

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 Chomsky 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 $A \rightarrow BC$ and $A \rightarrow a$, 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 $S \rightarrow \varepsilon$ where S is the start variable, for technical reasons. Note that this means that we allow $S \rightarrow \varepsilon$ 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. That's 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. Let's 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 $S_0 \rightarrow S$ 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 $A \rightarrow \varepsilon$. 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. 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 $A \rightarrow \varepsilon$ and the rule $B \rightarrow uAv$ where u and v are not both the empty string. First we remove $A \rightarrow \varepsilon$. Then we add to this rule the rule $B \rightarrow uv$. (Make sure that you don't delete the original rule $B \rightarrow uAv$. If, on the other hand we had the rule $A \rightarrow \varepsilon$ and $B \rightarrow A$, then we would remove the $A \rightarrow \varepsilon$ and replace the rule $B \rightarrow A$ with the rule $B \rightarrow \varepsilon$. 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 $A \rightarrow B$, then whenever the rule $B \rightarrow v$ appears, we will replace it by $A \rightarrow v$ (unless this rule was already removed in this step.) 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 variable 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 $A \rightarrow u_1u_2\dots u_n$ where $n > 2$. Then we convert this to a set of rules, $A \rightarrow u_1A_1$, $A_1 \rightarrow u_2A_2$, \dots , $A_{k-2} \rightarrow u_{k-1}u_k$. 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 $A \rightarrow u_1B$ where u_1 is a terminal, then we replace this by $A \rightarrow U_1B$ and $U \rightarrow u_1$.

I. EXAMPLE CONVERSION TO CHOMSKY NORMAL FORM

Lets work out an example. Consider the grammar

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

First we add a new start state:

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

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

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

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

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

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

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

Next lets remove $S \rightarrow S$:

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

Further we can eliminate $S_0 \rightarrow S$:

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

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

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

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

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

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

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

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

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

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 $V_1 \rightarrow a$ and $V_2 \rightarrow b$:

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

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