal form. If $\pi : X \rightarrow X_{i_0} \cdots X_{i_{k-1}}$ is a production of P with $k \geq 3$, consider $k-2$ new nonterminals $Z_0^\pi, \ldots, Z_{k-3}^\pi$ and the productions

$$X \rightarrow X_{i_0} Z_0^\pi, Z_0^\pi \rightarrow X_{i_1} Z_1^\pi, \cdots, Z_{k-3}^\pi \rightarrow X_{i_{k-2}} X_{i_{k-1}}$$

Define the grammar $G' = (A_N \cup A', A_T, S, P')$, where A' consists of all symbols Z_ℓ^π, and P' consists of all productions of the form $X \rightarrow a$ or $X \rightarrow X_{i_0}X_{i_1}$, and of productions obtained from productions of P having the form $X \rightarrow X_{i_0} \cdots X_{i_{k-1}}$ with $k \geq 3$, by applying the method described above. It is easy to see that G' is equivalent to G and that G' is in Chomsky normal form.

Example

Let $G = (\{S_0, S_1, S_2\}, \{a, b\}, S_0, P)$ be the context-free grammar, where P contains the following productions:

$$S_0 \to aS_2, S_0 \to bS_1, S_1 \to a, S_1 \to aS_0, S_1 \to bS_1S_1,$$
$$S_2 \to b, S_2 \to bS_0, S_2 \to aS_2S_2.$$

By introducing the new nonterminal symbols X_a, X_b we obtain the grammar $G_1 = (\{S_0, S_1, S_2, X_a, X_b\}, \{a, b\}, S_0, P_1)$, where P_1 consists of

$$S_0 \to X_aS_2, S_0 \to X_bS_1, S_1 \to a, S_1 \to X_aS_0, S_1 \to X_bS_1S_1,$$
$$S_2 \to b, S_2 \to X_bS_0, S_2 \to X_aS_2S_2, X_a \to a, X_b \to b.$$

(Example cont'd)

G_1 is equivalent to G, has no chain productions and every production that contains a terminal symbol is of the form $X \to a$. This grammar has two productions, $S_1 \to X_b S_1 S_1$ and $S_2 \to X_a S_2 S_2$, that violate Chomsky normal form, so we introduce the new nonterminals Z_0, Z_1.
Applying the technique introduced before to these productions results in the set of productions P' given by:

$$\begin{array}{l} S_0 \to X_a S_2,\ S_0 \to X_b S_1,\ S_1 \to a,\ S_1 \to X_a S_0, \\ S_1 \to X_b Z_0,\ Z_0 \to S_1 S_1,\ S_2 \to b,\ S_2 \to X_b S_0, \\ S_2 \to X_a Z_1,\ Z_1 \to S_2 S_2,\ X_a \to a,\ X_b \to b. \end{array}$$

The resulting grammar $G' = (\{S_0, S_1, S_2, X_a, X_b, Z_0, Z_1\}, \{a, b\}, S_0, P')$ is in Chomsky normal form and is equivalent to G.

Using Chomsky normal form we can prove an important decidability result for the class $\mathcal{L}_2$. To this end, we need the following technical result relating the length of a word to the length of its derivation.

Lemma

Let $G = (A_N, A_T, S, P)$ be a context-free grammar in Chomsky normal form. Then, if $S \overset{*}{\underset{\alpha}{\Rightarrow}} x$ we have $|\alpha| \leq 2|x| - 1$.

Proof

We prove a slightly stronger statement, namely that if $X \overset{*}{\underset{\alpha}{\Rightarrow}} x$ for some $X \in A_N$, then $|\alpha| \leq 2|x| - 1$.
The argument is by induction on $n = |x| \geq 1$. If $n = 1$, we have $x = a$ for $a \in A_T$ and the derivation $X \overset{*}{\underset{\alpha}{\Rightarrow}} x$ consists in the application of the production $\pi : X \to a$. Therefore, $|\alpha| = 1$ and the inequality is satisfied.

