Chomsky Normal Form (CNF)

Tanjila Alam Sathi

Lecturer, CSE Department

Simplification of CFG

In CFG, sometimes all the production rules and symbols are not needed for the derivation of strings. Besides, there may also be some NULL productions or UNIT productions.

Elimination of these UNIT or NULL productions and symbols is called -> Simplification of CFG.

CNF

- **$A \rightarrow BC$ or $A \rightarrow a$**
- Where **A, B**, & C are variables & **a** is a terminal

❑ CFG to CNF: Need Simplification

1. Eliminate *useless symbols*, those variables/terminals that do not appear in any derivation of a terminal string from the start symbol
2. Eliminate *ε-productions*, those of the form $A \rightarrow \varepsilon$ for some variable A
3. Eliminate *unit productions*, those of the form $A \rightarrow$ **B for variables A & B**

Steps to convert a CFG to CNF:

1. If the start symbol S occurs on some right hand side, create a new start symbol S' and a new production S'->S

2. Remove Null Productions of the form $A \rightarrow \varepsilon$

3. Remove Unit productions of the form $A \rightarrow \mathbf{B}$

4. Replace each production A->B1…………Bn where n >2,with A->B1C where C->B2,……Bn. Repeat this step for all productions having two or more symbols on the right side.

5. If the right side of any production is in the form A->aB where 'a' is a terminal and A and B non-terminals,then the production is replaced by A->XB and X->a. Repeat this step for every production which is of the form A->aB.

Eliminating ε-productions

- **Basis:** If $A \rightarrow \varepsilon$ is a production of G, then A is nullable

- **Induction**: If there is a production

  $B \rightarrow C_1 C_2 \ldots C_k$ such that each C is a variable and each C is nullable, then B is nullable

Procedures

- CFG $G = (V, T, P, S)$.
- Determine all the nullable symbols of G
- Construct a new grammar $G_1 = (V, T, P_1, S)$
- For each production $A \rightarrow X_1X_2 \ldots X_k$ of P, suppose that m of the k X_i's are nullable symbols.
- The new grammar will have 2^m versions of this productions, where nullable X_i's in all possible combinations are present/absent
- Remove $A \rightarrow \varepsilon$

Example

- $S \rightarrow AB$
- $A \rightarrow aAA \mid \varepsilon$
- $B \rightarrow bBB \mid \varepsilon$

- A & B nullable
- $S \rightarrow AB$ nullable
- 04 possible combinations (present/absent of A & B)
- Ignore absent
- 03 combinations of S
- **$S \rightarrow AB \mid A \mid B$**

- $A \rightarrow aAA$
- 04 possible combinations (AA)
- $A \rightarrow aAA \mid aA \mid aA \mid a$
- Ignore one aA
- $B \rightarrow bBB \mid bB \mid b$

G_1

- $S \rightarrow AB \mid A \mid B$
- $A \rightarrow aAA \mid aA \mid a$
- $B \rightarrow bBB \mid bB \mid b$

Eliminating Unit Productions

- Any production of the form **A → B** , where A and B are variables, is called a unit production.

- **Basis**: (A, A) is a unit pair of any variable A, if A $\Rightarrow^*$ A by 0 steps.

- **Induction**: Let's (A, B) be a unit pair, and let B → C is a production, where A, B, and C are variables, then we can conclude that (A, C) is also a unit pair.

Example

$$
\begin{array}{lcl}
I & \rightarrow & a \mid b \mid Ia \mid Ib \mid I0 \mid I1 \\
F & \rightarrow & I \mid (E) \\
T & \rightarrow & F \mid T * F \\
E & \rightarrow & T \mid E + T
\end{array}
$$

From basis: (E, E), (T, T), (F, F) & (I, I) – unit pairs

From induction:

1. (E, E) & production E $\rightarrow$ T $\Rightarrow$ unit pair (E, T)

- From induction:

1. (E, E) & production $E \rightarrow T \Rightarrow$ unit pair (E, T)
2. (E, T) & production $T \rightarrow F \Rightarrow$ unit pair (E, F)
3. (E, F) & production $F \rightarrow I \Rightarrow$ unit pair (E, I)
4. (T, T) & production $T \rightarrow F \Rightarrow$ unit pair (T, F)
5. (T, F) & production $F \rightarrow I \Rightarrow$ unit pair (T, I)
6. (F, F) & production $F \rightarrow I \Rightarrow$ unit pair (F, I)

