CSE 322 - Introduction to Formal Methods in Computer Science
Chomsky Normal Form

Dave Bacon
Department of Computer Science & Engineering, University of Washington

A useful form for dealing with context free grammars is the Chomksy normal form. This is a particular form of writing a CFG which is useful for understanding CFGs and for proving things about them. It also makes the parse tree for derivations using this form of the CFG a binary tree. And as a CS major, I know you really love binary trees!

So what is Chomsky normal form? A CFG is in Chomsky normal form when every rule is of the form $A \to BC$ and $A \to a$, where a is a terminal, and A, B, and C are variables. Further B and C are not the start variable. Additionally we permit the rule $S \to \varepsilon$ where S is the start variable, for technical reasons. Note that this means that we allow $S \to \varepsilon$ as one of many possible rules.

Okay, so if this is the Chomsky normal form what is it good for? Well as a first fact, note that parse trees for a derivation using such a grammar will be a binary tree. Thats nice. It will help us down the road. Okay, so if it might be good for something, we can ask the natural question: is it possible to convert an arbitrary CFG into an equivalent grammar which is of the Chomsky normal form? The answer, it turns out, is yes. Lets see how such a conversion would proceed.

A. A new start variable

The first step is simple! We just add a new start variable S_0 and the rule $S_0 \to S$ where S is the original start variable. By doing this we guarantee that the start variable doesn't occur on the right hand side of a rule.

B. Eliminate the ε rules

Next we remove the ε rule. We do this as follows. Suppose we are removing the ε rule $A \to \varepsilon$. We remove this rule. But now we have to "fix" the rules which have an A on their right-hand side. We do this by, for each occurrence of A on the right hand side, adding a rule (from the same starting variable) which has the A removed. Further if A is the only thing occurring on the right hand side, we replace this A with ε. Of course this latter fact will have created a new ε rule. So we do this *unless we have previously removed* $A \to \varepsilon$. But onward we press: simply repeat the above process over and over again until all ε rules have been removed.

For example, suppose our rules contain the rule $A \to \varepsilon$ and the rule $B \to uAv$ where u and v are not both the empty string. First we remove $A \to \varepsilon$. Then we add to this rule the rule $B \to uv$. (Make sure that you don't delete the original rule $B \to uAv$. If, on the other hand we had the rule $A \to \varepsilon$ and $B \to A$, then we would remove the $A \to \varepsilon$ and replace the rule $B \to A$ with the rule $B \to \varepsilon$. Of course we now have to eliminate this rule via the same procedure.

C. Remove the unit rules

Next we need to remove the unit rules. If we have the rule $A \to B$, then whenever the rule $B \to u$ appears, we will add the rule $A \to u$ (unless this rule was already replaced.) Again we do this repeatedly until we eliminate all unit rules.

D. Take care of rules with more than two terminals or variables

At this point we have converted our CFG to one which has no ε transitions, and where all rules are either of the form variables goes to terminal, or of the form variable goes to string of variables and terminals with two or more symbols. These later rules are of the appropriate Chomsky normal form. To convert the remaining rules to proper form, we introduce extra variables. In particular suppose $A \to u_1 u_2 \dots u_n$ where $n > 2$. Then we convert this to a set of rules, $A \to u_1 A_1$, $A_1 \to u_2 A_2$, $\dots$, $A_{k-2} \to u_{k-1} u_k$. Now we need to take care of the rules with two elements on the right hand side. If both of the elements are variables, then we are fine. But if any of them are terminals, we

add a new variable and a new rule to take care of these. For example, if we have $A \to u_1 B$ where u_1 is a terminal, then we replace this by $A \to U_1 B$ and $U \to u_1$.

