

Grammars (part II)

Prof. Dan A. Simovici

UMB

- 1 Elimination of erasure productions in context-free grammars
- 2 Elimination of Chain Productions

Theorem

For every context-free grammar G , there is a context-free, λ -free grammar G' such that $L(G') = L(G) - \{\lambda\}$.

Proof.

Let $G = (A_N, A_T, S_0, P)$ be a context-free grammar. Consider the sequence $U_0, \dots, U_m, \dots$ of subsets of A_N defined by

$$\begin{aligned} U_0 &= \{X \mid X \in A_N \text{ and } X \rightarrow \lambda \in P\}, \\ U_{m+1} &= U_m \cup \{X \in A_N \mid X \rightarrow \alpha \in P \text{ for some } \alpha \in U_m^*\}, \end{aligned}$$

for $m \in \mathbb{N}$.

Since $U_0 \subseteq U_1 \subseteq \dots \subseteq A_N$, there is $k \in \mathbb{N}$ such that $U_k = U_{k+1}$. □

(Proof cont'd)

A simple argument (by induction on $h \geq 1$) shows that $U_k = U_{k+h}$ for every $h \geq 1$.

The **base step** is immediate.

Suppose that $U_k = U_{k+h}$ and let $X \in U_{k+h+1}$. If $X \in U_{k+h}$, then $X \in U_k$ by the inductive hypothesis. Otherwise, there is a production $X \rightarrow \alpha \in P$ such that $\alpha \in U_{k+h}^*$. By the inductive hypothesis, $\alpha \in U_k^*$, so $X \in U_{k+1} = U_k$. Therefore, $U_{k+h+1} = U_k$.

We claim that $X \xRightarrow{+}_G \lambda$ if and only if $X \in U_k$.

We prove by **strong induction** on $p \geq 1$ that if $X \xRightarrow{p}_G \lambda$, then $X \in U_k$.

For $p = 1$, if $X \xRightarrow{1}_G \lambda$, then $X \in U_0$ and $U_0 \subseteq U_k$.

(Proof cont'd)

Suppose that the statement is true for derivations $X \xRightarrow[G]{+} \lambda$ of length no greater than p and let $X \xRightarrow[G]{p+1} \lambda$. The first production applied in this derivation must have the form $X \rightarrow X_{i_1} \cdots X_{i_q}$; therefore, we have

$$X_{i_1} \cdots X_{i_q} \xRightarrow[G]{p} \lambda.$$

Hence, $X_{i_\ell} \xRightarrow[G]{p_\ell} \lambda$, where $p_\ell \leq p$ for $1 \leq \ell \leq q$. By the inductive hypothesis, we have $X_{i_\ell} \in U_k$, so $X_{i_1} \cdots X_{i_q} \in (U_k)^*$, which implies $X \in U_{k+1} = U_k$.

(Proof cont'd)

Conversely, it is easy to prove (by induction on n) that for every $X \in U_n$ we have $X \xrightarrow{+}_G \lambda$. We leave this argument to the reader. From this it follows that if $\theta \in U_k^*$, then $\theta \xrightarrow{*}_G \lambda$.

Consider now the set of productions P' , where

$$P' = \{X \rightarrow \alpha' \mid \alpha' \neq \lambda, \text{ there is } X \rightarrow \alpha \in P \text{ and } \alpha' \text{ is obtained from } \alpha \text{ by erasing 0 or more symbols from } U_k\}.$$

(Proof cont'd)

If G' is the context-free grammar $G' = (A_N, A_T, S_0, P')$, then $L(G') = L(G) - \{\lambda\}$. Indeed, suppose that $X \xrightarrow[p]{G'} \gamma$. Clearly, $\gamma \neq \lambda$ since G' has no erasure productions. We prove, by strong induction on p , that we have $X \xrightarrow[G]{*} \gamma$.

For $p = 0$, the statement is trivially true.

(Proof cont'd)

Assume that it holds for derivations of length less than or equal to p , and let $X \xRightarrow[p+1]{G'} \gamma$. If the first production applied in this derivation is

$X \rightarrow X_{i_0} \cdots X_{i_{h-1}}$, then $\gamma = \gamma_0 \cdots \gamma_{h-1}$, where $X_{i_j} \xRightarrow[p_j]{G'} \gamma_j$, $p_j \leq p$, for

$0 \leq j \leq h-1$. By the inductive hypothesis we have $X_{i_j} \xRightarrow{*}_G \gamma_j$ for

$0 \leq j \leq h-1$.

