SIMPLIFICATION OF CONTEXT FREE GRAMMAR

Dr. A. Beulah

AP/CSE

LEARNING OBJECTIVE

- To Understand the need of formal languages, and grammars (K3)
  - To Understand simplification of CFG

THREE WAYS TO SIMPLIFY/CLEAN A CFG

- Clean
  - Eliminate useless symbols

- Simplify
  - Eliminate ε-productions $A \cancel{\Rightarrow} \varepsilon$ $A \rightarrow \varepsilon$
  - Eliminate unit productions $A \cancel{\Rightarrow} B$ $\underbrace{A}_{NT} \rightarrow \underbrace{B}_{NT}$

ELIMINATING USELESS SYMBOLS

Dr. A. Beulah

AP/CSE

LEARNING OBJECTIVE

- To eliminate useless symbols

FIND USEFUL SYMBOLS

- A symbol X is reachable if there exists:

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

  Reed

- A symbol X is generating if there exists:

  $X \Rightarrow^* w$,

  for some $w \in T^*$

- For a symbol X to be "useful", it has to be both reachable and generating

  - $S \Rightarrow^* \alpha X \beta \Rightarrow^* w'$, for some $w' \in T^*$

- Omitting useless symbols obviously will not change the language generated by the grammar.

ALGORITHM TO DETECT USELESS SYMBOLS

1. Eliminate all symbols that are *not* generating

2. Eliminate all symbols that are *not* reachable

Is the order of these steps important, or can we switch?

EXAMPLE

- S → AB | a
- A → b

i) NG   w ∈ T*

B → NG   'B'

(i) ~~S → AB~~ | a
~~A → b~~

(ii) S → a
A → b

X

- *A, S* are generating (S → a, A → b)
- *B is not generating* (and therefore B is useless)
- Eliminating B… (i.e., remove all productions that involve B)

  S → a
  A → b

$S \overset{*}{\Rightarrow} x$

- Now, A is *not reachable* and therefore is useless

(i) Non G
(ii) Non R.

- Simplified G:

  S → a

What would happen if you reverse the order:
i.e., test reachability before generating?

Will fail to remove:
A → b

ALGORITHM TO DETECT USELESS SYMBOLS

$S \rightarrow aSb \mid A \mid \varepsilon$

$A \rightarrow aA$

(i) Non Generating

$S \rightarrow \varepsilon$ ✓

$A \rightarrow aA$ X

NG = A

$S \rightarrow aSb / \varepsilon$

(ii) non reach

$S \rightarrow aSb / \varepsilon$

ALGORITHM TO DETECT USELESS SYMBOLS

$S \rightarrow aSb \mid A \mid \varepsilon$

$A \rightarrow aA$

$S \rightarrow aSb \mid \varepsilon$

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$
  $A \rightarrow aA \mid aF$
  $B \rightarrow CF \mid b$
  $C \rightarrow cC \mid D$
  $D \rightarrow aD \mid BD \mid C$
  $E \rightarrow aA \mid BSA$
  $F \rightarrow bB \mid b$

Find useless symbols

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$
  $A \rightarrow aA \mid aF$
  $B \rightarrow CF \mid b$
  $C \rightarrow cC \mid D$
  $D \rightarrow aD \mid BD \mid C$
  $E \rightarrow aA \mid BSA$
  $F \rightarrow bB \mid b$
- First Find generating symbols

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$
  $A \rightarrow aA \mid aF$
  $B \rightarrow CF \mid b$
  $C \rightarrow cC \mid D$
  $D \rightarrow aD \mid BD \mid C$
  $E \rightarrow aA \mid BSA$
  $F \rightarrow bB \mid b$
- Find generating symbols
- B, F both generate terminals

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$
  $A \rightarrow aA \mid aF$
  $B \rightarrow CF \mid b$
  $C \rightarrow cC \mid D$
  $D \rightarrow aD \mid BD \mid C$
  $E \rightarrow aA \mid BSA$
  $F \rightarrow bB \mid b$
- Find generating symbols
- B, F both generate symbols
- S is generate symbol, since $S \rightarrow B$ and hence $S \Rightarrow^* b$

EXAMPLE

- $S \to AC \mid BS \mid B$
  $A \to aA \mid aF$
  $B \to CF \mid b$
  $C \to cC \mid D$
  $D \to aD \mid BD \mid C$
  $E \to aA \mid BSA$
  $F \to bB \mid b$
- Find generating symbols
- B, F both generate symbols
- S is generate symbol, since $S \to B$ and hence $S \Rightarrow^* b$
- A is generate symbol, since $A \to aF$ and hence $A \Rightarrow^* ab$