No more pairs that can be inferred

Procedures

- Given a CFG $G = (V, T, P, S)$, construct CFG $G_1 = (V, T, P_1, S)$:

1. Find all the unit pairs of G
2. For each unit pair (A, B) add to P_1 all the productions $\mathbf{A \rightarrow \alpha}$ where $\mathbf{B \rightarrow \alpha}$ is a nonunit production in P

-A=B is possible; in that way, P_1 contains all the nonunit production in P

Example

	Productions
(E, E)	$E \rightarrow E + T$
(E, T)	$E \rightarrow T * F$
(E, F)	$E \rightarrow (E)$
(E, I)	$E \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(T, T)	$T \rightarrow T * F$
(T, F)	$T \rightarrow (E)$
(T, I)	$T \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(F, F)	$F \rightarrow (E)$
(F, I)	$F \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
(I, I)	$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

Resulting grammars with no unit productions

Productions
$E \rightarrow E + T \mid T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$T \rightarrow T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$F \rightarrow (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

CFG to CNF

1. Arrangement of all bodies of length 2 or more to contain only variables.

   -for every terminal **a** create a new variable **A**. this variable has only 1 production $A \rightarrow a$

2. Breaking bodies of length 3 or more into a cascade productions, where each one has a body consisting of 2 variables.

-break productions $\mathbf{A \rightarrow B_1B_2\text{-----}B_k}$ for $k \geq 3$

$A \rightarrow B_1C_1, \quad C_1 \rightarrow B_2C_2, \ldots\ldots, C_{k-3} \rightarrow B_{k-2}C_{k-2}, \; C_{k-2} \rightarrow B_{k-1}B_k$

Example

$$\begin{aligned} I &\rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1 \\ F &\rightarrow I \mid (E) \\ T &\rightarrow F \mid T * F \\ E &\rightarrow T \mid E + T \end{aligned}$$

- 08 terminals: a, b, 0, 1, +, *, (, and)

No unit Productions
$E \rightarrow E + T \mid T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$T \rightarrow T * F \mid (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$F \rightarrow (E) \mid a \mid b \mid Ia \mid Ib \mid I0 \mid I1$
$I \rightarrow a \mid b \mid Ia \mid Ib \mid I0 \mid I1$

- Introduce new variables to represent terminals:

$$A \rightarrow a \quad B \rightarrow b \quad Z \rightarrow 0 \quad O \rightarrow 1$$

$$P \rightarrow + \quad M \rightarrow * \quad L \rightarrow (\quad R \rightarrow)$$

Example

- **Step1:** Make all bodies either a single terminal or multiple variables:

$$
\begin{array}{lll}
E & \rightarrow & EPT \mid TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO \\
T & \rightarrow & TMF \mid LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO \\
F & \rightarrow & LER \mid a \mid b \mid IA \mid IB \mid IZ \mid IO \\
I & \rightarrow & a \mid b \mid IA \mid IB \mid IZ \mid IO \\
A & \rightarrow & a \\
B & \rightarrow & b \\
Z & \rightarrow & 0 \\
O & \rightarrow & 1 \\
P & \rightarrow & + \\
M & \rightarrow & * \\
L & \rightarrow & (\\
R & \rightarrow &)
\end{array}
$$

- **Step 2**: Make all bodies either a single terminal or two variables:

$$
\begin{array}{lcl}
E & \rightarrow & EC_1 \mid TC_2 \mid LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO \\
T & \rightarrow & TC_2 \mid LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO \\
F & \rightarrow & LC_3 \mid a \mid b \mid IA \mid IB \mid IZ \mid IO \\
I & \rightarrow & a \mid b \mid IA \mid IB \mid IZ \mid IO \\
A & \rightarrow & a \\
B & \rightarrow & b \\
Z & \rightarrow & 0 \\
O & \rightarrow & 1 \\
P & \rightarrow & + \\
M & \rightarrow & * \\
L & \rightarrow & (\\
R & \rightarrow &) \\
C_1 & \rightarrow & PT \\
C_2 & \rightarrow & MF \\
C_3 & \rightarrow & ER
\end{array}
$$

Try Yourself

Problem from book:

7.1.2, 7. 1. 3, 7. 1. 4 & 7. 1. 5