Chomsky Normal Form

→ A standardized form for a CFG.

→ One of the main goals is to be able to check if string w is generated by grammar G.

→ What we want : Simple rules that are easy to check.

→ What we don't want : loops, Empty derivations, Useless rules.

Consider $S \to 0S1 \mid 1S0S1 \mid T$

$T \to S \mid \varepsilon$

1010000

10S10T00

10S1000

1O1000

Def 2.8 : A context-free grammar is in Chomsky Normal Form if every rule is of the form

$A \to BC$

$A \to a$

$A \to BC \to CDC$

$\to cDC$

$\to caC$

$\to cac$

where a is a terminal, A, B, C are variables and B and C are not the start variable. Also, we allow $S \to \varepsilon$.

One of the advantages of the Chomsky Normal Form is the "predictability" in the derivation of a string. Any string of length n requires exactly $2n-1$ steps in its derivation.

↓ EXERCISE!

Also Chomsky NF results in an efficient algorithm for checking if w is generated by G.

Natural Question at this stage. What if there is no equivalent CFG that is in the Chomsky Normal Form?

Theorem 2.9: Any CFL is generated by a CFG in Chomsky Normal Form.

Proof: We provide a process that demonstrates how to convert any CFG into Chomsky NF.

Optional Step: Remove useless symbols and productions

$S \to AB \mid a$ → B is a useless var.

$A \to a$

1. Add new start variable S_0.

$$S_0 \to S$$

This guarantees that S_0 does not appear in the RHS of any rule.

2. Remove ε rules.

Say there is the rule $A \to \varepsilon$. Then modify rules with A in RHS.

If $R \to uAv$ was a rule, then replace it with

$$R \to uAv \mid uv$$

(here u, v are string of terminals and variables)

If $R \to A$ is a rule, then replace it with

$$R \to A \mid \varepsilon$$

unless $R \to \varepsilon$ was already removed.

3. Remove unit rules like $A \to B$.

If there is a rule $B \to u$, add $A \to u$ unless we removed that rule.

4. Restructure rules with long RHS.

Example:

$$S \to TST \mid aB$$
$$T \to B \mid S$$
$$B \to b \mid \varepsilon$$

1. Add S_0 as new start variable.

$$S_0 \to S$$

$$S \rightarrow TST \mid aB$$
$$T \rightarrow B \mid S$$
$$B \rightarrow b \mid \varepsilon$$

2. Remove $B \rightarrow \varepsilon$

$$S_0 \rightarrow S$$
$$S \rightarrow TST \mid aB \mid a$$
$$T \rightarrow B \mid \varepsilon \mid S$$
$$B \rightarrow b$$

$T \rightarrow \varepsilon$

$$S_0 \rightarrow S$$
$$S \rightarrow TST \mid ST \mid TS \mid \cancel{S} \mid aB \mid a$$
$$T \rightarrow B \mid S$$
$$B \rightarrow b$$

redundant rule

Can be removed.

3. Unit Rules

$S_0 \to S$

$S_0 \to TST \mid TS \mid ST \mid aB \mid a$

$S \to TST \mid TS \mid ST \mid aB \mid a$

$T \to B \mid S$

$B \to b$

$T \to B$

$T \to b \mid S$

$T \to S$

$S_0 \to TST \mid TS \mid ST \mid aB \mid a$

$S \to TST \mid TS \mid ST \mid aB \mid a$

$T \to TST \mid TS \mid ST \mid aB \mid a \mid b$

$B \to b$

4. Restructure rules with more than one symbol in the RHS.

let $TS = Z$, $a = A$.

$$S_0 \rightarrow ZT \mid TS \mid ST \mid AB \mid a$$

$$S \rightarrow ZT \mid TS \mid ST \mid AB \mid a$$

$$T \rightarrow ZT \mid TS \mid ST \mid AB \mid a \mid b$$

$$Z \rightarrow TS$$

$$A \rightarrow a$$

$$B \rightarrow b$$

Now the above CFG is in Chomsky Normal Form and by construction, we have ensured that it is equivalent to the original grammar.