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$
  $A \rightarrow aA \mid aF$
  $B \rightarrow CF \mid b$
  $C \rightarrow cC \mid D$
  $D \rightarrow aD \mid BD \mid C$
  $E \rightarrow aA \mid BSA$
  $F \rightarrow bB \mid b$
- Find generating symbols
- B, F both generate symbols
- S is generate symbol, since $S \rightarrow B$ and hence $S \Rightarrow^* b$
- A is generate symbol, since $A \rightarrow aF$ and hence $A \Rightarrow^* ab$
- E is generate symbol, since $E \rightarrow aA$ and hence $E \Rightarrow^* aab$

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$
  $A \rightarrow aA \mid aF$
  $B \rightarrow CF \mid b$
  $C \rightarrow cC \mid D$
  $D \rightarrow aD \mid BD \mid C$
  $E \rightarrow aA \mid BSA$
  $F \rightarrow bB \mid b$
- Find generating symbols
- *C* and *D* are not generating symbols , so all rules containing *C* and *D* are **removed**

EXAMPLE

- $S \rightarrow AC \mid BS \mid B$

  $A \rightarrow aA \mid aF$

  $B \rightarrow CF \mid b$

  $C \rightarrow cC \mid D$

  $D \rightarrow aD \mid BD \mid C$

  $E \rightarrow aA \mid BSA$

  $F \rightarrow bB \mid b$

$S \Rightarrow AC$
$\Rightarrow aFC$
$\Rightarrow abC$
$\Rightarrow abcC$

abD

ω ω

- Find generating symbols
- *C* and *D* are not generating symbols , so all rules containing *C* and *D* are **removed**

EXAMPLE

- The new grammar is

$$S \rightarrow BS \mid B$$
$$A \rightarrow aA \mid aF$$
$$B \rightarrow b$$
$$E \rightarrow aA \mid BSA$$
$$F \rightarrow bB \mid b$$

- All non-terminals generate terminal strings
- Next, find reachable symbols

EXAMPLE

- The new grammar is

$$S \rightarrow BS \mid B$$
$$A \rightarrow aA \mid aF$$
$$B \rightarrow b$$
$$E \rightarrow aA \mid BSA$$
$$F \rightarrow bB \mid b$$

- All non-terminals generate terminal strings
- Next, find reachable symbols
- *S is a* reachable symbol, since it is the start symbol

EXAMPLE

- The new grammar is

$$S \rightarrow BS \mid B$$
$$A \rightarrow aA \mid aF$$
$$B \rightarrow b$$
$$E \rightarrow aA \mid BSA$$
$$F \rightarrow bB \mid b$$

- All non-terminals generate terminal strings
- Next, find reachable symbols
- *S is a* reachable symbol, since it is the start symbol
- *B is a* reachable symbol , since $S \rightarrow BS$, and hence *B* is derivable from *S*

EXAMPLE

- The new grammar is

  $S \rightarrow BS \mid B$

  $A \rightarrow aA \mid aF$

  $B \rightarrow b$

  $E \rightarrow aA \mid BSA$

  $F \rightarrow bB \mid b$

- All non-terminals generate terminal strings
- Next, find reachable symbols

- *A*, *E*, and *F* can not be derived from *S* or *B*, so all rules containing *A*, *E* and *F* are removed

EXAMPLE

- The new grammar is

$$S \to BS \mid B$$

$$B \to b$$

- The set of terminals of G_U is $\{b\}$, a is removed since it does not occur in any string in the language of G_U

EXAMPLE

- S→aAa
- A→Sb
- A→bCC
- A→DaA
- C→abb
- C→DD
- E→aC
- D→aDA

G = {S, A, C, E} ⇒ S → aAa

A → Sb | bCC

C → abb

E → aC

R = {S, A}

$\boxed{\begin{array}{l} S \to aAa \\ A \to Sb \end{array}}$

ELIMINATING ε-PRODUCTIONS

- **It is *not* possible to eliminate ε-productions for languages which include ε in their word set**

$A \rightarrow \varepsilon$

- Theorem: If G=(V,T,P,S) is a CFG for a language L, then L-{ε} has a CFG without ε-productions

- A non-terminal symbol that can **derive** the *null string* (ε) is called **nullable**.

