1 Chomsky Normal Form

Normal Forms for Grammars

It is typically easier to work with a context free language if given a CFG in a *normal form*.

Normal Forms
A grammar is in a normal form if its production rules have a special structure:

- *Chomsky Normal Form:* Productions are of the form $A \to BC$ or $A \to a$, where A, B, C are variables and a is a terminal symbol.
- *Greibach Normal Form* Productions are of the form $A \to a\alpha$, where $\alpha \in V^*$ and $A \in V$.

If ϵ is in the language, we allow the rule $S \to \epsilon$. We will require that S does not appear on the right hand side of any rules.

We will restrict our discussion to Chomsky Normal Form.

Main Result

Proposition 1. *For any non-empty context-free language L, there is a grammar G, such that $L(G) = L$ and each rule in G is of the form*

1. *$A \to a$ where $a \in \Sigma$, or*
2. *$A \to BC$ where neither B nor C is the start symbol, or*
3. *$S \to \epsilon$ where S is the start symbol (iff $\epsilon \in L$)*

Furthermore, G has no useless symbols.

Outline of Normalization

Given $G = (V, \Sigma, S, P)$, convert to CNF

- Let $G' = (V', \Sigma, S, P')$ be the grammar obtained after eliminating ϵ-productions, unit productions, and useless symbols from G.
- If $A \to x$ is a rule of G', where $|x| = 0$, then A must be S (because G' has no other ϵ-productions). If $A \to x$ is a rule of G', where $|x| = 1$, then $x \in \Sigma$ (because G' has no unit productions). In either case $A \to x$ is in a valid form.
- All remaining productions are of form $A \to X_1 X_2 \cdots X_n$ where $X_i \in V' \cup \Sigma$, $n \geq 2$ (and S does not occur in the RHS). We will put these rules in the right form by applying the following two transformations:
  1. Make the RHS consist only of variables
  2. Make the RHS be of length 2.

Make the RHS consist only of variables

Let $A \to X_1 X_2 \cdots X_n$, with X_i being either a variable or a terminal. We want rules where all the X_i are variables.

Example 2. Consider $A \to BbCdefG$. How do you remove the terminals?

For each $a, b, c \ldots \in \Sigma$ add variables $X_a, X_b, X_c, \ldots$ with productions $X_a \to a$, $X_b \to b$, ... Then replace the production $A \to BbCdefG$ by $A \to BX_bCX_dX_eX_fG$

For every $a \in \Sigma$

1. Add a new variable X_a
2. In every rule, if a occurs in the RHS, replace it by X_a
3. Add a new rule $X_a \to a$

Make the RHS be of length 2

- Now all productions are of the form $A \to a$ or $A \to B_1 B_2 \cdots B_n$, where $n \geq 2$ and each B_i is a variable.
- How do you eliminate rules of the form $A \to B_1 B_2 \ldots B_n$ where $n > 2$?
- Replace the rule by the following set of rules

$$\begin{array}{rcl} A & \to & B_1 B_{(2,n)} \\ B_{(2,n)} & \to & B_2 B_{(3,n)} \\ B_{(3,n)} & \to & B_3 B_{(4,n)} \\ & \vdots & \\ B_{(n-1,n)} & \to & B_{n-1} B_n \end{array}$$

  where $B_{(i,n)}$ are "new" variables.

An Example

Example 3. Convert: $S \to aA|bB|b$, $A \to Baa|ba$, $B \to bAAb|ab$, into Chomsky Normal Form.

1. Eliminate ϵ-productions, unit productions, and useless symbols. This grammar is already in the right form.
2. Remove terminals from the RHS of long rules. New grammar is: $X_a \to a$, $X_b \to b$, $S \to X_aA|X_bB|b$, $A \to BX_aX_a|X_bX_a$, and $B \to X_bAAX_b|X_aX_b$
3. Reduce the RHS of rules to be of length at most two. New grammar replaces $A \to BX_aX_a$ by rules $A \to BX_{aa}$, $X_{aa} \to X_aX_a$, and $B \to X_bAAX_b$ by rules $B \to X_bX_{AAb}$, $X_{AAb} \to AX_{Ab}$, $X_{Ab} \to AX_b$