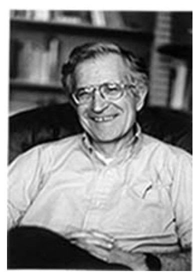
Chomsky Normal Form

Chomsky Normal Form

- Chomsky Normal Form
  - A context free grammar is in Chomsky Normal Form (CNF) if every production is of the form:
    - $A \rightarrow BC$
    - $A \rightarrow a$
    - Where A,B, and C are variables and a is a terminal.

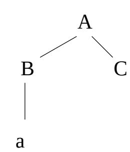
Theory Hall of Fame

- Noam Chomsky
  - The Grammar Guy
  - 1928 –
  - b. Philadelphia, PA
  - PhD – UPenn (1955)
    - Linguistics
  - Prof at MIT (Linguistics) (1955 - present)
  - Probably more famous for his leftist political views.



Chomsky Normal Form

- If we can put a CFG into CNF, then we can calculate the "depth" of the longest branch of a parse tree for the derivation of a string.



At most 2 branches at every node

Chomsky Normal Form

- 3 Step process:

1. Remove ε - Productions
2. Remove Unit Productions
3. Remove Useless Symbols

Removing ε -Productions

- A ε -Productions is a production of the form
  - $A \rightarrow \varepsilon$
  - Basic idea
    - Very similar to removing ε transitions from a NFA-ε
    - Find the set of all variables A such that $A \Rightarrow^* \varepsilon$ (set of nullable variables)
    - For all productions that contain a nullable variable on the right hand side, add a production that eliminates the nullable from the right hand side

Removing ε -Productions

- We must be a bit careful here
  - If ε is in a CFL, then the production $S \rightarrow \varepsilon$ must be in the production set.
  - The algorithm to be described will generate L – {ε}

Removing ε -Productions

- Step 1: Find the set of nullable variables:
  - Example:
    - $S \rightarrow AB$
    - $A \rightarrow aAA \mid \varepsilon$
    - $B \rightarrow bBB \mid \varepsilon$

    - All variables are nullable
      - A and B are nullable since $A \rightarrow \varepsilon$ and $B \rightarrow \varepsilon$
      - S is nullable since $S \rightarrow AB$ and A and B are nullable

Removing ε -Productions

- Step 2: Remove nullable variables
  - For all productions $A \rightarrow \beta$ where β contains nullable variables, add a new production with each nullable removed from β

Removing ε -Productions

Step 2: Remove nullable variables Example:

- $S \rightarrow AB$
- $A \rightarrow aAA \mid \varepsilon$
- $B \rightarrow bBB \mid \varepsilon$

- All variables are nullable

Removing ε -Productions

- Step 2: Remove nullable variables Example:
  - Consider: $S \rightarrow AB$
    - Add to P: $S \rightarrow A$ and $S \rightarrow B$
  - Consider: $A \rightarrow aAA$
    - Add to P: $A \rightarrow aA$ and $A \rightarrow a$
  - Consider: $B \rightarrow bBB$
    - Add to P: $B \rightarrow bB$ and $B \rightarrow b$

Removing ε -Productions

- Step 2: Remove nullable variables
  - Our grammar now looks like:
    - $S \rightarrow AB \mid A \mid B$
    - $A \rightarrow aAA \mid aA \mid a \mid \varepsilon$
    - $B \rightarrow bBB \mid bB \mid b \mid \varepsilon$

Removing ε -Productions

- Step 3: Remove your ε -Productions
  - Example:
    - Remove $A \rightarrow \varepsilon$ and $B \rightarrow \varepsilon$
    - Our final grammar looks like:
      - $S \rightarrow AB \mid A \mid B$
      - $A \rightarrow aAA \mid aA \mid a$
      - $B \rightarrow bBB \mid bB \mid b$
  - Questions?

Removing Unit Productions

- A Unit Productions is a production of the form
    - $A \rightarrow B$ where A and B are variable
  - Basic idea
    - Very similar to removing ε productions
    - For each variable A, find the set of all variables B such that $A \Rightarrow^* B$ by just following unit productions (A-derivable)
    - For all variables B that are A derivable and for all productions $B \rightarrow \alpha$, add the production $A \rightarrow \alpha$

Removing Unit Productions

- Step 0: Remove ε -Productions using the previous algorithm.