- Definition: A is "nullable" if $A \rightarrow^* \varepsilon$

NULLABLE NON-TERMINALS

- The set of nullable non-terminals of the grammar
  - $S \to ACA$

    $A \to aAa \mid B \mid C$

    $B \to bB \mid b$

    $C \to cC \mid \varepsilon$

    $A \to C \to \varepsilon$

    $A \Rightarrow^* \varepsilon$
  - C is nullable
    - since $C \to \boldsymbol{\varepsilon}$ and hence $C \Rightarrow^* \varepsilon$

NULLABLE NON-TERMINALS

- The set of nullable non-terminals of the grammar
  - $S \rightarrow ACA$

    $A \rightarrow aAa \mid B \mid C$

    $B \rightarrow bB \mid b$

    $C \rightarrow cC \mid \varepsilon$

$S \Rightarrow ACA$
$\Rightarrow \varepsilon CA$
$\Rightarrow \varepsilon A$
$S \overset{*}{\Rightarrow} \varepsilon$

  - C is nullable
    - since $C \rightarrow \boldsymbol{\varepsilon}$ and hence $C \Rightarrow^* \varepsilon$
  - A is nullable
    - since $A \rightarrow C$, and C is nullable

NULLABLE NON-TERMINALS

- The set of nullable non-terminals of the grammar
  - $S \to ACA$

    $A \to aAa \mid B \mid C$

    $B \to bB \mid b$

    $C \to cC \mid \varepsilon$

  - C is nullable
    - since $C \to \varepsilon$ and hence $C \Rightarrow^* \varepsilon$
  - A is nullable
    - since $A \to C$, and C is nullable
  - S is nullable since $S \to ACA$, and A and C are nullable

$N = \{S, A, C\}$

$C \to \varepsilon$

ALGORITHM TO DETECT ALL NULLABLE NT

- <u>Basis:</u>
  - If A → ε is a production in G, then A is nullable (note: A can still have other productions)
- <u>Induction:</u>
  - If there is a production B → $C_1C_2 \ldots C_k$, where *every* C_i is nullable, then B is also nullable

ELIMINATING ε-PRODUCTIONS

- If $\varepsilon \notin L(G)$, we can eliminate all productions $A \to \varepsilon$
- For every B referring to A:

$$\begin{array}{l} B \to \alpha\, A\, \beta \mid \ldots \\ A \to \varepsilon \mid \ldots \end{array} \quad \Longrightarrow \quad \begin{array}{l} B \to \alpha\, \beta \mid \alpha\, A\, \beta \mid \ldots \\ A \to \ldots \end{array}$$

- For example, if $B \to \varepsilon$ and $A \to BABa$
- Then after eliminating the rule $B \to \varepsilon$, new rules for A will be added
  - $A \to BABa$
  - $A \to ABa$
  - $A \to BAa$
  - $A \to Aa$

$A \to BABa$

$B \to \varepsilon$ x

B

$A \to BABa$

x ✓

$A \to Aa$

$A \to BAa$

$A \to ABa$

$A \to BABa$

EXAMPLE

- Let G be
  - $S \rightarrow SaB \mid aB$
    $B \rightarrow bB \mid \varepsilon$

$B \rightarrow \varepsilon$

B

$S \rightarrow SaB$

$S \rightarrow aB$

$B \rightarrow bB$

$S \rightarrow Sa / SaB$

$S \rightarrow a \mid aB$

$B \rightarrow bB / b$

EXAMPLE

- Let G be
  - $S \rightarrow SaB \mid aB$
    $B \rightarrow bB \mid \varepsilon$
- After removing ε -productions, the new grammar will be
  - $S \rightarrow SaB \mid Sa \mid aB \mid a$
    $B \rightarrow bB \mid b$

- The removal of e-productions *increases the number of rules* but *reduces the length of derivations*.