(Proof cont'd)

Suppose that the statement holds for words of length less than n, and let $x \in L(G)$ be a word such that $|x| = n$, where $n > 1$. Let the first production applied be $X \to YZ$; then we can write $x = uv$, there $Y \overset{*}{\underset{\beta}{\Rightarrow}} u$ and $Z \overset{*}{\underset{\gamma}{\Rightarrow}} v$ and $|\alpha| = |\beta| + |\gamma| + 1$, because the productions used in the last two derivations are exactly the ones used in $X \overset{*}{\underset{\alpha}{\Rightarrow}} x$. Applying the inductive hypothesis we obtain

$$|\alpha| = |\beta| + |\gamma| + 1 \leq 2|u| - 1 + 2|v| - 1 + 1 = 2(|u| + |v|) - 1 = 2|x| - 1.$$

Theorem

There is an algorithm to determine for a context-free grammar $G = (A_N, A_T, S, P)$ and a word $x \in A_T^*$ whether or not $x \in L(G)$.

Proof.

Construct a grammar G' equivalent to G such that one of the following two cases occurs:

1. if $\lambda \notin L(G)$ then G' is λ-free;
2. if $\lambda \in L(G)$ then G' contains a unique erasure production $S' \to \lambda$, where S' is the start symbol of G' and S' does not occur in any right member of any production of G'.

□

(Proof cont'd)

If $x = \lambda$, then $x \in L(G)$ if and only if $S \to \lambda$ is a production in G'. Suppose that $x \neq \lambda$. Let G_1 be a context-free grammar in Chomsky normal form such that $L(G_1) = L(G') - \{\lambda\} = L(G) - \{\lambda\}$. We have $x \in L(G_1)$ if and only if $x \in L(G)$. By the previous Lemma, if $S \overset{*}{\underset{\alpha}{\Rightarrow}} x$, then $|\alpha| \leq 2|x| - 1$, so we can decide if $x \in L(G)$ by listing all derivations of length at most $2|x| - 1$.

- As an alternative to writing a sequence of derivation steps, we consider describe context-free derivations using labeled ordered trees, so-called derivation trees.
- The labels of the leaves of an A-labeled ordered tree, when read from left-to-right, spell out a word in A^*.

Definition of Derivation Trees

Definition

Let $G = (A_N, A_T, S, P)$ be a λ-free context-free grammar, and let $d = (\gamma_0, \ldots, \gamma_m)$ be a derivation in G, where $\gamma_0 = X \in A_N$ and $\gamma_i \in (A_N \cup A_T)^*$ for $0 \leq i \leq m$. Let $A = A_N \cup A_T$.
The *derivation tree of the derivation d* is an A-labeled, ordered tree $\mathtt{T}_d$ defined inductively as follows:

Def. cont'd

1. If $m = 0$, then $\mathtt{T}_d$ consists of only one node labeled by $(0, X)$.
2. Suppose that $m \geq 1$ and that $\gamma_1 = X_0 \ldots X_{n-1}$, where $X_0 \ldots X_{n-1} \in (A_N \cup A_T)^*$. Let $\mathtt{T}_i$ be the A-labeled ordered tree that corresponds to the derivation $(X_i, \ldots, \alpha_i)$ for $0 \leq i \leq n-1$, where $\alpha = \alpha_0 \cdots \alpha_{n-1}$. Then, $\mathtt{T}_d$ is $\langle \mathtt{T}_0, \ldots, \mathtt{T}_{n-1}; X \rangle$.

The *set of derivation trees of G* is the set

$$\text{TREES}(G) = \{\mathtt{T}_d \mid d \text{ is a derivation in } G\}.$$

A derivation tree $\mathtt{T}_d \in \text{TREES}(G)$ is *complete* if $\text{word}(\mathtt{T}_d) \in A_T^*$, i.e. if all its leaves are labeled by terminal symbols of the grammar.
The set of complete derivation trees of G is denoted by $\text{TREES}_c(G)$.

