Normal Forms

Normal Forms

- **Chomsky Normal Form (CNF)**
- Def.: A context-free grammar G = (V, Σ, R, S) is said to be in Chomsky Normal Form (CNF), iff every rule in R is of one of the following forms:
- $X \rightarrow a$ where $a \in \Sigma$, or
- $X \rightarrow BC$ where B and $C \in V$

Example1

» Which one is Chomsky Normal Form ? Why?

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

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

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

Chomsky
Normal Form

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

Not Chomsky
Normal Form

Converting to Chomsky Normal Form

- There exists 5-steps algorithm to convert a CFG G into a new grammar Gc such that: $L(G) = L(Gc) - \{\varepsilon\}$
- *convertChomsky*(G:CFG) =
- 1- G′ = *removeEps*(G:CFG) $S \rightarrow \varepsilon$
- 2- G″ = *removeUnits*(G′:*CFG*) $A \rightarrow B$
- 3- G‴ = *removeMixed*(G″:*CFG*) $A \rightarrow aB$
- 4- G′^{v} = *removeLong*(G‴ :*CFG*) $S \rightarrow ABCD$
- 5- G′^{v} : $L(G) = L(G'^{v}) - \{\varepsilon\}$

  G^{v} = atmostoneEps(G′^{v}:CFG) $S^* \rightarrow S$, $S \rightarrow \varepsilon$

Remove Epsilon

- Find the set N of nullable variables in G.
- X is nullable iff either $X \rightarrow \varepsilon$ or
- $(X \rightarrow A\ ,\ A \rightarrow \varepsilon)$: $X \rightarrow \varepsilon$
- Ex1: G: $S \rightarrow aACa$
- $A \rightarrow B \mid a$
- $B \rightarrow C \mid c$
- $C \rightarrow cC \mid \varepsilon$
- Now, since $C \rightarrow \varepsilon$, C is nullable
- since $B \rightarrow C$, B is nullable
- since $A \rightarrow B$, A is nullable
- Therefore N = { A,B,C}

Remove Epsilon

- According to N, the rules will be:
- $S \rightarrow aACa$
- $A \rightarrow B \mid a \mid \varepsilon$
- $B \rightarrow C \mid c \mid \varepsilon$
- $C \rightarrow cC \mid \varepsilon$
- Now, removeEps returns G′:
- $S \rightarrow aACa \mid aAa \mid aCa \mid aa$
- $A \rightarrow B \mid a$
- $B \rightarrow C \mid c$
- $C \rightarrow cC \mid c$

Remove Unit Productions

- Def.: unit production is a rule whose right hand side consists of a single nonterminal symbol. Ex: $A \rightarrow B$
- Remove any unit production from G′.
- Ex1(continue), G‘:
- $S \rightarrow aACa \mid aAa \mid aCa \mid aa$
- $A \rightarrow B \mid a$
- $B \rightarrow C \mid c$
- $C \rightarrow cC \mid c$

Remove Unit Productions

- Now by Appling *removeUnits*(G':*CFG*) :
- Remove A → B But B → C | c, so Add A → C | c
- Remove B → C Add B → cC (B → c, already there)
- Remove A → C Add *A* → c*C* (*A* → c, already there)
- So *remove*Units returns G'' :
- *S* → a*AC*a | a*A*a | a*C*a | aa
- *A* → a | c | c*C*
- *B* → c | c*C*
- *C* → c*C* | c

Remove Mixed

- Def.: mixed is a rule whose right hand side consists of combination of terminals or terminals with nonterminal symbol.
- Create a new nonterminal T_a for each terminal $a \in \Sigma$
- For each T_a , add the rule $T_a \rightarrow a$
- Ex1(continue), G'' :
- $S \rightarrow aACa \mid aAa \mid aCa \mid aa$
- $A \rightarrow a \mid c \mid cC$
- $B \rightarrow c \mid cC$
- $C \rightarrow cC \mid c$

Remove Mixed

- Now, by apply *removeMixed*(G'':*CFG*), G''' :
- $S \rightarrow T_a A C T_a \mid T_a A T_a \mid T_a C T_a \mid T_a T_a$
- $A \rightarrow$ a | c | $T_c C$
- $B \rightarrow$ c | $T_c C$
- $C \rightarrow T_c C$ | c
- $T_a \rightarrow$ a
- $T_c \rightarrow$ c

Remove Long

- Def.: long is a rule whose right hand side consists of more than two nonterminal symbol.
- R: $A \rightarrow BCDE$
- By remove long, it will be:
- $A \rightarrow BM_2$
- $M_2 \rightarrow CM_3$
- $M_3 \rightarrow DE$
- Ex1(continue), G'' :
- $S \rightarrow aACa \mid aAa \mid aCa \mid aa$
- $A \rightarrow a \mid c \mid cC$
- $B \rightarrow c \mid cC$
- $C \rightarrow cC \mid c$

