CNF AND GNF

COMP 4200 – Formal Language

Self unit productions can be ignored

2| S → aAa | bBb , A → b | a , B → b , C → b

SIMPLIFY THE GIVEN GRAMMAR

- S → 0A | 1B | C A → 0S | 00 B → 1 | A C → 01
- S → ~~AB~~/CA ~~B → BC/AB~~ A → a C → ~~a~~B/b
- S→ASB | e A→aAS | a B→SbS | A | bb
- S→aAa | bBb | e A→C | a B→C | b C→C | b ~~D→A | B | ab~~

1 S → 0A | 1B | 01 A → 0S | 00 B → 1 | 0S | 00

C → 0 + -

2 S → CA A → a C → b

3 S → ASB | AB A → aAS | a | aA

B → SbS | A | bb | bS | Sb | b

B → SbS | bb | bS | Sb | b | aAS | a | aA

Safe Sequence

1. Remove ε
2. Remove Unit Production
3. Remove Useless

CHOMSKY NORMAL FORM

CHOMSKY NORMAL FORM

Chomsky Normal Form. A grammar where every production is either of the form $A \to BC$ or $A \to c$ (where A, B, C are arbitrary variables and c an arbitrary symbol).

Example: $S \to AS \mid a$ $A \to SA \mid b$

(If language contains ε, then we allow $S \to \varepsilon$ where S is start symbol, and forbid S on RHS.)

new start variable

$A \to BC$

$A \to a$

WHY CHOMSKY NORMAL FORM?

- The key advantage is that in Chomsky Normal Form, every derivation of a string of n letters has exactly $2n - 1$ steps.
- Thus, one can determine if a string is in the language by exhaustive search of all derivations.

CHOMSKY NORMAL FORM

- Every context free grammar/ CFL without ε can be generated by a grammar for which every production in the form of $A \rightarrow AA$ and $A \rightarrow a$ where A is variable and a is a terminal.
- CNF format:
  - Non-terminal $\rightarrow$ non-terminal non-terminal
  - non-terminal $\rightarrow$ terminal
- The conversion to Chomsky Normal Form has four main steps:
  1. Get rid of all ε productions.
  2. Get rid of all unit productions (where RHS is one variable).
  3. Replace every production that is too long by shorter productions.
  4. Move all terminals to productions where RHS is one terminal.

CAN ALWAYS PUT CFG INTO CHOMSKY NORMAL FORM

- Recall: CFG in Chomsky normal form if each rule has form:
  - $A \rightarrow BC$ or $A \rightarrow x$ or $S \rightarrow \varepsilon$ where $A \in V$; $B, C \in V - \{S\}$; $x \in \Sigma$.
- Theorem: Every CFL can be described by a CFG in Chomsky normal form.
- Proof Idea:
  - Start with CFG $G = (V, \Sigma, R, S)$.
  - Replace, one-by-one, every rule that is not "Chomsky".
  - Need to take care of: Start variable (not allowed on RHS of rules) ε-rules ($A \rightarrow \varepsilon$ not allowed when A isn't start variable) all other violating rules ($A \rightarrow B$, $A \rightarrow aBc$, $A \rightarrow BCDE$)

- A context-free grammar is in Chomsky normal form if every rule is of the form

$A \rightarrow BC$

$A \rightarrow a$

where a is any terminal and A, B, and C are any variables—except that B and C may not be the start variable.

We can also have $S \rightarrow \varepsilon$, where S is the start variable.

Theorem: Any context-free language is generated by a context-free grammar in Chomsky normal form.

- Add a new start variable S0 and the rule $S0 \rightarrow S$, where S was the original start variable. This change guarantees that the start variable doesn't occur on the right-hand side of a rule.
- Remove an ε-rule $A \rightarrow \varepsilon$, where A is not the start variable. If $R \rightarrow uAv$ is a rule in which u and v are strings of variables and terminals, we add rule $R \rightarrow uv$. We do so for each occurrence of an A, so the rule $R \rightarrow uAvAw$ causes us to add $R \rightarrow uvAw$, $R \rightarrow uAvw$, and $R \rightarrow uvw$. If we have the rule $R \rightarrow A$, we add $R \rightarrow \varepsilon$ unless we had previously removed the rule $R \rightarrow \varepsilon$.

