Chomsky Normal Form (of CFGs)

Previous two lectures

- Introduced context free grammars and CFLs
- Formal definition of CFGs
- Examples of CFGs
- *Designing CFGs*
  - Four techniques:
    1. Break down into simpler pieces
    2. Use DFAs if language at hand is regular
    3. Handle links ($R \longrightarrow uRv$)
    4. Handle recursion (e.g. G_4 : arithmetic expression)
- *Ambiguity*
- **Today**: Chomsky normal form

Chomsky Normal Form

DEFINITION 2.8

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. In addition, we permit the rule $S \rightarrow \varepsilon$, where S is the start variable.

Chomsky Normal Form

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

Proof Idea: We can convert any grammar G into CNF

The Conversion procedure (overview):

The conversion has several stages wherein rules that violate the conditions are replaced with equivalent ones that are satisfactory.

- First, we add a new start variable
- Then, we eliminate all **ε-rules** of the form $A \longrightarrow \varepsilon$
- We also eliminate all **unit rules** of the form $A \longrightarrow B$
- In both cases, we patch up the grammar to be sure that it still generates the same language
- Finally, we convert the remaining rules into the proper form

The conversion procedure: details

1. First, we add a **new start variable** S_0 and the rule $S_0 \longrightarrow S$, where S was the original start variable. (This change guarantees that the start variable doesn't occur on the RHS of the rule).
2. We take care of all **ε-rules**.
   a) We remove an ε-rule $A \longrightarrow \varepsilon$, where A is not the start variable.
   b) Then for each occurrence of A on the RHS of a rule, we add a new rule with that occurrence deleted. In other words, if $R \longrightarrow uAv$ is a rule in which u and v are strings of variables and terminals, we add rule $R \longrightarrow uv$.

   We do so for each occurrence of an A, so the rule $R \longrightarrow uAvAw$ causes us to add

   $R \longrightarrow uvAw$,

   $R \longrightarrow uAvw$, and

   $R \longrightarrow uvw$.

   If we have the rule $R \longrightarrow A$, we add $R \longrightarrow \varepsilon$ unless we had previously removed the rule $R \longrightarrow \varepsilon$

   We repeat these steps until we eliminate all ε-rules not involving the start variable

The conversion procedure: details

3. We handle all **unit** rules.
   a) We remove a unit rule $A \longrightarrow B$.
   b) Then whenever a rule $B \longrightarrow u$ appears, we add the rule $A \longrightarrow u$ unless this was a unit rule previously removed. As before, u is a string of variables and terminals.

   We repeat these steps until we eliminate all unit rules.

4. We convert **all remaining rules** into the proper form.
   a) We replace each rule $A \longrightarrow u_1\ u_2\ u_k$, where $k \geq 3$ and each u_i is a variable or terminal symbol, with the rules $A \longrightarrow u_1A_1$ $A_1 \longrightarrow u_2A_2$ $A_2 \longrightarrow u_3A_3$, …, $A_{k-2} \longrightarrow u_{k-1}u_k$.

      The A_i's are new variables.
   b) We replace any terminal u_i in the preceding rule(s) with the new variable U_i and add the rule $U_i \longrightarrow u_i$

Example 1

Convert the grammar G_6 given below into CNF

$$
\begin{aligned}
S &\to ASA \mid \mathtt{a}B \\
A &\to B \mid S \\
B &\to \mathtt{b} \mid \varepsilon
\end{aligned}
$$

Example 1

1. The original CFG G_6 is shown on the left. The result of applying the first step to make a new start variable appears on the right.

$$
\begin{aligned}
S &\to ASA \mid \mathtt{a}B \\
A &\to B \mid S \\
B &\to \mathtt{b} \mid \varepsilon
\end{aligned}
\qquad\qquad
\begin{aligned}
\boldsymbol{S_0} &\boldsymbol{\to} \boldsymbol{S} \\
S &\to ASA \mid \mathtt{a}B \\
A &\to B \mid S \\
B &\to \mathtt{b} \mid \varepsilon
\end{aligned}
$$

Note: In this step and subsequent steps, Rules shown in **bold** have just been added. Rules shown in grey have just been removed.

