

Lecture Notes 9: Chomsky Normal Form*Raghunath Tewari*

IIT Kanpur

Normal forms are CFGs whose substitution rules have a special form. Usually normal forms are general enough in the sense that any CFL will have a CFG in that normal form. Normal forms have a nice combinatorial structure which are useful in proving properties about CFLs.

1 Chomsky Normal Form

A CFG is said to be in *Chomsky Normal Form* (in short, CNF) if the following are true.

1. Every rule is of the form

- $A \rightarrow BC$, or
- $A \rightarrow a$,

where A, B, C are variables and a is a terminal.

2. The start variable is not present in the right hand side of any rule.
3. The rule $S \rightarrow \epsilon$ may be present (depending on whether the language has ϵ or not).

1.1 Converting a CFG to a grammar in Chomsky Normal Form

Let $G = (V, \Sigma, P, S)$ be a CFG. Below we give an algorithm to convert G into a CFG in Chomsky Normal Form.

1. Removing S from RHS of rules

If S appears on the RHS of some rule, add a new start variable S_0 and the rule $S_0 \rightarrow S$.

2. Removing ϵ -rules

Pick an ϵ -rule $A \rightarrow \epsilon$ and remove it (where A is not the start variable).

- Now for every occurrence of A on the RHS of some of all rules, add a rule deleting that occurrence of A . If A occurs multiple times on the RHS of a rule, then multiple rules might be added.
- If we have the rule $R \rightarrow A$, then add $R \rightarrow \epsilon$, unless the rule $R \rightarrow \epsilon$ has been removed previously.
- Repeat until no more ϵ -rules remain, except possibly involving the start variable.

Example: Suppose a grammar had the following rules:

$$\begin{aligned} A &\rightarrow \epsilon \\ B &\rightarrow uAv \\ C &\rightarrow u_1Av_1Aw_1 \end{aligned}$$

Then the grammar formed by removing the rule $A \rightarrow \epsilon$ will have the corresponding set of rules

$$\begin{array}{ll}
B & \rightarrow uAv \\
C & \rightarrow u_1Av_1Aw_1 \\
B & \rightarrow uv \quad (\text{rule added}) \\
C & \rightarrow u_1v_1Aw_1 \quad (\text{rule added with first occurrence of } A \text{ removed}) \\
C & \rightarrow u_1Av_1w_1 \quad (\text{rule added with second occurrence of } A \text{ removed}) \\
C & \rightarrow u_1v_1w_1 \quad (\text{rule added with both occurrence of } A \text{ removed})
\end{array}$$

3. Removing unit rules

Remove a rule $A \rightarrow B$ and for all rules of the form $B \rightarrow u$ add the rule $A \rightarrow u$, unless $A \rightarrow u$ is a unit rule that has already been removed. Repeat until no more unit rules remain.

Example: Suppose a grammar had the following rules:

$$\begin{array}{ll}
A & \rightarrow B \\
B & \rightarrow u
\end{array}$$

Then the grammar formed by removing the rule $A \rightarrow B$ will have the corresponding set of rules

$$\begin{array}{ll}
B & \rightarrow u \\
A & \rightarrow u \quad (\text{rule added})
\end{array}$$

4. Shortening the RHS

For every rule of the form $A \rightarrow u_1u_2 \dots u_k$ for $k \geq 3$, where $u_i \in V \cup T$, replace the rule with

$$\begin{array}{ll}
A & \rightarrow u_1A_1 \\
A_1 & \rightarrow u_2A_2 \\
A_2 & \rightarrow u_3A_3 \\
& \vdots \\
& \vdots \\
& \vdots \\
A_{k-2} & \rightarrow u_{k-1}u_k
\end{array}$$

Here A_i 's are the new variables added to the grammar.

5. Replacing certain terminals

If there is a rule of the form $A \rightarrow uv$ where at least one of either u or v is a terminal symbol (say u), then replace the rule $A \rightarrow uv$ with

$$\begin{array}{ll}
A & \rightarrow Uv \\
U & \rightarrow u
\end{array}$$

where U is a new variable added to the grammar. Repeat until no such rules remain.

1.2 An Example – CFG to CNF

Consider the following CFG where S is the start variable:

$$\begin{aligned} S &\longrightarrow ASB \\ A &\longrightarrow aASA \mid a \mid \epsilon \\ B &\longrightarrow SbS \mid A \mid bb \end{aligned}$$

We will convert the above grammar into a grammar in CNF. The rules/variables that get added at each step are shown in **bold** font.

1. Adding a new start variable S_0 , since S appears on RHS of some rules.

$$\begin{aligned} \mathbf{S_0} &\longrightarrow S \\ S &\longrightarrow ASB \\ A &\longrightarrow aASA \mid a \mid \epsilon \\ B &\longrightarrow SbS \mid A \mid bb \end{aligned}$$

2. Eliminating $A \longrightarrow \epsilon$.

$$\begin{aligned} S_0 &\longrightarrow S \\ S &\longrightarrow ASB \mid \mathbf{SB} \\ A &\longrightarrow aASA \mid a \mid \mathbf{aSA} \mid \mathbf{aAS} \mid \mathbf{aS} \\ B &\longrightarrow SbS \mid A \mid bb \mid \epsilon \end{aligned}$$

3. Eliminating $B \longrightarrow \epsilon$.

$$\begin{aligned} S_0 &\longrightarrow S \\ S &\longrightarrow ASB \mid SB \mid \mathbf{AS} \mid \mathbf{S} \\ A &\longrightarrow aASA \mid a \mid aSA \mid aAS \mid aS \\ B &\longrightarrow SbS \mid A \mid bb \end{aligned}$$

