Theory of Formal Languages and Automata
Lecture 10

Mahdi Dolati

Sharif University of Technology

Fall 2023

November 4, 2023

Simplification

It is possible to remove certain types of rules:

- Useless,
- ε-rule,
- Unit.

It does not mean a reduction of the total number of rules.

Theorem

Suppose CFG $G = (V, T, S, P)$ *contains:*

$$(*) A \to x_1 B x_2, \tag{1}$$
$$(**) B \to y_1 \mid y_2 \mid \ldots \mid y_n. \tag{2}$$

where $A \neq B$ *and (**) is the only rule that has B as the left side. If we build* $\widehat{G} = (V, T, S, \widehat{P})$ *by deleting rule (*) and adding*

$$A \to x_1 y_1 x_2 \mid x_1 y_2 x_2 \mid \ldots \mid x_1 y_n x_2, \tag{3}$$

then, L(G)=L($\widehat{G}$*).*

Definition

Let G=(V,T,S,P) be a CFG. A variable $A \in V$ is said to be useful iff there is at least one $w \in L(G)$ such that

$$S \overset{*}{\Rightarrow} xAy \overset{*}{\Rightarrow} w, \tag{4}$$

where $x, y \in (V \cup T)^*$.

- A variable is useless if it is not useful.
- A production rule is useless if it contains a useless variable.

Two reasons for being useless:

1. Is not reachable from the start variable,
2. Can not derive any string.

Theorem

For any CFG $G = (V, T, S, P)$ *there is an equivalent grammar* $\widehat{G} = (\widehat{V}, \widehat{T}, S, \widehat{P})$ *that does not contain any useless variables or productions.*

Example (Useless)

Variable C can not derive any string:

$$S \to aS \mid A \mid C,$$
$$A \to a,$$
$$B \to aa,$$
$$C \to aCb,$$

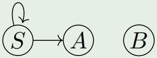
$$S \to aS \mid A,$$
$$A \to a,$$
$$B \to aa.$$

Simplification

Useless

Example (Useless Cont.)

Draw the dependency graph of the grammar where nodes are variables and (A, B) edge exists if there is a production rule $A \to xBy$.



$$S \to aS \mid A,$$
$$A \to a,$$
$$B \to aa.$$

$$S \to aS \mid A,$$
$$A \to a.$$

Definition

- A rule of the form $A \to \varepsilon$ is a ε-rule.
- A variable for which the derivation $A \overset{*}{\Rightarrow} \varepsilon$ is possible is called nullable.

Theorem

For any CFG G where $\varepsilon \notin L(G)$ there is an equivalent grammar $\widehat{G}$ having no ε-rules.

Proof.

First find the set V_N of all nullable variables:

1. For all $A \to \varepsilon$, put A into V_n,
2. For all $B \to A_a \dots A_n$ where A_1, ..., and A_n are in V_n, put B into V_n.

Second consider all rules of the form

$$A \to x_1 \dots x_m, \qquad m \geq 1, \tag{5}$$

where $x_i \in V \cup T$. Put such rules and all versions obtained from removing all combinations of nullable variables into $\widehat{P}$. If all x_i are nullable do not put $A \to \varepsilon$ into $\widehat{P}$. □

Example

$$S \to ABaC,\quad A \to BC,\quad B \to b \mid \varepsilon,\quad C \to D \mid \varepsilon,\quad D \to d.$$

$$S \to ABaC \mid BaC \mid AaC \mid ABa \mid aC \mid Aa \mid Ba \mid a,\quad A \to B \mid C \mid BC,\quad B \to b,\quad C \to D,\quad D \to d.$$

$V_n = \{B, C, A\}$

Definition

Any rule of the form $A \to B$ where $A, B \in V$ is called a unit rule or unit-production.

Theorem

For any CFG $G = (V, T, S, P)$ without ε-rules there is an equivalent grammar $\widehat{G} = (\widehat{V}, \widehat{T}, S, \widehat{P})$ that does not have any unit rules.

