Chomsky normal form

Back (for review), by popular demand
What is Chomsky Normal Form?

It’s a way to express the rules of a CFG

Every CFG may be written in normal form
What is the structure of Chomsky Normal Form?

CFG is in Chomsky normal form if every rule takes form:

- \( A \rightarrow BC \)
- \( A \rightarrow a \)

- B and C may not be the start variable
- Only the start variable can transition to \( \varepsilon \)
- Each variable goes to two other variables or two one terminal
- The start variable may not point back to the start variable
Consider a typical CFG

- Variables and terminals mix
  - • A -> xBy
- Some variables point to other single variables
  - • A -> C
- Start variable can point to itself
  - • S -> SS | y
- Any variable can transition to ε
  - • B -> ε
Converting to Chomsky Normal Form

• \( S_0 \rightarrow S \) where \( S \) was original start variable

• Remove \( A \rightarrow \epsilon \)

• For each multiple-occurrence of \( A \), add new rules with \( A \) deleted
  \[
  R \rightarrow uAvAw \quad \text{change to} \quad R \rightarrow uvAw \mid uAvw \mid uvw
  \]

• Shortcut all unit rules
  Given \( A \rightarrow B \) and \( B \rightarrow u \), add \( A \rightarrow u \)

• Replace rules \( A \rightarrow u_1u_2u_3 \ldots u_k \) with:
  \[
  A \rightarrow u_1A_1, A_1 \rightarrow u_2A_2, A_2 \rightarrow u_3A_3, \ldots, A_{k-2} \rightarrow u_{k-1}u_k
  \]
Let’s Chomsky-ize a non-Chomsky form grammar

G1:
S -> AB
A -> cAn | c
B -> BB | n

Convert 3-variable rules to 2-variable rules

S -> AB
A -> U_cD | c
B -> BB | n
U_c -> c
U_n -> n
D -> AU_n

Replace terminals-variable mixes with variables only

S -> AB
A -> U_cA U_n | c
B -> BB | n
U_c -> c
U_n -> n
More on G1:

G1:

S -> AB
A -> $U_c D | c$
B -> BB | n
$U_c$ -> c
$U_n$ -> n
D -> AU_n

This already fits normal form!

Typical to add new start state

Replace S in $S_0 -> S$ rule

Final answer!

<table>
<thead>
<tr>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0$ -&gt; $S$</td>
</tr>
<tr>
<td>$S$ -&gt; AB</td>
</tr>
<tr>
<td>$A$ -&gt; $U_c D</td>
</tr>
<tr>
<td>$B$ -&gt; BB</td>
</tr>
<tr>
<td>$U_c$ -&gt; c</td>
</tr>
<tr>
<td>$U_n$ -&gt; n</td>
</tr>
<tr>
<td>$D$ -&gt; AU_n</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0$ -&gt; AB</td>
</tr>
<tr>
<td>$S$ -&gt; AB</td>
</tr>
<tr>
<td>$A$ -&gt; $U_c D</td>
</tr>
<tr>
<td>$B$ -&gt; BB</td>
</tr>
<tr>
<td>$U_c$ -&gt; c</td>
</tr>
<tr>
<td>$U_n$ -&gt; n</td>
</tr>
<tr>
<td>$D$ -&gt; AU_n</td>
</tr>
</tbody>
</table>
Reminder:

CFG is in Chomsky normal form if every rule takes form:

\[
A \rightarrow BC \\
A \rightarrow a
\]

• B and C may not be the start variable
• Only the start variable can transition to ε

• Each variable goes to two other variables or two one terminal
• The start variable may not point back to the start variable
Let’s get more complicated with grammar G2

G2:
S -> AB
A -> cAn | c | ε
B -> BB | n

Let’s add new start state first this time:
S₀ -> S
S -> AB
A -> cAn | c | ε
B -> BB | n

To remove ε, first plug it in wherever it applies
More on G2

Finish removing $\varepsilon$

$S_0 \rightarrow S$

$S \rightarrow AB \mid B$

$S \rightarrow AB \mid \varepsilon B$

$A \rightarrow cAn \mid c \mid c\varepsilon n$

$B \rightarrow BB \mid n$

Replace terminals-variable mixes with variables only

$S_0 \rightarrow S$

$S \rightarrow AB \mid B$

$A \rightarrow U_c A U_n \mid c \mid U_c U_n$

$B \rightarrow BB \mid n$

$U_c \rightarrow c$

$U_n \rightarrow n$
More on G2

Convert 3-variable rules to 2-variable rules

Replace single variables on right side (S, B)