ELIMINATING ε-PRODUCTIONS

Let L be the language represented by the following CFG G:

S → AB
A → aAA | ε
B → bBB | ε

S → AB
S → ε | A | B | AB

Nullable symbols: {A, B}

New grammar constructed from G as follows:

B → b | bB | bB | bBB
==> B → b | bB | bBB
Similarly, A → a | aA | aAA
Similarly, S → A | B | AB

Simplified grammar

S → A | B | AB
A → a | aA | aAA
B → b | bB | bBB

+

S → ε

EXAMPLE

- Let G $S \rightarrow ACA$

$$A \rightarrow aAa \mid B \mid C$$
$$B \rightarrow bB \mid b$$
$$C \rightarrow cC \mid \varepsilon$$

EXAMPLE

- Let G $S \rightarrow ACA$
  $A \rightarrow aAa \mid B \mid C$
  $B \rightarrow bB \mid b$
  $C \rightarrow cC \mid \varepsilon$
- The equivalent essentially grammar is
  $S \rightarrow ACA \mid CA \mid AA \mid AC \mid A \mid C \mid \varepsilon$
  $A \rightarrow aAa \mid aa \mid B \mid C$
  $B \rightarrow bB \mid b$
  $C \rightarrow cC \mid c$
- Since $S \Rightarrow^* \varepsilon$ in G, the rule $S \rightarrow \varepsilon$ is allowed in G_L, but all other ε-productions are replaced

EXAMPLE

- Let G be
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB$
  - $B \rightarrow ab \mid aAbC \mid aAb \mid CC$
  - $C \rightarrow \varepsilon$

EXAMPLE

- Let G be
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB$
  - $B \rightarrow ab \mid aAbC \mid aAb \mid CC$
  - $C \rightarrow \varepsilon$
- We eliminate $C \rightarrow \varepsilon$ by replacing:
  - $B \rightarrow CC$ into $B \rightarrow CC$, $B \rightarrow C$, and $B \rightarrow \varepsilon$
  - $B \rightarrow aAbC$ into $B \rightarrow aAbC$ and $B \rightarrow aAb$

- Since $C \rightarrow \varepsilon$ is only C production
  - only $B \rightarrow \varepsilon$ and $B \rightarrow aAb$ retained.

- The new grammar:
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB$
  - $B \rightarrow \varepsilon \mid ab \mid aAb$

EXAMPLE

- The new grammar:
  - $S \to aS \mid SS \mid bA$
  - $A \to BB$
  - $B \to \varepsilon \mid ab \mid aAb$

EXAMPLE

- The new grammar:
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB$
  - $B \rightarrow \varepsilon \mid ab \mid aAb$
- We eliminate $B \rightarrow \varepsilon$ by replacing
  - $A \rightarrow BB$ into $A \rightarrow BB$, $A \rightarrow B$, and $A \rightarrow \varepsilon$
- Since there are other B productions, these are all retained
- The new grammar:
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB \mid B \mid \varepsilon$
  - $B \rightarrow ab \mid aAb$

EXAMPLE

- The new grammar:
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB \mid B \mid \varepsilon$
  - $B \rightarrow ab \mid aAb$

EXAMPLE

- The new grammar:
  - $S \rightarrow aS \mid SS \mid bA$
  - $A \rightarrow BB \mid B \mid \varepsilon$
  - $B \rightarrow ab \mid aAb$

- Finally we eliminate $A \rightarrow \varepsilon$ by replacing
  - $B \rightarrow aAb$ into $B \rightarrow aAb, B \rightarrow ab$
  - $S \rightarrow bA$ into $S \rightarrow bA \mid b$

- The final CFG is:
  - $S \rightarrow aS \mid SS \mid bA \mid b$
  - $A \rightarrow BB \mid B$
  - $B \rightarrow ab \mid aAb$

ELIMINATING ε-PRODUCTIONS

<u>Given:</u> $G=(V,T,P,S)$

<u>Algorithm:</u>

1. Detect all nullable variables in G
2. Then construct $G_1=(V,T,P_1,S)$ as follows:
   - For each production of the form: $A \rightarrow X_1X_2\ldots X_k$, where $k\geq1$, suppose ***m*** out of the ***k*** X_i's are nullable symbols, then G_1 will have **2^m** versions for this production

     i.e, all combinations where each X_i is either present or absent
   - Alternatively, if a production is of the form: $A \rightarrow \varepsilon$, then remove it

ELIMINATING UNIT PRODUCTIONS

- Rules having the form $A \rightarrow B$ are called **unit rules**
- Consider the rules
  - $A \rightarrow aA \mid a \mid B$
  - $B \rightarrow bB \mid b \mid C$

Prod ↑
Der. ↓

- The unit rule $A \rightarrow B$ indicates that any string derivable from B is also derivable from A
- The **removal of unit** rules *increases the number of rules* but *reduces the length of derivations*.

