CSC 339 – Theory of Computation
Fall 2023

8. Normal Forms of Context-Free Grammars

Outline

- Simplification of CFG
- Chomsky normal form
- Greinbach normal form

Substitution Rule

$S \to aB$

$A \to aaA$

$A \to abBc$

$B \to aA$

$B \to b$

Substitute

$B \to b$

Equivalent grammar

$S \to aB \mid ab$

$A \to aaA$

$A \to abBc \mid abbc$

$B \to aA$

3

$$S \to aB \mid ab$$

$$A \to aaA$$

$$A \to abBc \mid abbc$$

$$B \to aA$$

Substitute

$$B \to aA$$

$$S \to \cancel{aB} \mid ab \mid aaA$$

$$A \to aaA$$

$$A \to \cancel{abB}c \mid abbc \mid abaAc$$

Equivalent grammar

4

In general: $A \to xBz$

$B \to y_1$

Substitute

$B \to y_1$

$A \to xBz \mid xy_1z$

equivalent grammar

5

Nullable Variables

$\varepsilon-\text{production}:\quad X \to \varepsilon$

Nullable Variable: $Y \Rightarrow \ldots \Rightarrow \varepsilon$

Example:

$$S \to aMb$$

$$M \to aMb$$

$$M \to \varepsilon$$

Nullable variable

$\varepsilon-\text{production}:$

6

Removing $\varepsilon -$productions

$$S \to aMb$$

$$M \to aMb$$

$$\cancel{M \to \varepsilon}$$

Substitute

$$M \to \varepsilon$$

$$S \to aMb \mid ab$$

$$M \to aMb \mid ab$$

After we remove all the $\varepsilon -$productions
all the nullable variables disappear
(except for the start variable)

Unit-Productions

Unit Production: $X \rightarrow Y$

(a single variable in both sides)

Example:

$$S \rightarrow aA$$

$$A \rightarrow a$$

$$\boxed{A \rightarrow B}$$

$$\boxed{B \rightarrow A}$$

Unit Productions

$$B \rightarrow bb$$

Removal of unit productions:

$S \to aA$

$A \to a$

$\cancel{A \to B}$

$B \to A$

$B \to bb$

Substitute

$A \to B$

$S \to aA \mid aB$

$A \to a$

$B \to A \mid B$

$B \to bb$

9

Unit productions of form $X \to X$
can be removed immediately

$$S \to aA \mid aB$$
$$A \to a$$
$$B \to A \mid \cancel{B}$$
$$B \to bb$$

Remove
$B \to B$

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

10

$$S \to aA \mid aB$$

$$A \to a$$

$$\cancel{B \to A}$$

$$B \to bb$$

Substitute $B \to A$

$$S \to aA \mid aB \mid aA$$

$$A \to a$$

$$B \to bb$$

Remove repeated productions

$$S \rightarrow aA \mid aB \mid \cancel{aA}$$
$$A \rightarrow a$$
$$B \rightarrow bb$$

Final grammar

$$S \rightarrow aA \mid aB$$
$$A \rightarrow a$$
$$B \rightarrow bb$$

Useless Productions

$$S \to aSb$$
$$S \to \varepsilon$$
$$S \to A$$
$$A \to aA$$

Useless Production

Some derivations never terminate...

$$S \Rightarrow A \Rightarrow aA \Rightarrow aaA \Rightarrow \ldots \Rightarrow aa\ldots aA \Rightarrow \ldots$$

Another example:

$$S \rightarrow A$$

$$A \rightarrow aA$$

$$A \rightarrow \varepsilon$$

$$B \rightarrow bA$$ Useless Production

Not reachable from S

In general:

If there is a derivation

$$S \Rightarrow \ldots \Rightarrow xAy \Rightarrow \ldots \Rightarrow w \quad \in L(G)$$

consists of terminals

Then variable A is useful.

Otherwise, variable A is useless.

15

A production $A \rightarrow x$ is useless
if any of its variables is useless.

	$S \rightarrow aSb$	
	$S \rightarrow \varepsilon$	Productions
Variables	$S \rightarrow A$	useless
useless	$A \rightarrow aA$	useless
useless	$B \rightarrow C$	useless
useless	$C \rightarrow D$	useless

Removing Useless Variables and Productions

Example of a grammar:

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

First: find all variables that can produce strings with only terminals or ε (possible useful variables)

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

Round 1: $\{A, B\}$

(the right hand side of productions that have only terminals)

Round 2: $\{A, B, S\}$

(the right hand side of a production has terminals and variables of previous round)