Example

Let

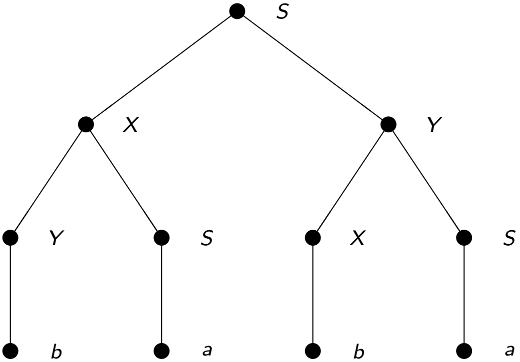
$$G = (\{S, X, Y\}, \{a, b\}, S, \{S \to XY, S \to a, X \to YS, \\ Y \to XS, X \to b, Y \to b\})$$

be a context-free grammar in Chomsky normal form. The derivation tree of

$$S \Rightarrow XY \Rightarrow YSY \Rightarrow YSXS \Rightarrow bSXS \Rightarrow baXS \Rightarrow babS \Rightarrow baba$$

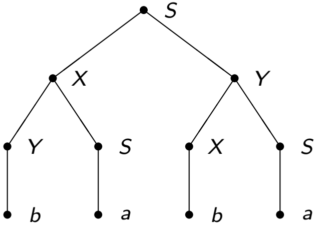
is given next:

$$S \Rightarrow XY \Rightarrow YSY \Rightarrow YSXS \Rightarrow bSXS \Rightarrow baXS \Rightarrow babS \Rightarrow baba$$



Every derivation in a context-free grammar $G = (A_N, A_T, S, P)$ is described by a derivation tree. Conversely, if $\mathtt{T}$ is a derivation tree such that $\text{word}(\mathtt{T}) = x \in A_T^*$ then, in general, several distinct derivations exist for the word x.

Example



This derivation tree also describes the derivation: $S \Rightarrow XY \Rightarrow XXS \Rightarrow XXa \Rightarrow YSXa \Rightarrow bSXa \Rightarrow baXa \Rightarrow baba$ is the same grammar $G = (\{S, X, Y\}, \{a, b\}, S, \{S \to XY, S \to a, X \to YS, Y \to XS, X \to b, Y \to b\})$.

Theorem

Let $G = (A_N, A_T, S, P)$ be a context-free grammar, and let $\mathtt{T} \in \mathrm{TREES}_c(G)$ be a complete derivation tree whose root is labeled by X, where the word spelled by $\mathtt{T}$, $\mathrm{word}(\mathtt{T}) = u \in A_T^*$. There is a unique leftmost (rightmost) derivation $X \overset{*}{\underset{G,left}{\Rightarrow}} u$. Moreover, the lengths of the leftmost and the rightmost derivations equal the number of internal nodes of $\mathtt{T}$.

Proof

The argument for leftmost derivations is by induction on the height of $\mathtt{T}$. If $\text{height}(\mathtt{T}) = 1$, then the derivation that corresponds to $\mathtt{T}$ is (X, u), which is an one-step leftmost derivation.
Suppose that the statement holds for complete derivation trees of height less than n, and let $\mathtt{T}$ be a complete derivation tree in G such that $\text{height}(\mathtt{T}) = n$. Then, $\mathtt{T} = \langle \mathtt{T}_0, \ldots, \mathtt{T}_{k-1}; X\rangle$, where $\text{height}(\mathtt{T}_i) < n$ for $0 \leq i \leq k-1$. Also, the root of $\mathtt{T}_i$ is labeled by the symbol $X_i \in A_N \cup A_T$ and its leaves are labeled by the terminal word u_i for $0 \leq i \leq k-1$, where $u_0 \cdots u_{k-1} = u$.

(Proof cont'd)

By the inductive hypothesis, for each of the trees $\mathtt{T}_i$, there is a unique leftmost derivation d_i:

$$X_i \Rightarrow w_{i0} \Rightarrow \cdots \Rightarrow w_{i\ell_i - 1} = u_i$$

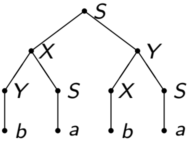
and the length of d_i is equal to the number of internal nodes of $\mathtt{T}_i$ for $0 \leq i \leq k-1$.
Then, we obtain the following leftmost derivation that corresponds to $\mathtt{T}$:

$$\begin{aligned} X &\Rightarrow X_0 X_1 \cdots X_{k-1} \\ &\Rightarrow w_{00} X_1 \cdots X_{k-1} \Rightarrow \cdots \Rightarrow u_0 X_1 \cdots X_{k-1} \\ &\Rightarrow u_0 w_{10} \cdots X_{k-1} \Rightarrow \cdots \Rightarrow u_0 u_1 \cdots X_{k-1} \\ &\vdots \\ &\Rightarrow u_0 u_1 \cdots w_{k-1\,0} \Rightarrow \cdots \Rightarrow u_0 u_1 \cdots u_{k-1}. \end{aligned}$$

