

Lecture 18 – Normal Forms of Context-Free Grammars

COSE215: Theory of Computation

Jihyeok Park



2023 Spring

- A **context-free grammar (CFG)** is a 4-tuple:

$$G = (V, \Sigma, S, R)$$

where

- V : a finite set of **variables** (nonterminals)
 - Σ : a finite set of **symbols** (terminals)
 - $S \in V$: the **start variable**
 - $R \subseteq V \times (V \cup \Sigma)^*$: a set of **production rules**.
-
- How to **simplify** a CFG?

Let's put it in **Chomsky normal form (CNF)**!

Definition (Chomsky Normal Form)

A CFG is in **Chomsky normal form (CNF)** if all productions are of the form for some $A, B, C \in V$ and $a \in \Sigma$:

$$A \rightarrow BC \quad \text{OR} \quad A \rightarrow a$$

(If $\epsilon \in L(G)$, then $S \rightarrow \epsilon$ is allowed with forbidden S on RHSs.)

$$S \rightarrow 0ABC \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid C$$

$$B \rightarrow 0B \mid 1$$

$$C \rightarrow CC \mid \epsilon$$

$$D \rightarrow 1D \mid AA$$

Is it possible to put this CFG in CNF? **Yes!**

$$S \rightarrow XS_1 \mid XB \mid YB \mid BB \quad A \rightarrow AA_1 \mid BA_2 \quad B \rightarrow XB \mid 1$$

$$S_1 \rightarrow AB$$

$$A_1 \rightarrow BA_2$$

$$X \rightarrow 0$$

$$A_2 \rightarrow BX$$

$$Y \rightarrow 1$$

1. Chomsky Normal Form (CNF)

- Eliminating ϵ -Productions

 - Nullable Variables

- Eliminating Unit Productions

 - Unit Pairs

- Eliminating Useless Variables

 - Generating Variables

 - Reachable Variables

- Putting CFG in CNF

Is it possible to eliminate ϵ -**productions**?

$$A \rightarrow \epsilon$$

However, it is impossible to eliminate when the language of the CFG contains the empty word (i.e., $\epsilon \in L(G)$).

Let's construct a new CFG G' from G such that

$$L(G') = L(G) \setminus \{\epsilon\}$$

by eliminating ϵ -productions:

- 1 Find all **nullable variables**.
- 2 Construct a new CFG with productions produced by replacing nullable variables with ϵ in all combinations, except for the ϵ -production.

Definition (Nullable Variables)

For a given CFG $G = (V, \Sigma, S, R)$, a variable $A \in V$ is **nullable** if

$$A \Rightarrow^* \epsilon$$

We can inductively define the set of **nullable variables**:

- **(Basis Case)** If $A \rightarrow \epsilon \in R$, then A is nullable.
- **(Induction Case)** If $A \rightarrow X_1 X_2 \cdots X_n \in R$ and X_1, X_2, \dots, X_n are all nullable, then A is nullable.

Consider the following CFG:

$$S \rightarrow 0ABC \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid C$$

$$B \rightarrow 0B \mid 1$$

$$C \rightarrow CC \mid \epsilon$$

$$D \rightarrow 1D \mid AA$$

- 1 Find all **nullable variables**: $\{A, C, D\}$
- 2 Construct a new CFG with productions produced by replacing nullable variables with ϵ in all combinations, except for the ϵ -production:

$$S \rightarrow 0ABC \mid 0BC \mid 0AB \mid 0B \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid BB0 \mid C$$

$$B \rightarrow 0B \mid 1$$

$$C \rightarrow CC \mid C$$

$$D \rightarrow 1D \mid 1 \mid AA \mid A$$

Is it possible to eliminate **unit productions**?

$$A \rightarrow B$$

Yes, we can do it by following the steps below:

- ① Find all **unit pairs**.
- ② Construct a new CFG by adding all possible non-unit productions of B to A for each unit pair (A, B) .