ELIMINATING UNIT PRODUCTIONS

- To eliminate the unit rule, add *A* rules that directly generate the same strings as *B*
  - Add a rule $A \rightarrow u$ for each $B \rightarrow u$ and deleting $A \rightarrow B$ from the grammar

$$\begin{aligned} A &\rightarrow B \\ B &\rightarrow \alpha \mid \ldots \end{aligned} \quad \Longrightarrow \quad \begin{aligned} A &\rightarrow \alpha \mid \ldots \\ B &\rightarrow \alpha \mid \ldots \end{aligned}$$

EXAMPLE

- Consider the productions
  - $A \rightarrow aA \mid a \mid B$
  - $B \rightarrow bB \mid b \mid d$

$A \rightarrow B$

$B \rightarrow bB$
$B \rightarrow b$
$B \rightarrow d$

$A \rightarrow bB \mid b \mid d$

- The new productions after eliminating the unit production $A \rightarrow B$
  - $A \rightarrow aA \mid a \mid bB \mid b \mid d$
  - $B \rightarrow bB \mid b \mid d$

$A \rightarrow aA \mid a \mid bB \mid b \mid d$
$B \rightarrow bB \mid b \mid d$

- We add new rules to *A* by replacing *B* in *A* with all its *RHS* rules

EXAMPLE

- $S \rightarrow ACA \mid CA \mid AA \mid AC \mid A \mid C \mid \varepsilon$

  $A \rightarrow aAa \mid aa \mid B \mid C$

  $B \rightarrow bB \mid b$

  $C \rightarrow cC \mid c$

$S \rightarrow A$
$S \rightarrow C$
$A \rightarrow B$
$A \rightarrow C$

$A \rightarrow C$

$A \rightarrow cC \mid c$

$A \rightarrow B$

$A \rightarrow bB \mid b$

$S \rightarrow C$

$S \rightarrow cC \mid c$

$S \rightarrow A$

$S \rightarrow aAa \mid aa \mid cC \mid c \mid bB \mid b$

EXAMPLE

- $S \to ACA \mid CA \mid AA \mid AC \mid A \mid C \mid \varepsilon$
  $A \to aAa \mid aa \mid B \mid C$
  $B \to bB \mid b$
  $C \to cC \mid c$

- The new equivalent grammar (without unit rules)
- $S \to \underline{ACA \mid CA \mid AA \mid AC} \mid \underline{aAa \mid aa \mid bB \mid b \mid cC \mid c} \mid \varepsilon$
  $A \to aAa \mid aa \mid \underline{bB \mid b \mid cC \mid c}$
  $B \to bB \mid b$
  $C \to cC \mid c$

↑ produ

↓ deriva

EXAMPLE

- $E \rightarrow T \mid E + T$
- $T \rightarrow F \mid F * T$
- $F \rightarrow a \mid (E)$

EXAMPLE

- $E \rightarrow T \mid E + T$
- $T \rightarrow F \mid F * T$
- $F \rightarrow a \mid (E)$

- $E \rightarrow T \mid E + T$
- $T \rightarrow a \mid (E) \mid F * T$
- $F \rightarrow a \mid (E)$

EXAMPLE

- $E \rightarrow T \mid E + T$
- $T \rightarrow F \mid F * T$
- $F \rightarrow a \mid (E)$

- $E \rightarrow T \mid E + T$
- $T \rightarrow a \mid (E) \mid F * T$
- $F \rightarrow a \mid (E)$

- $E \rightarrow a \mid (E) \mid F * T \mid E + T$
- $T \rightarrow a \mid (E) \mid F * T$
- $F \rightarrow a \mid (E)$

TEST YOUR KNOWLEDGE

- Suppose A→xBz and B → y, then the simplified grammar would be:

  a) A → xyz

  b) A → xBz|xyz

  c) A → xBz|B|y

  d) none of the mentioned

- Given Grammar: S → A, A → aA, A → e, B → bA

  Which among the following productions are Useless productions?

  a) S → A

  b) A → aA

  c) A → e

  d) B → bA

TEST YOUR KNOWLEDGE

- The Grammar can be defined as: G=(V, ∑, p, S)
  In the given definition, what does S represents?
  a) Accepting State
  b) Starting Variable
  c) Sensitive Grammar
  d) None of these

LEARNING OUTCOME

On successful completion of this topic, the student will be able to:

- Understand simplification of CFG (K3)

REFERENCE

- Hopcroft J.E., Motwani R. and Ullman J.D, “Introduction to Automata Theory, Languages and Computations”, Second Edition, Pearson Education, 2008