- Remove all unit rules $A \rightarrow B$.
  - Whenever a rule $B \rightarrow u$ appears, add the rule $A \rightarrow u$
  - Repeat these steps until all unit rules are eliminated.
- Convert all remaining rules rule $A \rightarrow u_1 u_2 \cdots u_k$, where $k \geq 3$ with the rules

  $A \rightarrow u_1 A_1$,

  $A_1 \rightarrow u_2 A_2$,

  .....

  $A_{k-2} \rightarrow u_{k-1} u_k$
- The A_i's are new variables. Replace any terminal u_i in the preceding rule(s) with the new variable U_i and add the rule $U_i \rightarrow u_i$.

Convert CFG to CNF: P, S→ASA/aB, A→B/S/ε, B→b/ε

$S \rightarrow ASA \mid aB \mid SA \mid AS \mid S \mid a$

ε removal

$A \rightarrow B \mid S$

$B \rightarrow b$

$S \rightarrow ASA \mid aB \mid SA \mid AS \mid a$

$A \rightarrow b \mid ASA \mid aB \mid AS \mid SA \mid a$

$B \rightarrow b$ ✓

Unit production removal

Can remove $S \rightarrow S$

XA

$S \to ASA \mid aB \mid SA \mid AS \mid a$

$A \to b \mid ASA \mid aB \mid AS \mid SA \mid a$

$B \to b$ ✓

$X \to AS$

$Y \to a$

> Also should be in CNF

reducing length of rules

$S \to XA \mid YB \mid AS \mid SA \mid a$

$A \to b \mid XA \mid YB \mid AS \mid SA \mid a$

$B \to b$

$X \to AS$

$Y \to a$

CNF Version

- Convert CFG to CNF: P, S→ASA/aB, A→B/S/ε, B→b/ε
- Step 1: Since S appears in RHS, we **add new state S0 and S0→S**
  - S0→S, S→ASA/aB, A→B/S/ε, B→b/ε
- Step 2: Remove Null productions:
  - A → ε, S→ASA/aB
  - S → SA/AS/S
  - B → ε, S→ASA/aB
  - S → a
  - P: S0→S, S→ASA/SA/AS/aB/a/S, A→B/S, B→b
- Step 3: Remove Unit productions:
  - S→S,S0→S,A→B,A→S
  - Removing S→S: S0→S,S→ASA/aB/a/AS/SA, A→B/S, B→b
  - Removing S0→S: S0→ ASA/aB/a/AS/SA,S→ASA/aB/a/AS/SA, A→B/S, B→b
  - Removing A→B: S0→ ASA/aB/a/AS/SA,S→ASA/aB/a/AS/SA, A→b/S, B→b
  - Removing A→S: S0→ ASA/aB/a/AS/SA,S→ASA/aB/a/AS/SA, A→b/ASA/aB/a/AS/SA, B→b

- Step 4: Find out the production that has more than two variables in RHS.
  - S0→ASA, S→ASA, A→ ASA, X→SA
  - After removing we get,
    - S0 → AX/aB/a/AS/SA
    - S → AX/aB/a/AS/SA
    - A → AX/aB/a/AS/SA/b
    - B → b
    - X → SA
- Step 5: Now change the production, S0 → aB, S→aB, A→aB, to remove 'a' make Y→a
  - S0→AX/YB/a/AS/SA
  - S→ AX/YB/a/AS/SA
  - A→b/ AX/YB/a/AS/SA
  - B→b
  - X→SA
  - Y→a

Convert the following CFG to CNF, S → ABA A → aA|ε , B → bB|ε

S → ABA | BA | AB | B | AA | A | ε Removal of ε

A → aA | a

B → bB | b

S_0 → S

S → ABA | AB | BA | B | AA | A

A → aA | a

B → bB | b

Unit Productions

$S_0 \to ABA \mid AB \mid BA \mid bB \mid b \mid AA \mid aA \mid a$

~~$S \to ABA \mid AB \mid BA \mid bB \mid b \mid AA \mid aA \mid a$~~ - Non reachable