Example 1

2. Remove ε-rules $B \to \varepsilon$, shown on the left, and $A \to \varepsilon$, shown on the right.

$$\begin{aligned} S_0 &\to S \\ S &\to ASA \mid \mathtt{a}B \mid \mathbf{a} \\ A &\to B \mid S \mid \boldsymbol{\varepsilon} \\ B &\to \mathtt{b} \mid \varepsilon \end{aligned}$$

$$\begin{aligned} S_0 &\to S \\ S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \mid \boldsymbol{S} \\ A &\to B \mid S \mid \varepsilon \\ B &\to \mathtt{b} \end{aligned}$$

Example 1

3a. Remove unit rules $S \to S$, shown on the left, and $S_0 \to S$, shown on the right.

$$\begin{aligned} S_0 &\to S \\ S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \mid S \\ A &\to B \mid S \\ B &\to \mathtt{b} \end{aligned}$$

$$\begin{aligned} S_0 &\to S \mid \boldsymbol{ASA} \mid \boldsymbol{\mathtt{a}B} \mid \boldsymbol{\mathtt{a}} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \\ S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\ A &\to B \mid S \\ B &\to \mathtt{b} \end{aligned}$$

Example 1

3b. Remove unit rules $A \to B$ and $A \to S$.

$$
\begin{aligned}
S_0 &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
A &\to B \mid S \mid \mathbf{b} \\
B &\to \mathtt{b}
\end{aligned}
\qquad
\begin{aligned}
S_0 &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
S &\to ASA \mid \mathtt{a}B \mid \mathtt{a} \mid SA \mid AS \\
A &\to S \mid \mathtt{b} \mid \boldsymbol{ASA} \mid \mathbf{a}\boldsymbol{B} \mid \mathbf{a} \mid \boldsymbol{SA} \mid \boldsymbol{AS} \\
B &\to \mathtt{b}
\end{aligned}
$$

Example 1

4. Convert the remaining rules into the proper form by adding additional variables and rules. The final grammar in Chomsky normal form is equivalent to G_6. (Actually the procedure given in Theorem 2.9 produces several variables U_i and several rules $U_i \to \mathtt{a}$. We simplified the resulting grammar by using a single variable U and rule $U \to \mathtt{a}$.)

$$
\begin{aligned}
S_0 &\to AA_1 \mid UB \mid \mathtt{a} \mid SA \mid AS \\
S &\to AA_1 \mid UB \mid \mathtt{a} \mid SA \mid AS \\
A &\to \mathtt{b} \mid AA_1 \mid UB \mid \mathtt{a} \mid SA \mid AS \\
A_1 &\to SA \\
U &\to \mathtt{a} \\
B &\to \mathtt{b}
\end{aligned}
$$

Example 2

Convert the grammar G_7 given below into CNF

$$S \rightarrow 0R0 | 1R1 | \epsilon$$

$$R \rightarrow 0R | 1R | \epsilon$$

Example 2

a) We remove an ε-rule $A \longrightarrow \varepsilon$, where A is not the start variable.
b) Then for each occurrence of A on the RHS of a rule, we add a new rule with that occurrence deleted. In other words, if $R \longrightarrow uAv$ is a rule in which u and v are strings of variables and terminals, we add rule $R \longrightarrow uv$.

Original grammar:

$$S \to 0R0|1R1|\epsilon$$
$$R \to 0R|1R|\epsilon$$

Create new start symbol:

$$S_0 \to S$$
$$S \to 0R0|1R1|\epsilon$$
$$R \to 0R|1R|\epsilon$$

Remove epsilon rules:

$$S_0 \to S|\epsilon$$
$$S \to 0R0|1R1$$
$$R \to 0R|1R|0|1$$

a) We remove a unit rule $A \longrightarrow B$.
b) Then whenever a rule $B \longrightarrow u$ appears, we add the rule $A \longrightarrow u$ unless this was a unit rule previously removed. As before, u is a string of variables and terminals.

Example 2

Remove unit rules:

$$S_0 \to 0R0|1R1|\epsilon$$
$$S \to 0R0|1R1$$
$$R \to 0R|1R|0|1$$

Example 2

a) We replace each rule $A \longrightarrow u_1\ u_2\ u_k$, where $k \geq 3$ and each u_i is a variable or terminal symbol, with the rules $A \longrightarrow u_1A_1$ $A_1 \longrightarrow u_2A_2$ $A_2 \longrightarrow u_3A_3$, …, $A_{k-2} \longrightarrow u_{k-1}u_k$.
The A_i's are new variables.

b) We replace any terminal u_i in the preceding rule(s) with the new variable U_i and add the rule $U_i \longrightarrow u_i$

Convert the remaining rules by adding additional variables and rules:

$$
\begin{aligned}
S_0 &\to XZ|YO|\epsilon \\
R &\to ZR|OR|0|1 \\
X &\to ZR \\
Y &\to OR \\
O &\to 1 \\
Z &\to 0
\end{aligned}
$$

Addendum

- Homework 5
- Return midterm1 papers