(Proof cont'd)

Furthermore, assume that the production $X \rightarrow X_{j_0} \cdots X_{j_{h-1}}$ was obtained from the production $X \rightarrow \theta_0 X_{j_0} \theta_1 \cdots X_{j_{h-1}} \theta_h$ from P , where $\theta_0, \dots, \theta_h \in (U_k)^*$. Our previous discussion allows us to infer the existence of the derivations $\theta_q \xRightarrow[G]{*} \lambda$ for $0 \leq q \leq h$. By combining the derivations obtained above, we have

$$\begin{array}{rcl}
 X & \xRightarrow[G]{} & \theta_0 X_{j_0} \theta_1 \cdots X_{j_{h-1}} \theta_h \\
 & \xRightarrow[G]{*} & X_{j_0} \cdots X_{j_{h-1}} \\
 & \xRightarrow[G]{*} & \gamma_0 \cdots \gamma_{h-1} = \gamma.
 \end{array}$$

This implies $L(G') \subseteq L(G) - \{\lambda\}$.

(Proof cont'd)

To prove the converse inclusion, consider a derivation $X \xrightarrow[G]{p} \gamma$, where $\gamma \neq \lambda$.

We claim that $X \xrightarrow[G']{*} \gamma$. The argument is by strong induction on $p \geq 0$.

The case $p = 0$ is trivially true. Assume that the statement holds for derivations of length of no more than p , and let $X \xrightarrow[G]{p+1} \gamma$, where $\gamma \neq \lambda$.

Let $\beta = X_{j_0} \cdots X_{j_{l-1}}$ be the word that follows X in the previous derivation, that is, $X \Rightarrow_G \beta \xrightarrow[G]{p} \gamma$.

(Proof cont'd)

We can write:

$$\gamma = \gamma_0 \cdots \gamma_{l-1},$$

where $X_{j_m} \xrightarrow[p_m]{G} \gamma_m$ and $p_m \leq p$ for $0 \leq m \leq l-1$.

If $\gamma_m \neq \lambda$, by the inductive hypothesis, we have $X_{j_m} \xrightarrow[G']{*} \gamma_m$. On the other hand, if $\gamma_m = \lambda$, we have $X_{j_m} \in U_k$. Let

$$\{h_0, \dots, h_{q-1}\} = \{h \mid 0 \leq h \leq l-1 \text{ and } \gamma_h \neq \lambda\}.$$

The definition of P' implies that we have the production $X \rightarrow X_{j_{h_0}} \cdots X_{j_{h_{q-1}}}$ in P' . Therefore,

$$X \xRightarrow[G']{} X_{j_{h_0}} \cdots X_{j_{h_{q-1}}} \xrightarrow[G']{*} \gamma_{h_0} \cdots \gamma_{h_{q-1}} = \gamma.$$

This implies $L(G) - \{\lambda\} \subseteq L(G')$.

Theorem

If G is a context-free grammar, then there is an equivalent context-free grammar 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 $S' \rightarrow \lambda$, where S' is the start symbol of G' , and S' does not occur in any right member of any production of G' .*

Proof

We have shown that for every context-free grammar G there is a context-free, λ -free grammar G_1 such that $L(G_1) = L(G) - \{\lambda\}$. If $\lambda \notin L(G)$, then the grammars G and G_1 are equivalent, and we can define G' as G_1 . This proves the first case of this theorem.

If $\lambda \in L(G)$, by the same theorem, we have the context-free, λ -free grammar $G_1 = (A_N, A_T, S_1, P)$ such that $L(G_1) = L(G) - \{\lambda\}$. Define the grammar G' by

$$G' = (A_N \cup \{S'\}, A_T, S', \{S' \rightarrow S_1, S' \rightarrow \lambda\} \cup P),$$

where S' is a new nonterminal symbol (i.e., that $S' \notin A_N$). It is immediate that G' satisfies the conditions of the second case of this Theorem and that $L(G') = L(G_1) \cup \{\lambda\} = L(G)$.