$A \to aA \mid a$

$B \to bB \mid b$

$X \to AB$

$Y \to b$

$Z \to a$

$S_0 \to XA \mid AB \mid BA \mid YB \mid b \mid AA \mid ZA \mid a$

$A \to ZA \mid a$

$B \to YB \mid b$

$X \to AB$

$Y \to b$

$Z \to a$

ANOTHER EXAMPLE

- Convert the following CFG to CNF.
  S → ABA A → aA|ε , B → bB|ε
- Step 1: Since S appears in RHS, we add new state S0 and S0→S
  - S0→S, S→ABA, A → aA|ε, B → bB|ε
- Step 2: Remove Null production.
  - A → ε, S→AB/BA/B, A→a
  - B → ε, S→AA/A, B→b
  - S0→S, S→ABA/AB/BA/AA/A/B, A→aA/a, B→bB/b
- Step 3: Remove Unit productions:
  - S0→S,S→B,S→A
  - Removing S→A, S→ABA/AB/BA/AA/a/B/aA, A→aA/a, B→bB/b
  - Removing S→B, S→, A→aA/a, B→bB/b
  - Removing S0→S, S0→ ABA/AB/BA/AA/a/b/aA/bB, A→aA/a, B→bB/b

- Step 4: Find out the production that has more than two variables in RHS.
  - S0→ABA, S→ABA, X→AB
  - After removing we get,
    - S0 → XA/AB/BA/AA/a/b/aA/bB,
    - S → XA/AB/BA/AA/a/b/aA/bB,
    - A → aA/a,
    - B → bB/b
    - X → AB
- Step 5: Now change the production, S0 → aA, S→aA, A→aA,B→bB, S0 → bB, S→bB to remove 'a' make Y→a,to remove 'b' make Z→b.
  - S0 → XA/AB/BA/AA/a/b/YA/ZB,
  - S → XA/AB/BA/AA/a/b/YA/ZB,
  - A → YA/a,
  - B → ZB/b
  - X → AB
  - Y→a
  - Z→b

EXAMPLE: CONVERT CFG INTO CHOMSKY NORMAL FORM

$S \rightarrow XSX \mid aY$

$X \rightarrow Y \mid S \mid \varepsilon$

$Y \rightarrow b \mid \varepsilon$

$X \rightarrow \varepsilon$

$S \rightarrow XSX \mid aY \mid SX \mid XS$

$X \rightarrow Y \mid S$

$Y \rightarrow b \mid \varepsilon$

$Y \rightarrow \varepsilon$

$S \rightarrow XSX \mid aY \mid SX \mid XS \mid a$

$X \rightarrow S \mid Y$

$Y \rightarrow b$

Unit Production

$S \rightarrow XSX \mid aY \mid SX \mid XS \mid a$

$X \rightarrow XSX \mid aY \mid SX \mid XS \mid a \mid b$

$Y \rightarrow b$

$Z \rightarrow XS$

$A \rightarrow a$

CNF

$S \rightarrow ZX \mid AY \mid SX \mid XS \mid a$

$X \rightarrow ZX \mid AY \mid SX \mid XS \mid a \mid b$

$Y \rightarrow b$

$Z \rightarrow XS$

$A \rightarrow a$

EXAMPLE: CONVERT CFG INTO CHOMSKY NORMAL FORM

- Initial CFG G0:

$S \rightarrow XSX \mid aY$

$X \rightarrow Y \mid S \mid \varepsilon$

$Y \rightarrow b \mid \varepsilon$

Step 1. Introduce new start variable S0 and new rule $S0 \rightarrow S$:

$S0 \rightarrow S$

$S \rightarrow XSX \mid aY$

$X \rightarrow Y \mid S \mid \varepsilon$

$Y \rightarrow b \mid \varepsilon$

- Step 2: Remove ε-rules for which left side is not start variable:
  - (i) remove $Y \rightarrow \varepsilon$
  - $S0 \rightarrow S$
  - $S \rightarrow XSX \mid aY \mid a$
  - $X \rightarrow Y \mid S \mid \varepsilon$
  - $Y \rightarrow b$
  - (ii) remove $X \rightarrow \varepsilon$
  - $S0 \rightarrow S$
  - $S \rightarrow XSX \mid aY \mid a \mid SX \mid XS \mid S$
  - $X \rightarrow Y \mid S$
  - $Y \rightarrow b$
- Step 3: Remove unit rules:
  - (i) remove unit rule, $S \rightarrow S$
    - $S0 \rightarrow S, S \rightarrow XSX \mid aY \mid a \mid SX \mid XS, X \rightarrow Y \mid S, Y \rightarrow b$
  - (ii) remove unit rule $S0 \rightarrow S$
    - $S0 \rightarrow XSX \mid aY \mid a \mid SX \mid XS, S \rightarrow XSX \mid aY \mid a \mid SX \mid XS, X \rightarrow Y \mid S, Y \rightarrow b$
  - (iii) remove unit rule $X \rightarrow Y$
    - $S0 \rightarrow XSX \mid aY \mid a \mid SX \mid XS, S \rightarrow XSX \mid aY \mid a \mid SX \mid XS, X \rightarrow S \mid b, Y \rightarrow b$
  - (iv) remove unit rule $X \rightarrow S$
    - $S0 \rightarrow XSX \mid aY \mid a \mid SX \mid XS, S \rightarrow XSX \mid aY \mid a \mid SX \mid XS, X \rightarrow b \mid XSX \mid aY \mid a \mid SX \mid XS, Y \rightarrow b$