Example (Unit Rule)

$$
\begin{aligned}
S_0 &\rightarrow aA \\
A &\rightarrow a \\
\boldsymbol{A} &\boldsymbol{\rightarrow} \boldsymbol{B} \\
B &\rightarrow A \\
B &\rightarrow bb
\end{aligned}
$$

$$
\begin{aligned}
S_0 &\rightarrow aA \mid \boldsymbol{aB} \\
A &\rightarrow a \\
B &\rightarrow A \mid \boldsymbol{B} \\
B &\rightarrow bb
\end{aligned}
$$

Example (Unit Rule Cont.)

$$S_0 \to aA \mid aB$$
$$A \to a$$
$$B \to A \mid \boldsymbol{B}$$
$$B \to bb$$

$$S_0 \to aA \mid aB$$
$$A \to a$$
$$B \to A$$
$$B \to bb$$

Example (Unit Rule Cont.)

$$S_0 \to aA \mid aB$$
$$A \to a$$
$$\boldsymbol{B \to A}$$
$$B \to bb$$

$$S_0 \to aA \mid aB \mid \boldsymbol{aA}$$
$$A \to a$$
$$B \to bb$$

Example (Unit Rule Cont.)

$$S_0 \to aA \mid aB \mid \boldsymbol{aA}$$
$$A \to a$$
$$B \to bb$$

$$S_0 \to aA \mid aB$$
$$A \to a$$
$$B \to bb$$

Proof.

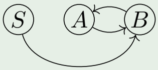
- Remove all rules of the form $A \to A$.
- For each A obtain all variables B such that $A \overset{*}{\Rightarrow} B$.
  - Use a dependency graph,
  - Nodes are variables,
  - Add edge (C, D) if rule $C \to D$ exists,
  - A derives B if there is a path from A to B.
- Put all non-unit rules of P into $\widehat{P}$.
- For all $A \overset{*}{\Rightarrow} B$ add the following rule to $\widehat{P}$:

$$A \to y_1 \mid \ldots \mid y_n,$$

where $B \to y_1 \mid \ldots \mid y_n$ is the set of all rules in $\widehat{P}$ with B on the left. □

Example

$$S \to Aa \mid B,$$
$$B \to A \mid bb,$$
$$A \to a \mid bc \mid B.$$



- $S \overset{*}{\Rightarrow} A$,
- $S \overset{*}{\Rightarrow} B$,
- $A \overset{*}{\Rightarrow} B$,
- $B \overset{*}{\Rightarrow} A$.

Add non-unit rules:

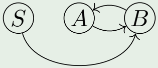
$$(\widehat{G}): S \to Aa,$$
$$B \to bb,$$
$$A \to a \mid bc.$$

Simplification

Unit

Example

$$S \to Aa \mid B,$$
$$B \to A \mid bb,$$
$$A \to a \mid bc \mid B.$$



- $S \overset{*}{\Rightarrow} A$,
- $S \overset{*}{\Rightarrow} B$,
- $A \overset{*}{\Rightarrow} B$,
- $B \overset{*}{\Rightarrow} A$.

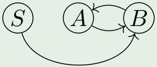
Consider $S \overset{*}{\Rightarrow} A$. Should add all $S \to y_i$ for all $A \to y_i$ from $\widehat{P}$.

$$(\widehat{G}) : S \to Aa, \quad B \to bb, \quad A \to a \mid bc.$$

$$(\widehat{G}) : S \to Aa \mid a \mid bc, \quad B \to bb, \quad A \to a \mid bc.$$

Example

$$S \to Aa \mid B,$$
$$B \to A \mid bb,$$
$$A \to a \mid bc \mid B.$$



- $S \overset{*}{\Rightarrow} A$,
- $S \overset{*}{\Rightarrow} B$,
- $A \overset{*}{\Rightarrow} B$,
- $B \overset{*}{\Rightarrow} A$.

Consider $S \overset{*}{\Rightarrow} B$. Should add all $S \to y_i$ for all $B \to y_i$ from $\widehat{P}$.

$$(\widehat{G}) : S \to Aa \mid a \mid bc, \quad B \to bb, \quad A \to a \mid bc.$$

$$(\widehat{G}) : S \to Aa \mid a \mid bc \mid bb, \quad B \to bb, \quad A \to a \mid bc.$$

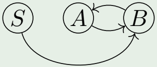
Simplification
Unit

Example

$S \to Aa \mid B,$

$B \to A \mid bb,$

$A \to a \mid bc \mid B.$



- $S \overset{*}{\Rightarrow} A,$
- $S \overset{*}{\Rightarrow} B,$
- $A \overset{*}{\Rightarrow} B,$
- $B \overset{*}{\Rightarrow} A.$

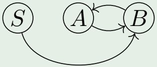
Consider $A \overset{*}{\Rightarrow} B$. Should add all $A \to y_i$ for all $B \to y_i$ from $\widehat{P}$.