(Proof cont'd)

If d is a leftmost derivation for T, then it must expand the nonterminals symbol $X_{i_0}, \ldots, X_{i_p}$ that occur in $X_0 \cdots X_{k-1}$. Thus, the derivation d must use the productions that occur in the leftmost derivations $d_{i_0}, \ldots, d_{i_{k-1}}$, respectively, in that order. This shows that the lefmost derivation is unique and the length of this derivation equals the number of internal nodes of T.

Example



For the derivation tree

$$S \Rightarrow XY \Rightarrow YSY \Rightarrow bSY \Rightarrow baY \Rightarrow baXS \Rightarrow babS \Rightarrow baba$$

is a leftmost derivation.

(Example cont'd)

The derivation

$$S \Rightarrow XY \Rightarrow XXS \Rightarrow XXa \Rightarrow Xba$$
$$\Rightarrow YSba \Rightarrow Yaba \Rightarrow baba$$

is the rightmost derivations.

If G is a context-free grammar and $x \in L(G)$, several distinct derivation trees may exist for x. In some cases, a considerable number of such distinct trees may exist.

Example

Let $G = (\{S\}, \{a\}, S, \{S \rightarrow SS, S \rightarrow a\})$ be a context-free grammar. It is not difficult to see that the language generated by G is $L(G) = \{a^m \mid m \geq 1\}$. Denote by $C(n)$ the number of derivation trees that describe derivations of the form $S \overset{*}{\underset{G}{\Rightarrow}} a^{n+1}$. We have $C(0) = 1$, and

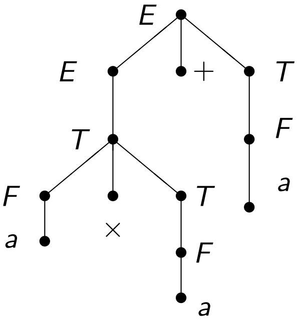
$$C(n) = \sum_{j=0}^{n-1} C(j)C(n-1-j),$$

It is possible to prove that $C(n) = \Theta\left(\frac{4^n}{n^{1.5}}\right)$.

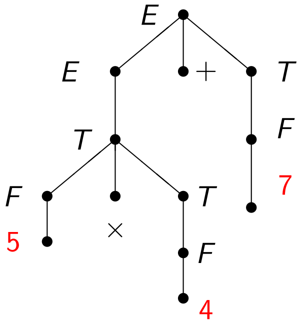
Derivation trees for arithmetic expressions relect implicitely the priority order of arithmetic operations.
Consider the context-free grammar

$$\begin{aligned} G &= (\{E, T, F\}, \{+, \times, (,)\}, E, \{E \to T, E \to E + T, \\ &\quad T \to F, T \to F \times T, F \to a, F \to (E)\}). \end{aligned}$$

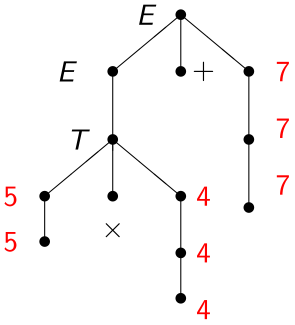
Derivation Tree for $a \times a + a$



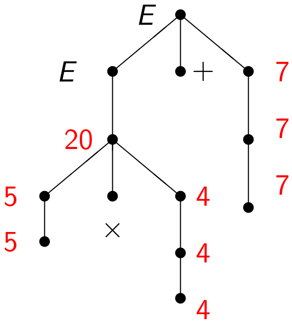
Computation of $5 \times 4 + 7$



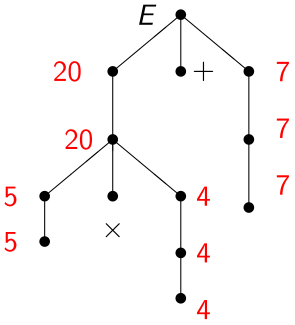
Computation of $5 \times 4 + 7$



Computation of $5 \times 4 + 7$



Computation of $5 \times 4 + 7$



Computation of $5 \times 4 + 7$