CSE 322

Introduction to Formal Models in Computer Science
Alternative Conversion to Chomsky Normal Form

Chomsky Normal Form: A context-free grammar $G = (V, \Sigma, R, S)$ is in Chomsky normal if and only if S does not appear on the right hand side of any rule, and all rules are of the form:

- $A \to BC$ for some $B, C \in V$,
- $A \to a$ for some $a \in \Sigma$, or
- $S \to \epsilon$

Theorem: Every CFL can be generated by some grammar in Chomsky Normal Form.

Proof: Let $G = (V, \Sigma, R, S)$ be a context-free grammar generating L. We give a several step construction for converting G to a grammar $G' = (V', \Sigma, R', S')$ in Chomsky Normal Form.

Step 1: Create a new start symbol S' and add the rule $S' \to S$.

Step 2: For each rule $A \to B_1 \ldots B_k$ with $k > 2$ (where each B_i is either a variable or a terminal), create new variables $T_2, \ldots T_{k-1}$ and replace the rule by rules $A \to B_1T_2$, $T_2 \to B_2T_3$,$\ldots T_{k-1} \to B_{k-1}B_k$. There are separate symbols T_i for each rule converted in this way.

Now all rules have right-hand sides of length at most 2.

Step 3: Figure out the set of variables $\mathcal{E}$ that can generate the empty string ϵ: If $A \to \epsilon$ is a rule then put A in $\mathcal{E}$. Then for every $A \in \mathcal{E}$ if $B \to w$ is a rule with $w \in \mathcal{E}^*$, also put $B \in \mathcal{E}$.

If $S' \in \mathcal{E}$ add the rule $S' \to \epsilon$ and remove all other rules $A \to \epsilon$. For every rule $A \to BC$ with $B \in \mathcal{E}$ add the rule $A \to C$. (C can either be a variable or terminal.) For every rule $A \to BC$ with $C \in \mathcal{E}$ add the rule $A \to B$. (B can either be a variable or terminal.)

Now all rules with right-hand side ϵ have been removed (except for the possible exception of $S' \to \epsilon$, which is allowed in Chomsky Normal Form).

Step 4: A *unit rule* is a rule of the form $A \to B$ where A and B are variables. We now only need to eliminate all unit rules. To do this we draw a directed graph of all the variables where there is an edge from A to B if $A \to B$ is a rule. For any variable A, let $V(A)$ be the set of variables reachable from A in this graph (by definition A always belongs to $V(A)$).

Call a right-hand side of a rule *interesting* if the rule is not a unit rule. To make the Chomsky normal form grammar, we define a new grammar with the same variables in which $A \to w$ if and only if w is an interesting right-hand side of some rule whose left-hand side is in $V(A)$.

This new grammar has all right-hand sides of length at most 2, no rules of form $A \to \epsilon$ for $A \neq S'$, and no unit rules.

Step 5: We are thus essentially done, except that we might have right-hand sides of length 2 which have a terminal as one of the two symbols, i.e., rules of the form $A \to aB$ or $A \to Ba$ or $A \to ab$. This is easily fixed by adding an entirely new variable T_a for each terminal a, adding the rule $T_a \to a$, and replacing all occurrences of a in right-hand sides of length 2 by T_a.

Clearly these rules keep the language generated the same and the final grammar produced is in Chomsky Normal Form. $\square$