Definition (Unit Pairs)

For a given CFG $G = (V, \Sigma, S, R)$, a pair of variables $(A, B) \in V \times V$ is a **unit pair** if

$$A \Rightarrow^* B$$

We can inductively define the set of **unit pairs**:

- **(Basis Case)** (A, A) is a unit pair for all $A \in V$.
- **(Induction Case)** If (A, B) is a unit pair and $B \rightarrow C \in R$, then (A, C) is a unit pair.

After eliminating ϵ -productions:

$$S \rightarrow 0ABC \mid 0BC \mid 0AB \mid 0B \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid BB0 \mid C$$

$$B \rightarrow 0B \mid 1$$

$$C \rightarrow CC \mid C$$

$$D \rightarrow 1D \mid 1 \mid AA \mid A$$

- 1 Find all **unit pairs**:

$$\{(S, S), (A, A), (A, C), (B, B), (C, C), (D, D), (D, A), (D, C)\}$$

- 2 Construct a new CFG by adding all possible non-unit productions of B to A for each unit pair (A, B) :

$$S \rightarrow 0ABC \mid 0BC \mid 0AB \mid 0B \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid BB0 \mid CC$$

$$B \rightarrow 0B \mid 1$$

$$C \rightarrow CC$$

$$D \rightarrow 1D \mid 1 \mid AA \mid ABB0 \mid BB0$$

What are useless variables?

- **Non-generating variables:** Variables that cannot derive any word.
- **Unreachable variables:** Variables unreachable from the start variable.

Is it possible to eliminate **useless variables**?

Yes, we can do it by following the steps below:

- 1 Find all **generating variables**.
- 2 Find all **reachable variables**.
- 3 Construct a new CFG by removing all productions that contain non-generating variables or come from unreachable variables.

Definition (Generating Variables)

For a given CFG $G = (V, \Sigma, S, R)$, a variable $A \in V$ is a **generating variable** if for some $w \in \Sigma^*$,

$$A \Rightarrow^* w$$

We can inductively define the set of **generating variables**:

- **(Basis Case)** There is no basis case.
- **(Induction Case)** If $A \rightarrow \alpha \in R$ and α contains only symbols or generating variables, then A is a generating variable.

Definition (Reachable Variables)

For a given CFG $G = (V, \Sigma, S, R)$, a variable $A \in V$ is a **reachable variable** if there exists a derivation:

$$S \Rightarrow^* \alpha A \beta$$

We can inductively define the set of **reachable variables**:

- **(Basis Case)** The start variable S is reachable variable.
- **(Induction Case)** If $A \in V$ is a reachable variable and $A \rightarrow \alpha \in R$, then all variables in α are reachable variables.

Eliminating Useless Variables – Example

After eliminating ϵ -productions and unit productions:

$$S \rightarrow 0ABC \mid 0BC \mid 0AB \mid 0B \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid BB0 \mid CC$$

$$B \rightarrow 0B \mid 1$$

$$C \rightarrow CC$$

$$D \rightarrow 1D \mid 1 \mid AA \mid ABB0 \mid BB0$$

- 1 Find all **generating variables**: $\{S, A, B, D\}$ – C is non-generating.
- 2 Find all **reachable variables**: $\{S, A, B, C\}$ – D is unreachable.
- 3 Construct a new CFG by removing all productions that contain non-generating variables or come from unreachable variables:

$$S \rightarrow 0AB \mid 0B \mid 1B \mid BB$$

$$A \rightarrow ABB0 \mid BB0$$

$$B \rightarrow 0B \mid 1$$

Our goal is to put a CFG in **Chomsky normal form (CNF)** consisting of:

$$A \rightarrow BC \quad \text{OR} \quad A \rightarrow a$$

(If $\epsilon \in L(G)$, then $S \rightarrow \epsilon$ is allowed with forbidden S on RHSs.)