4. Eliminating the unit rule $S \longrightarrow S$.

$$\begin{aligned} S_0 &\longrightarrow S \\ S &\longrightarrow ASB \mid SB \mid AS \\ A &\longrightarrow aASA \mid a \mid aSA \mid aAS \mid aS \\ B &\longrightarrow SbS \mid A \mid bb \end{aligned}$$

5. Eliminating the unit rule $B \longrightarrow A$.

$$\begin{aligned} S_0 &\longrightarrow S \\ S &\longrightarrow ASB \mid SB \mid AS \\ A &\longrightarrow aASA \mid a \mid aSA \mid aAS \mid aS \\ B &\longrightarrow SbS \mid bb \mid \mathbf{aASA} \mid \mathbf{a} \mid \mathbf{aSA} \mid \mathbf{aAS} \mid \mathbf{aS} \end{aligned}$$

6. Eliminating the unit rule $S_0 \rightarrow S$.

$$\begin{aligned}
S_0 &\rightarrow \mathbf{ASB} \mid \mathbf{SB} \mid \mathbf{AS} \\
S &\rightarrow ASB \mid SB \mid AS \\
A &\rightarrow aASA \mid a \mid aSA \mid aAS \mid aS \\
B &\rightarrow SbS \mid bb \mid aASA \mid a \mid aSA \mid aAS \mid aS
\end{aligned}$$

7. Adding variable U_1 and rule $U_1 \rightarrow AS$.

$$\begin{aligned}
S_0 &\rightarrow \mathbf{U_1B} \mid SB \mid AS \\
S &\rightarrow \mathbf{U_1B} \mid SB \mid AS \\
A &\rightarrow a\mathbf{U_1A} \mid a \mid aSA \mid \mathbf{aU_1} \mid aS \\
B &\rightarrow SbS \mid bb \mid \mathbf{aU_1A} \mid a \mid aSA \mid \mathbf{aU_1} \mid aS \\
\mathbf{U_1} &\rightarrow \mathbf{AS}
\end{aligned}$$

8. Adding variable U_2 and rule $U_2 \rightarrow aU_1$.

$$\begin{aligned}
S_0 &\rightarrow U_1B \mid SB \mid AS \\
S &\rightarrow U_1B \mid SB \mid AS \\
A &\rightarrow \mathbf{U_2A} \mid a \mid aSA \mid aU_1 \mid aS \\
B &\rightarrow SbS \mid bb \mid \mathbf{U_2A} \mid a \mid aSA \mid aU_1 \mid aS \\
U_1 &\rightarrow AS \\
\mathbf{U_2} &\rightarrow \mathbf{aU_1}
\end{aligned}$$

9. Adding variable U_3 and rule $U_3 \rightarrow aS$.

$$\begin{aligned}
S_0 &\rightarrow U_1B \mid SB \mid AS \\
S &\rightarrow U_1B \mid SB \mid AS \\
A &\rightarrow U_2A \mid a \mid \mathbf{U_3A} \mid aU_1 \mid aS \\
B &\rightarrow SbS \mid bb \mid U_2A \mid a \mid \mathbf{U_3A} \mid aU_1 \mid aS \\
U_1 &\rightarrow AS \\
U_2 &\rightarrow aU_1 \\
\mathbf{U_3} &\rightarrow \mathbf{aS}
\end{aligned}$$

10. Adding variable U_4 and rule $U_4 \rightarrow Sb$.

$$\begin{aligned}
S_0 &\rightarrow U_1B \mid SB \mid AS \\
S &\rightarrow U_1B \mid SB \mid AS \\
A &\rightarrow U_2A \mid a \mid U_3A \mid aU_1 \mid aS \\
B &\rightarrow \mathbf{U_4S} \mid bb \mid U_2A \mid a \mid U_3A \mid aU_1 \mid aS \\
U_1 &\rightarrow AS \\
U_2 &\rightarrow aU_1 \\
U_3 &\rightarrow aS \\
\mathbf{U_4} &\rightarrow \mathbf{Sb}
\end{aligned}$$

11. Adding variables V_1, V_2 and rules $V_1 \rightarrow a, V_2 \rightarrow b$.

$$\begin{aligned}
S_0 &\rightarrow U_1B \mid SB \mid AS \\
S &\rightarrow U_1B \mid SB \mid AS \\
A &\rightarrow U_2A \mid a \mid U_3A \mid \mathbf{V_1U_1} \mid \mathbf{V_1S} \\
B &\rightarrow U_4S \mid \mathbf{V_2V_2} \mid U_2A \mid a \mid U_3A \mid \mathbf{V_1U_1} \mid \mathbf{V_1S} \\
U_1 &\rightarrow AS \\
U_2 &\rightarrow \mathbf{V_1U_1} \\
U_3 &\rightarrow \mathbf{V_1S} \\
U_4 &\rightarrow \mathbf{SV_2} \\
\mathbf{V_1} &\rightarrow \mathbf{a} \\
\mathbf{V_2} &\rightarrow \mathbf{b}
\end{aligned}$$

Chomsky Normal Form of the given grammar is

$$\begin{aligned}
S_0 &\rightarrow U_1B \mid SB \mid AS \\
S &\rightarrow U_1B \mid SB \mid AS \\
A &\rightarrow U_2A \mid U_3A \mid V_1U_1 \mid V_1S \mid a \\
B &\rightarrow U_4S \mid V_2V_2 \mid U_2A \mid U_3A \mid V_1U_1 \mid V_1S \mid a \\
U_1 &\rightarrow AS \\
U_2 &\rightarrow V_1U_1 \\
U_3 &\rightarrow V_1S \\
U_4 &\rightarrow SV_2 \\
V_1 &\rightarrow a \\
V_2 &\rightarrow b
\end{aligned}$$

Exercise 1. Problem 2.14 from textbook.