Removing Unit Productions

- Step 1: For all variables A find the set of A-derivable variables:
  - Recursive definition of A-derivable
    1. If $A \rightarrow B$ then B is A-derivable
    2. If C is A derivable and $C \rightarrow B$ (and $B \neq A$), then B is A derivable
    3. No other variables are A-derivable.

Removing Unit Productions

- Step 1: For all variables A find the set of A-derivable variables:
  - Example:
    - $S \rightarrow S + T \mid T$
    - $T \rightarrow T * F \mid F$
    - $F \rightarrow (S) \mid a$
    - Let's find the set of S-derivable variables:
      - T is S derivable since $S \rightarrow T$
      - F is S derivable since $T \rightarrow F$ and T is S derivable

Removing Unit Productions

- Step 1: For all variables A find the set of A-derivable variables:
  - Example:
    - $S \rightarrow S + T \mid T$
    - $T \rightarrow T * F \mid F$
    - $F \rightarrow (S) \mid a$
    - S-derivable = {T, F}
    - T-derivable = {F}
    - F-derivable = ∅

Removing Unit Productions

- Step 2: For each variable A, if B is A-derivable, for each non-unit production $B \rightarrow \beta$, add the production $A \rightarrow \beta$

Removing Unit Productions

- Step 2:
  - Example:
    - $S \rightarrow S + T \mid T$
    - $T \rightarrow T * F \mid F$
    - $F \rightarrow (S) \mid a$
    - S-derivable = {T, F}
    - T-derivable = {F}
    - Add to P: $S \rightarrow T * F$, $S \rightarrow (S) \mid a$
    - : $T \rightarrow (S) \mid a$

Removing Unit Productions

- Step 2:
  - Our new grammar now looks like:
    - $S \rightarrow S + T \mid T * F \mid (S) \mid a \mid T$
    - $T \rightarrow T * F \mid (S) \mid a \mid F$
    - $F \rightarrow (S) \mid a$

Removing Unit Productions

- Step 3: Remove Unit Productions
  - Our final grammar looks like:
  - Our new grammar now looks like:
    - $S \rightarrow S + T \mid T * F \mid (S) \mid a$
    - $T \rightarrow T * F \mid (S) \mid a$
    - Remove $S \rightarrow T$, $T \rightarrow F$
  - Questions

Removing Useless Symbols

- A symbol X is <u>useful</u> for a grammar G = (V, T, P, S) if
  - $S \Rightarrow^* \alpha X \beta \Rightarrow^* w$ where $w \in L(G)$
- In other words, a <u>useful symbol</u> will be used somewhere in the derivation of a string in the language.
- Any symbol that is not useful is <u>useless</u>.
- Useless symbols do not add to the language generated by a grammar, so it's okay to remove them.

Removing Useless Symbols

- Definitions:
  - We say a symbol X is <u>generating</u> if:
    - $X \Rightarrow^* w$ for some $w \in L(G)$
  - We say a symbol X is <u>reachable</u> if:
    - $S \Rightarrow^* \alpha X \beta$ for some α, β
- Symbols that are useful must be both generating and reachable.
  - Such symbols (and assoc. productions) can be removed

Removing useless symbols

- Algorithm:
  1. Eliminate all non generating symbols
  2. Eliminate all non reachable symbols from resultant grammar.

Removing useless symbols

- Finding generating symbols
  1. All symbols in T are generating
  2. If $A \rightarrow \alpha$ and all symbols in α are generating, then A is generating.
  3. No other symbols are generating.

Removing useless symbols

- Finding reachable symbols
  1. S is reachable
  2. If A is reachable, and $A \rightarrow \alpha$, then all variables in α are reachable.

Removing Useless Symbols

- Example:

  $S \rightarrow AB \mid a$

  $A \rightarrow b$

  B is useless since it is not generating

  Eliminate it

Removing useless symbols

- Example:

  $S \rightarrow a$

  $A \rightarrow b$

  – Now A is not reachable, eliminate it!

  $S \rightarrow a$

  Note that you must eliminate non-generating symbols before non-reachable symbols.

Recall our goal

- Chomsky Normal Form
  - A context free grammar is in Chomsky Normal Form (CNF) if every production is of the form:
    - $A \rightarrow BC$
    - $A \rightarrow a$
    - Where A,B, and C are variables and a is a terminal.