- Step 4: Replace problematic terminals a by variable U with $U \rightarrow a$.
  - $S0 \rightarrow XSX \mid UY \mid a \mid SX \mid XS$
  - $S \rightarrow XSX \mid UY \mid a \mid SX \mid XS$
  - $X \rightarrow b \mid XSX \mid UY \mid a \mid SX \mid XS$
  - $Y \rightarrow b$
  - $U \rightarrow a$
- Shorten long RHS to sequence of RHS's with only 2 variables each
  - $S0 \rightarrow XX_1 \mid UY \mid a \mid SX \mid XS$
  - $S \rightarrow XX_1 \mid UY \mid a \mid SX \mid XS$
  - $X \rightarrow b \mid XX_1 \mid UY \mid a \mid SX \mid XS$
  - $Y \rightarrow b$
  - $U \rightarrow a$
  - $X_1 \rightarrow SX$ which is a CFG in Chomsky normal form.

April 3rd Exam 2

TRY YOURSELF!

- Convert the following CFG into CNF.

S → aAD

A→aB | bAB

B → b

D → d

X→a

CNF

S → XAD

A → XB | BAB

B → b

D → d

X → a

Y → XA

Z → BA

S → YD

A → XB | ZB

B → b

D → d

X → a

Y → XA

Z → BA

- Convert the following CFG into CNF.

S → 0A0 | 1B1 | BB

A → C

B → S | A

C → S | ε

- Convert the given CFG to CNF,

S → AbA

A → Aa | ε

$S \to 0A0 \mid 1B1 \mid BB$

$A \to C$

$B \to S \mid A$

$C \to S \mid \varepsilon$

$C \to \varepsilon$

$S \to 0A0 \mid 1B1 \mid BB$

$A \to \varepsilon \mid C$

$B \to S \mid A$

$C \to S$

$A \to \varepsilon$

$S \to 0A0 \mid 1B1 \mid BB \mid 00$

$A \to C$

$B \to S \mid A \mid \varepsilon$

$C \to S$

$B \to \varepsilon$

$S \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11 \mid B \mid \varepsilon$

$A \to C$

$B \to S \mid A$

$C \to S$

$S_0 \to S$

$S \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11 \mid B$

$A \to C$

$B \to S \mid A$

$C \to S$

Unit Pro.

$S_0 \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11$

~~$S \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11$~~

$A \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11$

~~$B \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11$~~

~~$C \to 0A0 \mid 1B1 \mid BB \mid 00 \mid 11$~~

$S_0 \to 0S_00 \mid 1S_01 \mid S_0S_0 \mid 00 \mid 11$

$X \to 0 \quad Y \to 1$

$S_0 \to XS_0X \mid YS_0Y \mid S_0S_0 \mid XX \mid YY$

$X \to 0 \quad Y \to 1 \quad Z \to XS_0 \quad U \to YS_0$

$S_0 \to ZX \mid UX \mid S_0S_0 \mid XX \mid YY$

$X \to 0 \quad Y \to 1 \quad Z \to XS_0 \quad U \to YS_0$

$S \rightarrow AbA$

$A \rightarrow Aa \mid \varepsilon$

$S \rightarrow AbA$

$A \rightarrow Aa \mid \varepsilon$

$A \rightarrow \varepsilon$

$S \rightarrow AbA \mid bA \mid Ab \mid b$

$A \rightarrow Aa \mid a$

$X \rightarrow a \quad Y \rightarrow b$

$S \rightarrow AYA \mid YA \mid AY \mid b$

$A \rightarrow AX \mid a$

$X \rightarrow a$

$Y \rightarrow b$

$Z \rightarrow AY$

$S \rightarrow ZA \mid YA \mid AY \mid b$

$A \rightarrow AX \mid a$

$X \rightarrow a$

$Y \rightarrow b$

$Z \rightarrow AY$

S → AaA

A → aaBa | CDA | CD

B → bB

C → Ca | D

D → bD | ε

SUMMARY

There are special forms for CFGs such as Chomsky Normal Form, where every production has the form $A \rightarrow BC$ or $A \rightarrow c$. The algorithm to convert to this form involves

(1) determining all nullable variables and getting rid of all ε-productions,

(2) getting rid of all variable unit productions,

(3) breaking up long productions, and

(4) moving terminals to unit productions.