Example

Let $G = (\{S, X, Y, Z\}, \{a, b\}, S, \{S \rightarrow XYZ, X \rightarrow YZ, X \rightarrow aYb, X \rightarrow a, Y \rightarrow \lambda, Y \rightarrow b, Z \rightarrow \lambda, Z \rightarrow c\})$ be a context-free grammar that contains erasure productions. The sequence of subsets of $\{S, X, Y, Z\}$ is

$$U_0 = \{Y, Z\}, U_1 = \{Y, Z, X\}, U_2 = \{Y, Z, X, S\}, U_3 = U_2.$$

Therefore, the set of productions P' is given by

$$\begin{aligned} P' = \{ & S \rightarrow XYZ, S \rightarrow YZ, S \rightarrow XZ, S \rightarrow XY, S \rightarrow X, S \rightarrow Y, \\ & S \rightarrow Z, X \rightarrow YZ, X \rightarrow Y, X \rightarrow Z, X \rightarrow aYb, X \rightarrow ab, \\ & X \rightarrow a, Y \rightarrow b, Z \rightarrow c \} \end{aligned}$$

Observe that the productions of P' are obtained by erasing zero, one, or more of the symbols X, Y, Z from the rules of P .

The previous theorem shows that it is possible to limit the erasure productions in context-free grammars that generate a language L to a single production that has the start symbol as its left member, without restricting the generality.

Corollary

Every context-free language is a context-sensitive language; in other words, $\mathcal{L}_2 \subseteq \mathcal{L}_1$.

Proof.

This is an immediate consequence of a previous theorem and the definitions of $\mathcal{L}_1$ and $\mathcal{L}_2$. □

Definition

Let $G = (A_N, A_T, S, P)$ be a context-free grammar. A *chain production* is a production $X \rightarrow Y$, where $X, Y \in A_N$.

Theorem

Let $G = (A_N, A_T, S, P)$ be a context-free grammar. There is a context-free grammar G_1 such that G_1 is equivalent to G and G_1 does not contain chain productions.

Proof

We assume initially that G is λ -free. Let X be a nonterminal symbol. To eliminate productions of the form $X \rightarrow Y$ consider the following sequence of sets:

$$\begin{aligned} U_0^X &= \{X\} \\ U_{n+1}^X &= U_n^X \cup \{Z \in A_N \mid Y \rightarrow Z \in P \text{ for some } Y \in U_n^X\} \end{aligned}$$

It is clear that the sequence $U_0^X, \dots, U_n^X, \dots$ is an increasing sequence of subsets of A_N . The finiteness of A_N implies the existence of a number i such that $U_i^X = U_{i+1}^X$. Then, by induction on $\ell \geq 1$, we can easily prove that $U_i^X = U_{i+\ell}^X$ for $\ell \geq 1$.

We shall prove that $U_i^X = \{Z \in A_N \mid X \xRightarrow{*}_G Z\}$.

A straightforward argument by induction on n shows that

$$U_n^X \subseteq \{Z \in A_N \mid X \xRightarrow{*}_G Z\} \text{ for } n \in \mathbb{N}. \text{ In particular,}$$

$$U_i^X \subseteq \{Z \in A_N \mid X \xRightarrow{*}_G Z\}.$$

(Proof cont'd)

To prove the converse inclusion, we prove that if a derivation $X \xRightarrow[k]{G} Z$, then $Z \in U_i^X$. The argument is by induction on k . For $k = 0$, $Z = X$, and $Z \in U_0^X \subseteq U_i^X$, so the conclusion follows. Suppose that the statement holds for derivations of length k , and let $X \xRightarrow[k+1]{G} Z'$. Since the grammar has no erasure rules, we can write $X \xRightarrow[k]{G} Z \xrightarrow{G} Z'$. By the inductive hypothesis, $Z \in U_i^X$; the existence of the production $Z \rightarrow Z'$ implies that $Z \in U_{i+1}^X = U_i^X$. Thus, $\{Z \in A_N \mid X \xRightarrow{*}{G} Z\} \subseteq U_i^X$.

(Proof cont'd)

Denote the set $\{Z \in A_N \mid X \xRightarrow[G]{*} Z\}$ by U_*^X . The context-free grammar $G_1 = (A_N, A_T, S, P_1)$ is defined by

$$P_1 = \{X \rightarrow \alpha \mid Z \rightarrow \alpha \in P \text{ for some } Z \in U_*^X \text{ and } \alpha \notin A_N\}.$$

It is clear that the grammar G_1 has no chain productions and is equivalent to G .

