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 ε
- 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 $A \rightarrow \varepsilon$

We won't use this rule in this class, will use all others

- For each multiple-occurrence of A, add new rules with A deleted

  $R \rightarrow uAvAw$ change to $R \rightarrow uvAw \mid uAvw \mid uvw$

- Shortcut all unit rules

  Given $A \rightarrow B$ and $B \rightarrow u$, add $A \rightarrow u$

- Replace rules $A \rightarrow u_1 u_2 u_3 \ldots u_k$ with:

  $A \rightarrow u_1 A_1, A_1 \rightarrow u_2 A_2, A_2 \rightarrow u_3 A_3, \ldots, A_{k-2} \rightarrow u_{k-1} u_k$

Let's Chomsky-ize a non-Chomsky form grammar

G1:

S -> AB

A -> cAn | c

B -> BB | n

Replace terminals-variable mixes with variables only

S -> AB

A -> $U_c A U_n$ | c

B -> BB | n

U_c -> c

U_n -> n

Convert 3-variable rules to 2-variable rules

S -> AB

A -> $U_c D$ | c

B -> BB | n

U_c -> c

U_n -> n

D -> $A U_n$

More on G1:

G1:

S -> AB

A -> **U_cD** | c

B -> BB | n

U_c -> c

U_n -> n

D -> AU_n

This already fits normal form!

Typical to add new start state

S_0 -> S

S -> AB

A -> **U_cD** | c

B -> BB | n

U_c -> c

U_n -> n

D -> AU_n

Replace S in S_0 -> S rule

Final answer!

S_0 -> AB

S -> AB

A -> **U_cD** | c

B -> BB | n

U_c -> c

U_n -> n

D -> AU_n

Reminder:

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

$$A \to BC$$

$$A \to a$$

- 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:

S_0 -> S

S -> AB

A -> cAn | c | ε

B -> BB | n

To remove ε, first plug it in wherever it applies

S_0 -> S

S -> AB | **ε**B

A -> cAn | c | ε | c**ε**n

B -> BB | n

More on G2

S_0 -> S
S -> AB | εB
A -> cAn | c | ε | cεn
B -> BB | n

Finish removing ε

S_0 -> S
S -> AB | B
A -> cAn | c | cn
B -> BB | n

Replace terminals-variable mixes with variables only

S_0 -> S
S -> AB | B
A -> U_c A U_n | c | U_c U_n
B -> BB | n
U_c -> c
U_n -> n

More on G2

S_0 -> S
S -> AB | B
A -> U_c A U_n | c | $U_c U_n$
B -> BB | n
U_c -> c
U_n -> n

Convert 3-variable rules to 2-variable rules

S_0 -> S
S -> AB | B
A -> U_cD | c | $U_c U_n$
B -> BB | n
U_c -> c
U_n -> n
D -> AU_n

Replace single variables on right side (S, B)

S_0 -> AB | BB | n
S -> AB | BB | n
A -> U_cD | c | $U_c U_n$
B -> BB | n
U_c -> c
U_n -> n
D -> AU_n

G2 final answer

S_0 -> AB | BB | n

S -> AB | **BB | n**

A -> **U_cD** | c | **U_c U_n**

B -> BB | n

U_c -> c

U_n -> n

D -> AU_n

Produces same CFL as:

S -> AB

A -> cAn | c | ε

B -> BB | n

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 even **more** complicated

G3:

S -> AB | BS

A -> cAn | c | ε

B -> BB | n

Add new start state:

S_0 -> S

S -> AB | BS

A -> cAn | c | ε

B -> BB | n

Remove ε

S_0 -> S

S -> AB | BS | **B**

A -> cAn | c | **cn**

B -> BB | n

More on G3

Replace terminals-variable mixes with variables only

$\mathbf{S_0 \text{ -> } S}$

S -> AB | BS | **B**

A -> cAn | c | **cn**

B -> BB | n

$\mathbf{S_0 \text{ -> } S}$

S -> AB | BS | **B**

A -> $\mathbf{U_c}$A$\mathbf{U_n}$ | c | $\mathbf{U_cU_n}$

B -> BB | n

$\mathbf{U_c \text{ -> } c}$

$\mathbf{U_n \text{ -> } n}$

More on G3

Replace single variables on right side (S, B)

S_0 -> S	**S_0 -> AB | BS | BB | n**
S -> AB | BS | **B**	S -> AB | BS | **BB | n**
A -> **U_c**A**U_n** | c | **U_cU_n**	A -> **U_c**A**U_n** | c | **U_cU_n**
B -> BB | n	B -> BB | n
U_c -> c	**U_c -> c**
U_n -> n	**U_n -> n**

More on G3

Convert 3-variable rules to 2-variable rules

S_0 -> AB | BS | BB | n
S -> AB | BS | **BB | n**
A -> **U_cAU_n** | c | **U_cU_n**
B -> BB | n
U_c -> c
U_n -> n

S_0 -> AB | BS | BB | n
S -> AB | BS | **BB | n**
A -> **U_cD** | c | **U_cU_n**
B -> BB | n
U_c -> c
U_n -> n
D -> AU_n

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