Chomsky Normal Form

- Given a CFG G, there is an equivalent CFG, G' in Chomsky Normal form such that
  - L(G') = L(G) – {ε}

Chomsky Normal Form

- Step 1:
  - Remove ε -Productions
- Step 2:
  - Remove Unit Productions
- Step 3:
  - Remove useless symbols

Chomsky Normal Form

- After steps 1 – 3 :
  - All productions are of the form:
    - $A \to a$ where A is a variable and a is a terminal
    - $A \to \beta$ where $|\beta| \geq 2$ and β contains variables and/or terminals.
  - Step 4: Derive terminals from new variables:
    - For all productions of the 2nd type: $A \to \beta$, for all terminals a in β, create a new variable X_a
    - Add a new production $X_a \to a$
    - Replace a in β with X_a

Chomsky Normal Form

- Step 4:
  - Let's go back to our first example:
    - $S \to AB \mid A \mid B$
    - $A \to aAA \mid aA \mid a$
    - $B \to bBB \mid bB \mid b$
  - Removing unit transitions:
    - $S \to AB \mid aAA \mid aA \mid a \mid bBB \mid bB \mid b$
    - $A \to aAA \mid aA \mid a$
    - $B \to bBB \mid bB \mid b$
  - Note that S, A, and B are all useful.

Chomsky Normal Form

- Step 4:
  - Define new productions: $X_a \to a$ and $X_b \to b$ and replace instance of a with X_a , similarly for b
    - $S \to AB \mid aAA \mid aA \mid a \mid bBB \mid bB \mid b$
    - $A \to aAA \mid aA \mid a$
    - $B \to bBB \mid bB \mid b$
  - New:
    - $S \to AB \mid X_a AA \mid X_a A \mid a \mid X_b BB \mid X_b B \mid b$
    - $A \to X_a AA \mid X_a A \mid a$
    - $B \to X_b BB \mid X_b B \mid b$
    - $X_a \to a$
    - $X_b \to b$

Chomsky Normal Form

- After steps 1 – 4 :
  - All productions are of the form:
    - $A \to a$ where A is a variable and a is a terminal
    - $A \to \beta$ where $|\beta| \geq 2$ and β contains only variables.
  - Step 5:
    - For all productions of type 2 where $|\beta| > 2$, replace the production with a series of new productions each having exactly 2 variables on the right
    - Best illustrated with an example

Chomsky Normal Form

- Step 4:
  - The production:
    - $A \rightarrow BCDBCE$
  - Would be replaced with
    - $A \rightarrow BY_1$
    - $Y_1 \rightarrow CY_2$
    - $Y_2 \rightarrow DY_3$
    - $Y_3 \rightarrow BY_4$
    - $Y_4 \rightarrow CE$

Chomsky Normal Form

- Step 4:
  - Back to our example
    - $S \rightarrow AB \mid \underline{X_a AA} \mid X_a A \mid a \mid \underline{X_b BB} \mid X_b B \mid b$
    - $A \rightarrow \underline{X_a AA} \mid X_a A \mid a$
    - $B \rightarrow \underline{X_b BB} \mid X_b B \mid b$
    - $X_a \rightarrow a$
    - $X_b \rightarrow b$
  - Add productions
    - $Y_1 \rightarrow AA$
    - $Y_2 \rightarrow BB$

Chomsky Normal Form

- Step 4:
  - Our final grammar
    - $S \rightarrow AB \mid X_a Y_1 \mid X_a A \mid a \mid X_b Y_2 \mid X_b B \mid b$
    - $A \rightarrow X_a Y_1 \mid X_a A \mid a$
    - $B \rightarrow X_b Y_2 \mid X_b B \mid b$
    - $Y_1 \rightarrow AA$
    - $Y_2 \rightarrow BB$
    - $X_a \rightarrow a$
    - $X_b \rightarrow b$
  - Questions

CNF

- Any grammar can be placed into CNF
- Why bother?
  - Remember that awful CFG we generated last week?
    - Simplification
  - Gives upper limit on size of parse tree
    - Pumping Lemma will need this.

Questions?

- Next time
  - The Return of the pumping lemma