$$(\widehat{G}) : S \to Aa \mid a \mid bc \mid bb, \quad B \to bb, \quad A \to a \mid bc.$$

$$(\widehat{G}) : S \to Aa \mid a \mid bc \mid bb, \quad B \to bb, \quad A \to a \mid bc \mid bb.$$

Example

$$S \to Aa \mid B,$$
$$B \to A \mid bb,$$
$$A \to a \mid bc \mid B.$$



- $S \overset{*}{\Rightarrow} A$,
- $S \overset{*}{\Rightarrow} B$,
- $A \overset{*}{\Rightarrow} B$,
- $B \overset{*}{\Rightarrow} A$.

Consider $B \overset{*}{\Rightarrow} A$. Should add all $B \to y_i$ for all $A \to y_i$ from $\widehat{P}$.

$$(\widehat{G}) : S \to Aa \mid a \mid bc \mid bb,$$
$$B \to bb,$$
$$A \to a \mid bc \mid bb.$$

$$(\widehat{G}) : S \to Aa \mid a \mid bc \mid bb,$$
$$B \to bb \mid a \mid bc,$$
$$A \to a \mid bc \mid bb.$$

Simplification

Theorem

If CF language L does not contain ε, there is a CFG that $LG = L$ and does not have any useless rules, ε-rules, or unit rules.

Proof.

Note that

- removal of unit rules does not create ε-rules,
- removal of useless rules does not create ε-rules or unit rules.

Thus, use previous theorems and obtain the grammar by using the following sequence of steps:

1. Remove ε-rules,
2. Remove unit rules.
3. Remove useless rules.

□

Definition (Chomsky normal form (CNF))

A CFG is in CNF if every rule is of the form:

$$A \to BC$$
$$A \to a$$

- a is any terminal,
- A, B and C are variables,
- B and C may not be the start variable.

Following rule is also valid:

$$S \to \varepsilon$$

Theorem

Grammars in CNF generate all context-free languages.

Proof idea:

- Convert any CF grammar into CNF:
    - Add a new start,
    - Eliminate all ε-rules,
    - Eliminate all unit rules.

Proof.

1. Add new start variable that only appears at the left-hand side of rule,
   - $S_0 \to S$.
2. For all $A \neq S_0$ that appear at the left-hand side of a ε-rule:
   - Find all rules that A appears in their right-hand side,
   - Add a new version of such rules where A is removed (consider all combinations),
   - Remove the ε-rule.
3. Fix number of symbols at the right-hand side of rules to two:

$$A \to u_1 u_2 \dots u_k \equiv \begin{array}{rl} A & \to u_1 A_1 \\ A_1 & \to u_2 A_2 \\ \dots & \\ A_{k-2} & \to u_{k-1} A_k \end{array}$$

4. Ensure two symbols are variables:

$$A_{i-1} \to u_i A_i \equiv \begin{array}{rl} A_{i-1} & \to U_i A_i \\ U_i & \to u_i \end{array}$$

□

Example (CNF Procedure)

$$S \to ASA \mid aB$$
$$A \to B \mid S$$
$$B \to b \mid \varepsilon$$

$$S_0 \to S$$
$$S \to ASA \mid aB$$
$$A \to B \mid S$$
$$B \to b \mid \varepsilon$$

Example (CNF Procedure Cont.)

$$S_0 \to S$$
$$S \to ASA \mid aB$$
$$A \to B \mid S$$
$$B \to b \mid \varepsilon$$

$$S_0 \to S$$
$$S \to ASA \mid aB \mid a$$
$$A \to B \mid S \mid \varepsilon$$
$$B \to b$$

Example (CNF Procedure Cont.)

$$
\begin{aligned}
S_0 &\to S \\
S &\to ASA \mid aB \mid a \\
A &\to B \mid S \mid \varepsilon \\
B &\to b
\end{aligned}
$$

$$
\begin{aligned}
S_0 &\to S \\
S &\to ASA \mid aB \mid a \mid \boldsymbol{SA} \mid \boldsymbol{AS} \mid \boldsymbol{S} \\
A &\to B \mid S \\
B &\to b
\end{aligned}
$$

Example (CNF Procedure Cont.)

$$S_0 \to S$$
$$S \to ASA \mid aB \mid a \mid SA \mid AS \mid \boldsymbol{S}$$
$$A \to B \mid S$$
$$B \to b$$

$$S_0 \to S$$
$$S \to ASA \mid aB \mid a \mid SA \mid AS$$
$$A \to B \mid S$$
$$B \to b$$