GREIBACH NORMAL FORM (GNF)

GREIBACH NORMAL FORM (GNF)

- Every context free grammar/ CFL without ε can be generated by a grammar for which every production in the form of $A \rightarrow a\alpha$ and $A \rightarrow a$ where A is variable,a is a terminal and α is collection of variables
- GNF FORMAT
  - Non-terminal → terminal .any no of non-terminal
  - Non-terminal → terminal
- A CFG is in GNF if the productions are in form,
  - $A \rightarrow b$
  - $A \rightarrow bC_1C_2.....C_n$

- Steps:
  - Check if the given CFG has any UNIT productions or NULL productions and remove if any.
  - Check if CFG is already in CNF and convert it to CNF if not.
  - Change the names of nonterminal symbols into some A_i in ascending order of i.
- Example:
  - S→CA/BB
  - B→b/SB
  - C→b
  - A→a
- Step 1 and 2 are satisfied.
- Step 3: S,B,C,A → non-terminal symbols.
  - Replace, S with A_1
  - C with A_2
  - A with A_3
  - B with A_4

$A_1 \rightarrow A_2A_3/A_4A_4$
$A_4 \rightarrow b/A_1A_4$
$A_2 \rightarrow b$
$A_3 \rightarrow a$

- Step 4: After the rules so that the non-terminal are in ascending order such that, if the production is of the form $A_i \rightarrow A_jX$ then, i<j and should never be i>j.
  - $A_1 \rightarrow A_2A_3/A_4A_4$ => (1<2 and 1<4) => ☑
  - $A_4 \rightarrow b/A_1A_4$ => (4>1) => we need to resolve this.
  - Replace A_1
  - $A_4 \rightarrow b/ A_2A_3A_4/A_4A_4\,A_4$ => (4>2) => again we need to resolve this issue.
  - Replace value of A_2
  - $A_4 \rightarrow b/ bA_3A_4/A_4A_4\,A_4$ => (4≥4) => this needs to be resolved, called Left recursion.

    This is GNF

- Step 5: Remove Left Recursion
  - Introduce new variable to remove left recursion
  - $A_4 \rightarrow b/ bA_3A_4/A_4A_4\,A_4$
  - $Z \rightarrow A_4A_4Z/ A_4A_4$ => take variable that follow problematic ones and write it along with new variable.
  - Rewriting,
  - $A_4 \rightarrow b/ bA_3A_4/bZ/bA_3\,A_4Z$

- Step 5 continuation: We are not allowed to have variable in beginning
  - $A_1 \rightarrow A_2A_3/A_4\,A_4$
  - $A_4 \rightarrow b/\ bA_3A_4/bZ/bA_3\,A_4Z$
  - $Z \rightarrow A_4A_4Z/\ A_4A_4$
  - $A_2 \rightarrow b$
  - $A_3 \rightarrow a$
- Modifying:
  - $A_1 \rightarrow A_2A_3/A_4\,A_4$ => replace A_4
  - $A_1 \rightarrow bA_3/\ bA_4/bA_3\,A_4A_4/bZA_4/bA_3\,A_4ZA_4$
  - $A_4 \rightarrow b/\ bA_3A_4/bZ/bA_3\,A_4Z$
  - $Z \rightarrow bA_4/\ bA_3\,A_4A_4/bZA_4/bA_3\,A_4ZA_4/bA_4A_4z/bA_3A_4A_4Z/bzA_4Z/bA_3A_4ZA_4Z$
  - $A_2 \rightarrow b$
  - $A_3 \rightarrow a$

TRY YOURSELF!

- Convert the given CFG to GNF:
  - S→CA
  - A→a
  - C→aB/b
- CFG to GNF
  - S → AA | 0
  - A → SS | 1
- Convert the given CFG to GNF:
  - S→AB
  - A→BS/b
  - B→ SA