CS310

Chomsky Normal Form

Section: 2.1

October 14, 2016

Quick Review

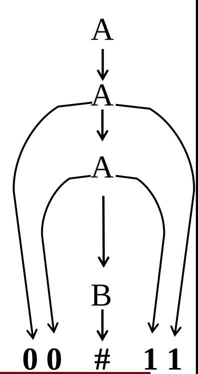
- (CFG) 4-tuple (V, ∑, R, S)
- –V finite set of variables
- –∑ finite set of terminals
- –R set of rules of form:
- variable **->** (string of variables and terminals)
- –S ∈ V, start variable
- –L(G) = { w ∈ ∑* | S → * w}
- w is in ∑* and can be derived from S

Example

A → 0A1

A → B

B → #



Chomsky Normal Form

•CNF presents a grammar in a standard, simplified form:

$A \rightarrow BC$

$A \rightarrow a$

1. where A,B,C are variables and B and C are not the start variable AND
2. *a* is a terminal

NOTE: The rule S →ε (S Start Variable) is also allowed so the language can generate the empty string (optional)

Chomsky Normal Form Examples

Which CFG's are in Chomsky Normal Form?

- $S \rightarrow XM$

  $M \rightarrow SY$

  $X \rightarrow x$

  $Y \rightarrow y$

- $S \rightarrow xSy$

  $S \rightarrow xy$

CNF Benefits

- Easier to prove statements about CFG's when in CNF
- Any CFG can be converted to CNF
- Remove productions:
  1. $A \rightarrow \varepsilon$ to empty
  2. $A \rightarrow B$ Unit rule
  3. $A \rightarrow s$, s contains a terminal and $|s| > 1$
  4. $A \rightarrow s$, $|s| > 2$, $s \in \{ V \cup \Sigma \}^*$

Step 1: Removing $A \rightarrow \varepsilon$

Example:

$S \rightarrow UAV$

$A \rightarrow \varepsilon$

In CNF? If not, how to fix?

Nullable variables

•A variable A is ***nullable*** if $A \rightarrow^{*} \varepsilon$

So we must:

- Find all nullable variables
- Remove all **ε** transitions
- If $T \rightarrow X_1AX_2$ and A is nullable then add $T \rightarrow X_1X_2$

Step 1 Example

$S \rightarrow TU$

$T \rightarrow AB$

$A \rightarrow aA \mid \varepsilon$

$B \rightarrow bB \mid \varepsilon$

$U \rightarrow ccA \mid B$

Nullable variables?

Productions removed?

Productions added?

Step 2: Removing A → B (Unit Productions)

A → B Must add A → s to compensate!

B → s

s ∈ { V U ∑ }*

- A variable B is **A-derivable** if A→ ***B**
- Must:
  - Find all A-derivable variables for each A
  - Remove all unit transitions
  - If B → s and B is A-derivable then add A → s

Step 2 Example

S → TU | T | U B → bB | b

T → AB | A | B U → ccA | B | cc

A → aA | a

S-derivable:

T-derivable:

U-derivable:

NOTE: Not concerned about A, B since not involved in unit production

Step 2 Example, cont.

$S \rightarrow TU \mid T \mid U$ $\quad B \rightarrow bB \mid b$

$T \rightarrow AB \mid A \mid B$ $\quad U \rightarrow ccA \mid B \mid cc$

$A \rightarrow aA \mid a$

Productions removed:

Productions added:

Step 2 Example, cont.

$S \rightarrow TU \mid T \mid U \mid \varepsilon$ $\quad B \rightarrow bB \mid b$

$T \rightarrow AB \mid A \mid B$ $\quad U \rightarrow ccA \mid B \mid cc$

$A \rightarrow aA \mid a$

Productions that still violate CNF?:

Steps 3,4 Example, cont.

Grammar should now have NO unit productions or null productions $A \rightarrow \varepsilon$ (except A may be start variable in null production)

Still need to fix productions with terminals on the RHS

Steps 3, 4: Remove $A \rightarrow S_1aS_2$

$A \rightarrow S_1aS_2$

$a \in \sum$, S_1 and S_2 are strings, at least one is not empty

Create

$X_a \rightarrow a$

$A \rightarrow S_1X_aS_2$

Then fix up $A \rightarrow S_1X_aS_2$

–why? what rule is violated?

–how?

Remove A-> $S_1X_aS_2$

$A \rightarrow S_1X_aS_2$

$A \rightarrow$

Note: In general, chain productions:

$A \rightarrow A_1A_2\ldots A_k$, then decompose as:

$A \rightarrow A_1W_1, W_1 \rightarrow A_2W_2\ldots\ldots W_{k-1} \rightarrow A_{k-1}A_k$

Step 3,4 Example, cont.

$S \rightarrow$ TU | aA | A | bB |AB | ccA | cc | ε

$T \rightarrow$ AB | aA | a| bB |b

$A \rightarrow$ aA | a

$B \rightarrow$ bB | b

$U \rightarrow$ ccA | cc \ bB \ b

$A \rightarrow$ aA | a

Finish!

Put in to CNF

$S \rightarrow ASA \mid aB$

$A \rightarrow B \mid S$

$B \rightarrow b \mid \varepsilon$

Should get a grammar equivalent to:

$S_0 \rightarrow AA_1 \mid X_aB \mid a \mid SA \mid AS$

$S \rightarrow AA_1 \mid X_aB \mid a \mid SA \mid AS$

$A \rightarrow b \mid a\ A_1 \mid X_aB \mid a \mid SA \mid AS$

$A1 \rightarrow SA$

$X_a \rightarrow a$

$B \rightarrow b$