Remove Long

- Ex1(continue), G''' :
- $S \rightarrow T_aACT_a \mid T_aAT_a \mid T_aCT_a \mid T_aT_a$
- $A \rightarrow$ a | c | T_cC
- $B \rightarrow$ c | T_cC
- $C \rightarrow T_cC$ | c
- $T_a \rightarrow$ a
- $T_c \rightarrow$ c

Remove Long

- Now, by apply *removeLong*(G''' :*CFG*), G'^{v} :
- $S \rightarrow T_aS_1 \mid T_aS_3 \mid T_aS_4 \mid T_aT_a$
- $S_1 \rightarrow AS_2 \quad S_2 \rightarrow CT_a \quad S_3 \rightarrow AT_a \quad S_4 \rightarrow CT_a$
- $A \rightarrow a \mid c \mid T_cC$
- $B \rightarrow c \mid T_cC$
- $C \rightarrow T_cC \mid c$
- $T_a \rightarrow a$
- $T_c \rightarrow c$

Add Epsilon

- Finally, *atmostoneEps*(G'^{v}:CFG) does not apply in Ex1, $G^{v} = G'^{v}$ since L does not contain ε (S is not nullable).
- $S \rightarrow T_aS_1 \mid T_aS_3 \mid T_aS_4 \mid T_aT_a$
- $S_1 \rightarrow AS_2 \quad S_2 \rightarrow CT_a \quad S_3 \rightarrow AT_a \quad S_4 \rightarrow CT_a$
- $A \rightarrow a \mid c \mid T_cC$
- $B \rightarrow c \mid T_cC$
- $C \rightarrow T_cC \mid c$
- $T_a \rightarrow a$
- $T_c \rightarrow c$

Example2

- Convert the following CFG to CNF:
- $S \rightarrow ABC$
- $A \rightarrow aC \mid D$
- $B \rightarrow bB \mid \varepsilon \mid A$
- $C \rightarrow Ac \mid \varepsilon \mid Cc$
- $D \rightarrow aa$

1-Remove epsilon

- N = {A,B,C} So
- add $S \rightarrow AB|BC|AC$, $A \rightarrow a$, $B \rightarrow b$, $C \rightarrow c$
- delete $B \rightarrow \varepsilon$, $C \rightarrow \varepsilon$
- The result is:
- $S \rightarrow ABC|\ AB|BC|AC$
- $A \rightarrow aC\ |\ D|a$
- $B \rightarrow bB\ |\ \ A|b$
- $C \rightarrow Ac\ |\ Cc|\ c$
- $D \rightarrow aa$

2- Remove units

» $S \rightarrow ABC|\ AB|BC|AC$

» $A \rightarrow aC\ |\ D\ |a$ ($A \rightarrow D$, $D \rightarrow aa$) remove $A \rightarrow D$

add $A \rightarrow aa$

» $B \rightarrow bB\ |\ \ A|b$ ($B \rightarrow A$, $A \rightarrow aC|aa|a$) remove $B \rightarrow A$

add $B \rightarrow aC|aa|a$

» $C \rightarrow Ac\ |\ Cc|c$

» $D \rightarrow aa$

2- Remove units

- » $S \rightarrow ABC|\ AB|BC|AC$
- » $A \rightarrow aC \mid aa \mid a$
- » $B \rightarrow bB \mid aC|aa|a|b$
- » $C \rightarrow Ac \mid Cc|c$
- » $D \rightarrow aa$

3- Remove Mixed

- $S \rightarrow ABC|\ AB|BC|AC$
- $A \rightarrow T_aC \mid T_aT_a \mid a$
- $B \rightarrow T_bB \mid T_aC|\ T_aT_a \mid a|b$
- $C \rightarrow AT_c \mid CT_c|c$
- $D \rightarrow T_aT_a$

4- Remove Long

- » $S \rightarrow AM_2 | M_2 | AC | AB$
- » $M_2 \rightarrow BC$
- » $A \rightarrow T_aC \mid T_aT_a \mid a$
- » $B \rightarrow T_bB \mid T_aC \mid T_aT_a \mid a \mid b$
- » $C \rightarrow AT_c \mid CT_c \mid c$
- » $D \rightarrow T_aT_a$
- » $T_a \rightarrow a$
- » $T_b \rightarrow b$
- » $T_c \rightarrow c$

5- Add Epsilon: N/A

- » $S \rightarrow AM_2 | \cancel{M_2} | BC | AC | AB$
- » $M_2 \rightarrow BC$
- » $A \rightarrow T_aC \mid T_aT_a \mid a$
- » $B \rightarrow T_bB \mid T_aC \mid T_aT_a \mid a \mid b$
- » $C \rightarrow AT_c \mid CT_c \mid c$
- » $D \rightarrow T_aT_a$
- » $T_a \rightarrow a$
- » $T_b \rightarrow b$
- » $T_c \rightarrow c$

Example3

- Convert the following CFG to CNF:
- A → BAB | B | ε
- B → 00 | ε
- 1- remove epsilon:
- A → BAB | B | BB | AB | BA
- B → 00
- 2- remove units:
- A → BAB | 00 | BB | AB | BA
- B → 00