This process can be generalized

18

Remove productions that use variables other than $\{A, B, S\}$

$$S \to aS \mid A \mid \cancel{C}$$
$$A \to a$$
$$B \to aa$$
$$\cancel{C \to aCb}$$

$$\Rightarrow$$

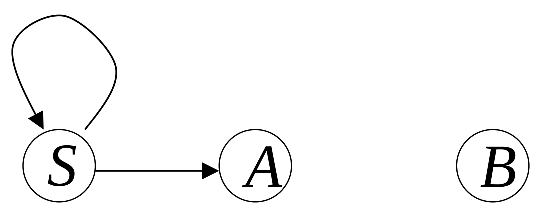
$$S \to aS \mid A$$
$$A \to a$$
$$B \to aa$$

19

Second: Find all variables reachable from S

Use a Dependency Graph where nodes are variables

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



unreachable

20

Keep only the variables
reachable from S

$$S \to aS \mid A$$
$$A \to a$$
$$\cancel{B \to aa}$$

Final Grammar

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

Contains only
useful variables

Summary

This sequence guarantees that unwanted variables and productions are removed.

- **Step 1:** Remove Nullable Variables
- **Step 2:** Remove Unit-Productions
- **Step 3:** Remove Useless Variables

Normal Forms for Context-free Grammars

Chomsky Normal Form

Each production has one of the following forms:

$$A \rightarrow BC \qquad \text{or} \qquad A \rightarrow a$$

variable variable terminal

24

Examples:

$$S \to AS$$
$$S \to a$$
$$A \to SA$$
$$A \to b$$

Chomsky
Normal Form

$$S \to AS$$
$$S \to AAS$$
$$A \to SA$$
$$A \to aa$$

Not Chomsky
Normal Form

25

Conversion to Chomsky Normal Form

Example: $S \to ABa$ Not Chomsky Normal Form

$A \to aab$

$B \to Ac$

Convert it to Chomsky Normal Form

1) Introduce new variables for the terminals:

$$T_a, T_b, T_c$$

$$S \to ABa$$
$$A \to aab$$
$$B \to Ac$$

$$\Rightarrow$$

$$S \to ABT_a$$
$$A \to T_a T_a T_b$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

2) Introduce new intermediate variable V_1
to break first production:

$$\boxed{S \to ABT_a}$$
$$A \to T_a T_a T_b$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

$$\Rightarrow$$

$$\boxed{\begin{array}{l} S \to AV_1 \\ V_1 \to BT_a \end{array}}$$
$$A \to T_a T_a T_b$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

28

3) Introduce intermediate variable: V_2

$$S \to AV_1$$
$$V_1 \to BT_a$$
$$\boxed{A \to T_aT_aT_b}$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

$$\Rightarrow$$

$$S \to AV_1$$
$$V_1 \to BT_a$$
$$\boxed{\begin{array}{l} A \to T_aV_2 \\ V_2 \to T_aT_b \end{array}}$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

29

Initial grammar

$$S \to ABa$$

$$A \to aab$$

$$B \to Ac$$

Final grammar in Chomsky Normal Form:

$$S \to AV_1$$

$$V_1 \to BT_a$$

$$A \to T_a V_2$$

$$V_2 \to T_a T_b$$

$$B \to AT_c$$

$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$

In general:

From any context-free grammar (which doesn't produce ε) not in Chomsky Normal Form,

we can obtain an equivalent grammar in Chomsky Normal Form.

The Procedure

First remove:

Nullable variables

Unit productions

Optionally, useless variables

Then, for every symbol: a

New variable: T_a

Add production: $T_a \rightarrow a$

In productions with length at least 2
replace a with T_a

Productions of form $A \rightarrow a$ do not need to change.

33

Replace any production $A \to C_1 C_2 \cdots C_n$

with
$$A \to C_1 V_1$$
$$V_1 \to C_2 V_2$$
$$\cdots$$
$$V_{n-2} \to C_{n-1} C_n$$

New intermediate variables: $V_1, V_2, \ldots, V_{n-2}$

34

Observations

- Chomsky normal forms are good for parsing and proving theorems
- It is easy to find the Chomsky normal form for any context-free grammar

Greinbach Normal Form

All productions have the form:

$$A \to a\, V_1 V_2 \cdots V_k \qquad k \geq 0$$

symbol

variables

36

Examples:

$$S \to cAB$$
$$A \to aA \mid bB \mid b$$
$$B \to b$$

Greinbach
Normal Form

$$S \to abSb$$
$$S \to aa$$

Not Greinbach
Normal Form

Conversion to Greinbach Normal Form:

$$S \to abSb$$
$$S \to aa$$

$$S \to aT_b S T_b$$
$$S \to aT_a$$
$$T_a \to a$$
$$T_b \to b$$

Greinbach
Normal Form

Observations

- Greinbach normal forms are very good for parsing strings (better than Chomsky normal forms).
- However, it is not always easy to find the Greinbach normal form of a grammar.