If G is not λ -free, then there exists an equivalent context-free grammar $G' = (A_N \cup \{S'\}, A_T, S', P' \cup \{S' \rightarrow \lambda\})$ where $S' \rightarrow \lambda$ is the unique erasure production of G' , and S' does not occur in any right member of any production of G' . The grammar $G'' = (A_N \cup \{S'\}, A_T, S', P')$ generates the language $L(G) - \{\lambda\}$. By applying the previous construction to G'' we obtain the grammar $G_1'' = (A_N \cup \{S'\}, A_T, S', P_1'')$ that has no chain rules and for which $L(G_1'') = L(G) - \{\lambda\}$. Then, the desired grammar G_1 is given by

$$G_1 = (A_N \cup \{S'\}, A_T, S_1, P_1'' \cup \{S' \rightarrow \lambda\}),$$

where S_1 is a new start symbol.

Example

The grammar

$$G = (\{S, X, Y\}, \{a, b, c\}, S, \{S \rightarrow X, S \rightarrow aX, X \rightarrow Y, X \rightarrow bY, S \rightarrow a, X \rightarrow b, Y \rightarrow c\})$$

is λ -free and contains some chain productions.

(Example cont'd)

We have $U_0^S = \{S\}$, $U_1^S = \{S, X\}$, $U_2^S = \{S, X, Y\}$, and $U_2^S = U_3^S = \dots$, so $U_*^S = \{S, X, Y\}$. Similar computations give $U_*^X = \{X, Y\}$ and $U_*^Y = \{Y\}$. The grammar

$$G_1 = (\{S, X, Y\}, \{a, b, c\}, S, \{S \rightarrow aX, S \rightarrow bY, S \rightarrow a, S \rightarrow b, S \rightarrow c, X \rightarrow c, X \rightarrow bY, X \rightarrow b, Y \rightarrow c\}).$$

is equivalent to G and has no chain productions.

Let $G = (A_N, A_T, S, P)$ be a context-free grammar, and let X be a nonterminal symbol. Denote by $L(G, X)$ the set of terminal words that can be generated from X in the grammar G , that is,

$$L(G, X) = \{x \in A_T^* \mid X \xrightarrow[G]{*} x\}.$$

Clearly, we have $L(G, S) = L(G)$.

Definition

Let $G = (A_N, A_T, S, P)$ be a context-free grammar. A symbol $s \in A_N \cup A_T$ is *accessible* if it occurs in a word $\alpha \in (A_N \cup A_T)^*$ such that $S \xRightarrow{*}_G \alpha$.

A symbol $X \in A_N$ is *productive* if $L(G, X) \neq \emptyset$.

Theorem

Let $G = (A_N, A_T, S, P)$ be a context-free grammar. There is a construction of an equivalent grammar $G' = (A'_N, A_T, S, P')$ such that $P' = \emptyset$ if $L(G) = \emptyset$, and if $L(G) \neq \emptyset$, then every symbol in A'_N is productive.

Proof

Define the sequence $U_0, \dots, U_n, \dots$ of subsets of A_N by

$$\begin{aligned} U_0 &= \{X \in A_N \mid X \rightarrow u \in P \text{ for some } u \in A_T^*\} \\ U_{n+1} &= U_n \cup \{X \in A_N \mid X \rightarrow \alpha \in P \text{ for some } \alpha \in (U_n \cup A_T)^*\} \end{aligned}$$

Note that $U_0 \subseteq U_1 \subseteq \dots \subseteq U_n \subseteq \dots \subseteq A_N$. Therefore, there is i such that $U_i = U_{i+1}$. An easy argument by induction on k shows that $U_i = U_{i+k}$ for $k \geq 1$. This part of the proof is left to the reader.

We claim that

$$\{X \in A_N \mid L(G, X) \neq \emptyset\} = U_i.$$

(Proof cont'd)

If $n = 0$, the conclusion follows from the definition of U_0 . Suppose that the inclusion holds for U_n and let $Y \in U_{n+1}$. If $Y \in U_n$ the conclusion is immediate. Otherwise, there is a production $Y \rightarrow \alpha$, where $\alpha = w_0 Z_0 w_1 Z_1 \cdots w_{p-1} Z_{p-1} w_p$, where $w_i \in A_T^*$ for $0 \leq i \leq p$ and $Z_j \in U_n$ for $0 \leq j \leq p-1$. By the inductive hypothesis, we have the derivations $Z_j \xRightarrow{*}_G z_j$, where $z_j \in A_T^*$ for $0 \leq j \leq p-1$. Thus, we obtain the derivation

$$Y \Rightarrow w_0 Z_0 w_1 Z_1 \cdots w_{p-1} Z_{p-1} w_p \xRightarrow{*}_G w_0 z_0 w_1 z_1 \cdots w_{p-1} z_{p-1} w_p \in A_T^*,$$

which gives the desired conclusion. In particular,

$$U_i \subseteq \{X \in A_N \mid X \xRightarrow{*}_G u \text{ for some } u \in A_T^*\}.$$

