CS116-Automata Theory and Formal Languages

Lecture 8

Simplifications of Context-Free Grammars

Computer Science Department

1st Semester 2025-2026

A Substitution Rule

Equivalent grammar

$S \rightarrow aB$

$A \rightarrow aaA$

$A \rightarrow abBc$

$B \rightarrow aA$

$B \rightarrow b$

Substitute

$B \rightarrow b$

$S \rightarrow aB \mid ab$

$A \rightarrow aaA$

$A \rightarrow abBc \mid abbc$

$B \rightarrow aA$

$$S \rightarrow aB \mid ab$$
$$A \rightarrow aaA$$
$$A \rightarrow abBc \mid abbc$$
$$B \rightarrow aA$$

Substitute

$$B \rightarrow aA$$

$$S \rightarrow \cancel{aB} \mid ab \mid aaA$$
$$A \rightarrow aaA$$
$$A \rightarrow \cancel{abBc} \mid abbc \mid abaAc$$

Equivalent grammar

In general:

$$A \rightarrow xBz$$

$$B \rightarrow y_1$$

Substitute

$B \rightarrow y_1$

$$A \rightarrow xBz \mid xy_1 z$$

equivalent grammar

Nullable Variables

$\lambda -$ production: $\qquad X \rightarrow \lambda$

Nullable Variable: $\qquad Y \Rightarrow \ldots \Rightarrow \lambda$

Example:

$$S \rightarrow aMb$$
$$M \rightarrow aMb$$
$$M \rightarrow \lambda$$

Nullable variable

$\lambda -$ production

Removing $\lambda-$productions

$$S \rightarrow aMb$$
$$M \rightarrow aMb$$
$$\cancel{M \rightarrow \lambda}$$

Substitute $M \rightarrow \lambda$

$$S \rightarrow aMb \mid ab$$
$$M \rightarrow aMb \mid ab$$

After we remove all the $\lambda-$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:

$$S \rightarrow aA$$

$$A \rightarrow a$$

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

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

$$B \rightarrow bb$$

Unit Productions

Removal of unit productions:

$S \to aA$

$A \to a$

$\cancel{A \to B}$

$B \to A$

$B \to bb$

Substitute

$A \to B$

$S \to aA \mid aB$

$A \to a$

$B \to A \mid B$

$B \to bb$

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

$$B \to B$$

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

$$S \to aA \mid aB$$

$$A \to a$$

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

$$B \to bb$$

Substitute

$B \to A$

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

$$A \to a$$

$$B \to bb$$

Remove repeated productions

Final grammar

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

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

Useless Productions

$$S \rightarrow aSb$$
$$S \rightarrow \lambda$$
$$S \rightarrow A$$
$$A \rightarrow aA$$

Useless Production

Some derivations never terminate...

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

Another grammar:

$$S \rightarrow A$$

$$A \rightarrow aA$$

$$A \rightarrow \lambda$$

$$B \rightarrow bA$$

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

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

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

Removing Useless Variables and Productions

Example Grammar:

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

First: find all variables that can produce strings with only terminals or λ (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 production that has 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

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

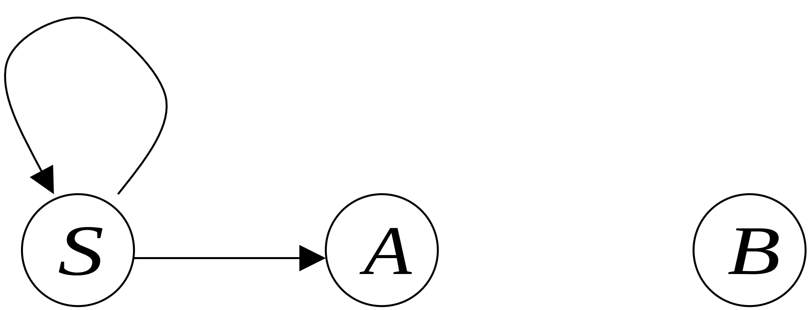
$$S \to aS \mid A \mid \cancel{C}$$
$$A \to a$$
$$B \to aa$$
$$\cancel{C \to aCb}$$

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

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

Keep only the variables reachable from S

Final Grammar

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

$$S \rightarrow aS \mid A$$
$$A \rightarrow a$$

Contains only useful variables

Removing All

Step 1: Remove Nullable Variables

Step 2: Remove Unit-Productions

Step 3: Remove Useless Variables

This sequence guarantees that unwanted variables and productions are removed

Normal Forms
for
Context-free Grammars

Chomsky Normal Form

Each productions has form:

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

variable variable terminal

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

Conversion to Chomsky Normal Form

Example:

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

Not Chomsky Normal Form

We will convert it to Chomsky Normal Form

Introduce new variables for the terminals:

$$T_a, T_b, T_c$$

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

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

Introduce new intermediate variable V_1 to break first production:

$$\boxed{S \to ABT_a}$$

$$A \to T_a T_a T_b$$

$$B \to AT_c$$

$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$

$$\boxed{\begin{array}{l} S \to AV_1 \\ V_1 \to BT_a \end{array}}$$

$$A \to T_a T_a T_b$$

$$B \to AT_c$$

$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$

Introduce intermediate variable: V_2

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

$$\Longrightarrow$$

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

Final grammar in Chomsky Normal Form:

Initial grammar

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

$$S \to AV_1$$
$$V_1 \to BT_a$$
$$A \to T_a V_2$$
$$V_2 \to T_a T_b$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

In general:

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

we can obtain:
an equivalent grammar
in Chomsky Normal Form

The Procedure

First remove:

Nullable variables

Unit productions

(Useless variables optional)

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!

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

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 form:

$$A \rightarrow a\; V_1 V_2 \cdots V_k \qquad\qquad k \geq 0$$

symbol

variables

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:

$$S \rightarrow abSb$$
$$S \rightarrow aa$$

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

Greinbach
Normal Form

Observations

- Greinbach normal forms are very good for parsing strings (better than Chomsky Normal Forms)

- However, it is difficult to find the Greinbach normal of a grammar