Example (CNF Procedure Cont.)

$$S_0 \to \boldsymbol{S}$$
$$S \to ASA \mid aB \mid a \mid SA \mid AS$$
$$A \to B \mid S$$
$$B \to b$$

$$S_0 \to \boldsymbol{ASA \mid aB \mid a \mid SA \mid AS}$$
$$S \to ASA \mid aB \mid a \mid SA \mid AS$$
$$A \to B \mid S$$
$$B \to b$$

Example (CNF Procedure Cont.)

$$
\begin{aligned}
S_0 &\to ASA \mid aB \mid a \mid SA \mid AS \\
S &\to ASA \mid aB \mid a \mid SA \mid AS \\
A &\to \boldsymbol{B} \mid S \\
B &\to b
\end{aligned}
\qquad
\begin{aligned}
S_0 &\to ASA \mid aB \mid a \mid SA \mid AS \\
S &\to ASA \mid aB \mid a \mid SA \mid AS \\
A &\to \boldsymbol{b} \mid S \\
B &\to b
\end{aligned}
$$

Example (CNF Procedure Cont.)

$$\begin{aligned} S_0 &\to ASA \mid aB \mid a \mid SA \mid AS \\ S &\to ASA \mid aB \mid a \mid SA \mid AS \\ A &\to b \mid \boldsymbol{S} \\ B &\to b \end{aligned}$$

$$\begin{aligned} S_0 &\to ASA \mid aB \mid a \mid SA \mid AS \\ S &\to ASA \mid aB \mid a \mid SA \mid AS \\ A &\to b \mid \boldsymbol{ASA \mid aB \mid a \mid SA \mid AS} \\ B &\to b \end{aligned}$$

Example (CNF Procedure Cont.)

$$
\begin{aligned}
S_0 &\to ASA \mid \boldsymbol{a}B \mid a \mid SA \mid AS \\
S &\to ASA \mid \boldsymbol{a}B \mid a \mid SA \mid AS \\
A &\to b \mid ASA \mid \boldsymbol{a}B \mid a \mid SA \mid AS \\
B &\to b
\end{aligned}
$$

$$
\begin{aligned}
S_0 &\to ASA \mid \boldsymbol{U}B \mid a \mid SA \mid AS \\
S &\to ASA \mid \boldsymbol{U}B \mid a \mid SA \mid AS \\
A &\to b \mid ASA \mid \boldsymbol{U}B \mid a \mid SA \mid AS \\
B &\to b \\
\boldsymbol{U} &\boldsymbol{\to} \boldsymbol{a}
\end{aligned}
$$

Example (CNF Procedure Cont.)

$$S_0 \to A\boldsymbol{SA} \mid UB \mid a \mid SA \mid AS$$
$$S \to A\boldsymbol{SA} \mid UB \mid a \mid SA \mid AS$$
$$A \to b \mid A\boldsymbol{SA} \mid UB \mid a \mid SA \mid AS$$
$$B \to b$$
$$U \to a$$

$$S_0 \to A\boldsymbol{A_1} \mid UB \mid a \mid SA \mid AS$$
$$S \to A\boldsymbol{A_1} \mid UB \mid a \mid SA \mid AS$$
$$A \to b \mid A\boldsymbol{A_1} \mid UB \mid a \mid SA \mid AS$$
$$B \to b$$
$$U \to a$$
$$\boldsymbol{A_1 \to SA}$$

- Put restrictions on the positions in which terminals and variables can appear.

Definition

A context-free grammar is said to be in Greibach normal form if all productions have the form

$$A \to ax, \tag{6}$$

where $a \in T$ and $x \in V^*$.

Theorem

For every context-free grammar G with $\varepsilon \notin L(G)$, there exists an equivalent grammar $\widehat{G}$ in Greibach normal form.

- In general it is not simple to:
  - convert a given grammar to Greibach normal form,
  - proof that this conversion can always be done.

Example

Following grammar is not in GNF.

$$S \to AB, \tag{7}$$
$$A \to aA \mid bB \mid b, \tag{8}$$
$$B \to b. \tag{9}$$

However, we can obtain the following grammar that is in GNF:

$$S \to aAB \mid bBB \mid bB, \tag{10}$$
$$A \to aA \mid bB \mid b, \tag{11}$$
$$B \to b. \tag{12}$$