(Proof cont'd)

To prove the converse inclusion we prove by strong induction on $m \geq 1$ that $X \xRightarrow[G]{m} u$ for $u \in A_T^*$ implies $X \in U_{m-1}$. The basis case, $m = 1$, is immediate.

Suppose that the statement holds for derivations of length less than or equal to m and consider a derivation $X \xRightarrow[G]{m+1} u$ for $u \in A_T^*$. If we write the first step of this derivation, we obtain

$$X \xRightarrow[G]{} w_0 Z_0 w_1 Z_1 \cdots w_{p-1} Z_{p-1} w_p \xRightarrow[G]{m} u,$$

where $w_0, \dots, w_{p-1}, u \in A_T^*$, and $Z_0, \dots, Z_{p-1} \in A_N$.

(Proof cont'd)

The word u can be written as $u = w_0 z_0 w_1 z_1 \cdots w_{p-1} z_{p-1} w_p$, where $Z_j \xrightarrow[G]{\ell_j} z_j$, $\ell_j \leq m$ for $0 \leq j \leq p-1$. By the inductive hypothesis, we have $Z_j \in U_{\ell_j-1} \subseteq U_{m-1}$, so $w_0 Z_0 w_1 Z_1 \cdots w_{p-1} Z_{p-1} w_p \in (U_{m-1} \cup A_T)^*$. Thus, $X \in U_m$.

(Proof cont'd)

Since $U_m \subseteq U_i$ for every $m \in \mathbb{N}$ and $m \geq 1$, we obtain the converse inclusion and, therefore, the desired equality.

Note that $S \in U_i$ if and only if $L(G) \neq \emptyset$. Define the set of productions P' by

$$P' = \begin{cases} \emptyset & \text{if } S \notin U_i \\ \{X \rightarrow \alpha \mid \alpha \in (U_i \cup A_T)^* \text{ and } X \rightarrow \alpha \in P\} & \text{otherwise.} \end{cases}$$

(Proof cont'd)

Since $P' \subseteq P$ it follows that $L(G') \subseteq L(G)$. Conversely, if $u \in L(G)$, then $S \xRightarrow{*}_G u$. Let $X \rightarrow \alpha$ be a production that occurs in this derivation. We have

$$S \xRightarrow{*}_G \beta X \gamma \xRightarrow{*}_G \beta \alpha \gamma \xRightarrow{*}_G u.$$

Therefore, every nonterminal symbol that occurs in α must be productive. This allows us to conclude that $\alpha \in (U_i \cup A_T)^*$, hence $X \rightarrow \alpha \in P'$. Since every production used in the derivation $S \xRightarrow{*}_G u$ belongs to P' , it follows that $u \in L(G')$, so $L(G) \subseteq L(G')$.

Corollary

The emptiness of the language $L(G)$ generated by a context-free grammar $G = (A_N, A_T, S, P)$ is decidable.

Proof.

Note that the start symbol S of a context-free grammar G is productive if and only if $L(G) \neq \emptyset$. Therefore, in order to decide if $L(G) = \emptyset$, it suffices to compute the set U_i . Then, $L(G) = \emptyset$ if and only if $S \notin U_i$. □

Example

Let $G = (\{S, X, Y, Z\}, \{a, b\}, S, \{S \rightarrow YZ, S \rightarrow XY, S \rightarrow XZ, Z \rightarrow ab, Y \rightarrow bc\})$ be a context-free grammar. The sequence $U_0, U_1, \dots$ is given by $U_0 = \{Y, Z\}$, $U_1 = \{S, Y, Z\}$, $U_1 = U_2 = \dots$. Therefore, the grammar $G' = (\{S, Y, Z\}, \{a, b\}, S, \{S \rightarrow YZ, Z \rightarrow ab, Y \rightarrow bc\})$ has only productive symbols and is equivalent to G .