CSE 322

Introduction to Formal Models in Computer Science

Conversion to Chomsky Normal Form

The following form is slightly more general than the version of Chomsky Normal Form given in the text. This form is just as good for the purposes of parsing and has the property that *every* CFL can be expressed in this 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' in Chomsky Normal Form.

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

Step 2: For each terminal symbol $a \in \Sigma$ that appears on the right side of a rule of G of size at least 2 create a new variable A, add the rule $A \to a$ and replace very occurence of a on the right side of a rule of size at least 2 by A.

Step 3: For each rule $A \to B_1 \ldots B_k$ with $k > 2$, create new non-terminals $T_2, \ldots T_{k-1}$ replace the rule by rules $A \to B_1 T_2$, $T_2 \to B_2 T_3$,$\ldots T_{k-1} \to B_{k-1} B_k$. (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 4: Figure out the set of nonterminals $\mathcal{E}$ 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 $A \to C$. For every rule $A \to BC$ with $C \in \mathcal{E}$ add the rule $A \to B$.

Step 5: A *unit rule* is a rule of the form $A \to B$ where A and B are nonterminals. We now only need to eliminate all unit rules. To do this we draw a directed graph of all the nonterminals where there is an edge from A to B if $A \to B$ is a rule. For any non-terminal A, let $\mathcal{D}(A)$ be the set of nonterminals reachable from A in this graph. (This is just like the $\mathcal{D}(A)$ in the text except we ignore terminals.)

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 non-terminals 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 $\mathcal{D}(A)$.

Clearly these rules keep the language generated the same. $\square$