» 3- remove mixed

» $A \rightarrow BC \mid T_0T_0 \mid BB \mid AB \mid BA$

» $C \rightarrow AB$, $B \rightarrow T_0T_0$, $T_0 \rightarrow 0$

» 4- Add epsilon (Optional, Generally this step is applicable for root variable i.e. S)

» $A \rightarrow BC \mid T_0T_0 \mid BB \mid AB \mid BA \mid \varepsilon$

» $C \rightarrow AB$

» $B \rightarrow T_0T_0$

» $T_0 \rightarrow 0$

Example4

» Convert the following CFG to CNF:

$$S \rightarrow ABa$$

$$A \rightarrow aab$$

$$B \rightarrow Ac$$

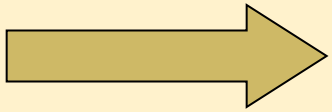
Introduce new variables for the terminals:

$$T_a, T_b, T_c$$

$$S \to ABa$$

$$A \to aab$$

$$B \to Ac$$



$$S \to ABT_a$$

$$A \to T_a T_a T_b$$

$$B \to AT_c$$

$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$

Introduce new intermediate variable V_1 to break first production:

$$S \to ABT_a$$

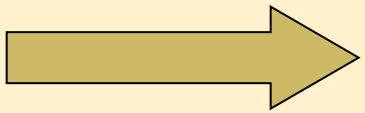
$$A \to T_a T_a T_b$$

$$B \to AT_c$$

$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$



$$S \to AV_1$$

$$V_1 \to BT_a$$

$$A \to T_a T_a T_b$$

$$B \to AT_c$$

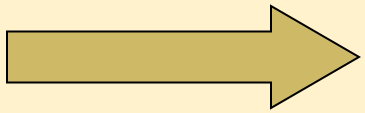
$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$

Introduce intermediate variable: V_2

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



$$S \to AV_1$$
$$V_1 \to BT_a$$
$$A \to T_aV_2$$
$$V_2 \to T_aT_b$$
$$B \to AT_c$$
$$T_a \to a$$
$$T_b \to b$$
$$T_c \to c$$

Final grammar in Chomsky Normal Form:

Initial grammar

$$S \to ABa$$

$$A \to aab$$

$$B \to Ac$$

$$S \to AV_1$$

$$V_1 \to BT_a$$

$$A \to T_aV_2$$

$$V_2 \to T_aT_b$$

$$B \to AT_c$$

$$T_a \to a$$

$$T_b \to b$$

$$T_c \to c$$

Example5

- Convert the following CFG to CNF:
- $E \rightarrow E + T$
- $E \rightarrow T$
- $T \rightarrow T * F$
- $T \rightarrow F$
- $F \rightarrow (E)$
- $F \rightarrow id$

1- Remove units

- Remove $E \rightarrow T$, add $E \rightarrow T * F \mid F$
- Remove $E \rightarrow F$, add $E \rightarrow (E) \mid id$
- Remove $T \rightarrow F$, add $T \rightarrow (E) \mid id$
- The result:
- $E \rightarrow E + T \mid T * F \mid (E) \mid id$
- $T \rightarrow T * F \mid (E) \mid id$
- $F \rightarrow (E) \mid id$

2- Remove Mixed

- $E \rightarrow E\ T_{+}\ T \mid T\ T_{*}\ F \mid T_{(}E\ T_{)} \mid id$
- $T \rightarrow T\ T_{*}\ F \mid T_{(}E\ T_{)} \mid id$
- $F \rightarrow T_{(}E\ T_{)} \mid id$
- $T_{(} \rightarrow ($
- $T_{)} \rightarrow)$
- $T_{+} \rightarrow +$
- $T_{*} \rightarrow *$

3- Remove Long

- $E \rightarrow E\ M_2 \mid T\ M_3 \mid T_(\ M_4 \mid id$
- $T \rightarrow T\ M_3 \mid T_(M_4 \mid id$
- $F \rightarrow T_(\ M_4 \mid id$
- $M_2 \rightarrow T_+\ T$
- $M_3 \rightarrow T_*\ F$
- $M_4 \rightarrow E\ T_)$
- $T_(\rightarrow ($
- $T_) \rightarrow)$
- $T_+ \rightarrow +$
- $T_* \rightarrow *$

5- Add Epsilon: N/A

- $E \rightarrow E\ M_2 \mid T\ M_3 \mid T_(\ M_4 \mid id$
- $T \rightarrow T\ M_3 \mid T_(M_4 \mid id$
- $F \rightarrow T_(\ M_4 \mid id$
- $M_2 \rightarrow T_+\ T$
- $M_3 \rightarrow T_*\ F$
- $M_4 \rightarrow E\ T_)$
- $T_(\rightarrow ($
- $T_) \rightarrow)$
- $T_+ \rightarrow +$
- $T_* \rightarrow *$

» Thank you for various WEB Resources!!