I. EXAMPLE CONVERSION TO CHOMSKY NORMAL FORM

Lets work out an example. Consider the grammar

$$\begin{aligned} S &\to ASB \\ A &\to aAS|a|\varepsilon \\ B &\to SbS|A|bb \end{aligned}$$

First we add a new start state:

$$\begin{aligned} S_0 &\to S \\ S &\to ASB \\ A &\to aAS|a|\varepsilon \\ B &\to SbS|A|bb \end{aligned}$$

Next we need to eliminate the ε rules. Eliminating $A \to \varepsilon$ yields

$$\begin{aligned} S_0 &\to S \\ S &\to ASB|SB \\ A &\to aAS|a|aS \\ B &\to SbS|A|bb|\varepsilon \end{aligned}$$

Now we have a new ε rule., $B \to \varepsilon$. Lets remove it

$$\begin{aligned} S_0 &\to S \\ S &\to ASB|SB|S|AS \\ A &\to aAS|a|aS \\ B &\to SbS|A|bb \end{aligned}$$

Next we need to remove all unit rules. Lets begin by removing $B \to A$:

$$\begin{aligned} S_0 &\to S \\ S &\to ASB|SB|S|AS \\ A &\to aAS|a|aS \\ B &\to SbS|bb|aAS|a|aS \end{aligned}$$

Next lets remove $S \to S$:

$$\begin{aligned} S_0 &\to S \\ S &\to ASB|SB|AS \\ A &\to aAS|a|aS \\ B &\to SbS|bb|aAs|a|aS \end{aligned}$$

Further we can eliminate $S_0 \to S$:

$$\begin{aligned} S_0 &\to ASB|SB|AS \\ S &\to ASB|SB|AS \\ A &\to aAS|a|aS \\ B &\to SbS|bb|aAs|a|aS \end{aligned}$$

Now we need to take care of the rules with more than three symbols. First replace $S_0 \to ASB$ by $S_0 \to AU_1$ and $U_1 \to SB$:

$$\begin{aligned} S_0 &\to AU_1|SB|AS \\ S &\to ASB|SB|AS \\ A &\to aAS|a|aS \\ B &\to SbS|bb|aAs|a|aS \\ U_1 &\to SB \end{aligned}$$

Next eliminate $S \to ASB$ in a similar form (technically we could reuse U_1, but lets not):

$$\begin{aligned} S_0 &\to AU_1|SB|AS \\ S &\to AU_2|SB|AS \\ A &\to aAS|a|aS \\ B &\to SbS|bb|aAs|a|aS \\ U_1 &\to SB \\ U_2 &\to SB \end{aligned}$$

Onward and upward, now fix $A \to aAS$ by introducing $A \to aU_3$ and $U_3 \to AS$.

$$\begin{aligned} S_0 &\to AU_1|SB|AS \\ S &\to AU_2|SB|AS \\ A &\to aU_3|a|aS \\ B &\to SbS|bb|aAs|a|aS \\ U_1 &\to SB \\ U_2 &\to SB \\ U_3 &\to AS \end{aligned}$$

Finally, fix the two $B \to$ rules:

$$\begin{aligned} S_0 &\to AU_1|SB|AS \\ S &\to AU_2|SB|AS \\ A &\to aU_3|a|aS \\ B &\to SU_4|bb|aU_5|a|aS \\ U_1 &\to SB \\ U_2 &\to SB \\ U_3 &\to AS \\ U_4 &\to bS \\ U_5 &\to AS \end{aligned}$$

Finally we need to work with the rules which have terminals and variables or two terminals. We need to introduce new variables for these. Let these be $V_1 \to a$ and $V_2 \to b$:

$$\begin{aligned} S_0 &\to AU_1|SB|AS \\ S &\to AU_2|SB|AS \\ A &\to V_1U_3|a|V_1S \\ B &\to SU_4|V_2V_2|V_1U_5|a|V_1S \\ U_1 &\to SB \\ U_2 &\to SB \\ U_3 &\to AS \\ U_4 &\to V_2S \\ U_5 &\to AS \\ V_1 &\to a \\ V_2 &\to b \end{aligned}$$

A quick examination shows us that we have ended up with a grammar in Chomsky normal form. (This can, of course, be simplified.)