\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow AB | B \\
A & \rightarrow U_c A U_n | c | U_c U_n \\
B & \rightarrow BB | n \\
U_c & \rightarrow c \\
U_n & \rightarrow n \\
D & \rightarrow AU_n
\end{align*}
\]
G2 final answer

\[
\begin{align*}
S_0 &\rightarrow AB \mid BB \mid n \\
S &\rightarrow AB \mid BB \mid n \\
A &\rightarrow U_c D \mid c \mid U_c U_n \\
B &\rightarrow BB \mid n \\
U_c &\rightarrow c \\
U_n &\rightarrow n \\
D &\rightarrow AU_n
\end{align*}
\]

Produces same CFL as:

\[
\begin{align*}
S &\rightarrow AB \\
A &\rightarrow cAn \mid c \mid \varepsilon \\
B &\rightarrow BB \mid n
\end{align*}
\]
Reminder:

CFG is in Chomsky normal form if every rule takes form:

\[ A \rightarrow BC \]
\[ A \rightarrow a \]

• B and C may not be the start variable
• Only the start variable can transition to \( \varepsilon \)

• Each variable goes to two other variables or two one terminal
• The start variable may not point back to the start variable
Let’s get even more complicated

G3:

S \rightarrow AB \mid BS
A \rightarrow cAn \mid c \mid \varepsilon
B \rightarrow BB \mid n

Add new start state:

S_0 \rightarrow S

Remove \varepsilon

S_0 \rightarrow S
S \rightarrow AB \mid BS \mid B
A \rightarrow cAn \mid c \mid cn
B \rightarrow BB \mid n
More on G3

Replace terminals-
variable mixes with
variables only

$$S_0 \rightarrow S$$

$$S \rightarrow AB \mid BS \mid B$$

$$A \rightarrow cAn \mid c \mid cn$$

$$B \rightarrow BB \mid n$$

$$S_0 \rightarrow S$$

$$S \rightarrow AB \mid BS \mid B$$

$$A \rightarrow U_cA U_n \mid c \mid U_c U_n$$

$$B \rightarrow BB \mid n$$

$$U_c \rightarrow c$$

$$U_n \rightarrow n$$
More on G3

Replace single variables on right side (S, B)

\[
\begin{align*}
S_0 & \rightarrow S \\
S & \rightarrow AB \mid BS \mid B \\
A & \rightarrow U_cA U_n \mid c \mid U_c U_n \\
B & \rightarrow BB \mid n \\
U_c & \rightarrow c \\
U_n & \rightarrow n
\end{align*}
\]

\[
\begin{align*}
S_0 & \rightarrow AB \mid BS \mid BB \mid n \\
S & \rightarrow AB \mid BS \mid BB \mid n \\
A & \rightarrow U_cA U_n \mid c \mid U_c U_n \\
B & \rightarrow BB \mid n \\
U_c & \rightarrow c \\
U_n & \rightarrow n
\end{align*}
\]
**More on G3**

Convert 3-variable rules to 2-variable rules

<table>
<thead>
<tr>
<th>Original 3-variable rules</th>
<th>New 2-variable rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0 \rightarrow AB \mid BS \mid BB \mid n$</td>
<td>$S_0 \rightarrow AB \mid BS \mid BB \mid n$</td>
</tr>
<tr>
<td>$S \rightarrow AB \mid BS \mid BB \mid n$</td>
<td>$S \rightarrow AB \mid BS \mid BB \mid n$</td>
</tr>
<tr>
<td>$A \rightarrow U_c A U_n \mid c \mid U_c U_n$</td>
<td>$A \rightarrow U_c D \mid c \mid U_c U_n$</td>
</tr>
<tr>
<td>$B \rightarrow BB \mid n$</td>
<td>$B \rightarrow BB \mid n$</td>
</tr>
<tr>
<td>$U_c \rightarrow c$</td>
<td>$U_c \rightarrow c$</td>
</tr>
<tr>
<td>$U_n \rightarrow n$</td>
<td>$U_n \rightarrow n$</td>
</tr>
<tr>
<td>$D \rightarrow AU_n$</td>
<td>$D \rightarrow AU_n$</td>
</tr>
</tbody>
</table>
Reminder:

CFG is in Chomsky normal form if every rule takes form:

\[ A \rightarrow BC \]
\[ A \rightarrow a \]

• B and C may not be the start variable
• Only the start variable can transition to \( \varepsilon \)
• Each variable goes to two other variables or two one terminal
• The start variable may not point back to the start variable