We can put a CFG in CNF by following the steps below:

- ① If S on RHSs, add a new start variable S' and a production $S' \rightarrow S$.
- ② Eliminate ϵ -productions, unit productions, and useless variables.
- ③ Arrange so that all RHSs whose length is greater than 1 consist only of variables. To do so, if terminal a appears in a RHS, then replace it with a new variable A and add a production $A \rightarrow a$.
- ④ Replace all RHSs whose length is greater than 2 with a chain of variables. To do so, if $A \rightarrow X_1X_2 \cdots X_n$ is a production with $n > 2$, then replace it with a sequence of productions:

$$A \rightarrow X_1A_1 \quad A_1 \rightarrow X_2A_2 \quad \cdots \quad A_{n-2} \rightarrow X_{n-1}X_n$$

- ⑤ If ϵ is in the original CFG, add a production $S \rightarrow \epsilon$ (or $S' \rightarrow \epsilon$).

- ① If S on RHSs, add a new start variable S' and a production $S' \rightarrow S$.
- ② Eliminate ϵ -productions, unit productions, and useless variables:

$$\begin{aligned} S &\rightarrow 0AB \mid 0B \mid 1B \mid BB \\ A &\rightarrow ABB0 \mid BB0 \\ B &\rightarrow 0B \mid 1 \end{aligned}$$

- ③ Arrange so that all RHSs whose length > 1 consist only of variables:

$$\begin{aligned} S &\rightarrow XAB \mid XB \mid YB \mid BB & X &\rightarrow 0 \\ A &\rightarrow ABBX \mid BBX & Y &\rightarrow 1 \\ B &\rightarrow XB \mid 1 \end{aligned}$$

- ④ Replace all RHSs whose length > 2 with a chain of variables:

$$\begin{aligned} S &\rightarrow XS_1 \mid XB \mid YB \mid BB & A &\rightarrow AA_1 \mid BA_2 & B &\rightarrow XB \mid 1 \\ S_1 &\rightarrow AB & A_1 &\rightarrow BA_2 & X &\rightarrow 0 \\ & & A_2 &\rightarrow BX & Y &\rightarrow 1 \end{aligned}$$

- ⑤ If ϵ is in the original CFG, add a production $S \rightarrow \epsilon$ (or $S' \rightarrow \epsilon$): **No.**

Putting CFG in CNF – Example 2

Let's put the following CFG in CNF:

$$S \rightarrow aSb \mid \epsilon$$

- ① If S on RHSs, add a new start variable S' and a production $S' \rightarrow S$.

$$S' \rightarrow S \quad S \rightarrow aSb \mid ab$$

- ② Eliminate ϵ -productions, unit productions, and useless variables:

$$S' \rightarrow aSb \mid ab \quad S \rightarrow aSb \mid ab$$

- ③ Arrange so that all RHSs whose length > 1 consist only of variables:

$$S' \rightarrow ASB \mid AB \quad S \rightarrow ASB \mid AB \quad A \rightarrow a \quad B \rightarrow b$$

- ④ Replace all RHSs whose length > 2 with a chain of variables:

$$S' \rightarrow AS_1 \mid AB \quad S \rightarrow AS_1 \mid AB \quad S_1 \rightarrow SB \quad A \rightarrow a \quad B \rightarrow b$$

- ⑤ If ϵ is in the original CFG, add a production $S \rightarrow \epsilon$ (or $S' \rightarrow \epsilon$): **Yes.**

$$S' \rightarrow \epsilon \mid AS_1 \mid AB \quad S \rightarrow AS_1 \mid AB \quad S_1 \rightarrow SB \quad A \rightarrow a \quad B \rightarrow b$$

1. Chomsky Normal Form (CNF)

- Eliminating ϵ -Productions

 - Nullable Variables

- Eliminating Unit Productions

 - Unit Pairs

- Eliminating Useless Variables

 - Generating Variables

 - Reachable Variables

- Putting CFG in CNF

- Properties of Context-Free Languages

Jihyeok Park

jihyeok_park@korea.ac.kr

<https://plrg.korea.ac.kr>