Chomsky Normal Form

We introduce Chomsky Normal Form, which is used to answer questions about context-free languages

Chomsky Normal Form

Chomsky Normal Form. *A grammar where every production is either of the form* $A \to BC$ *or* $A \to \mathtt{c}$ *(where* A, B, C *are arbitrary variables and* $\mathtt{c}$ *an arbitrary symbol).*

Example:

$$S \to AS \mid \mathtt{a}$$
$$A \to SA \mid \mathtt{b}$$

(If language contains ε, then we allow $S \to \varepsilon$ where S is start symbol, and forbid S on RHS.)

Why Chomsky Normal Form?

The key advantage is that in Chomsky Normal Form, every derivation of a string of n letters has exactly $2n - 1$ steps.

Thus: *one can determine if a string is in the language by exhaustive search of all derivations.*

Conversion

The conversion to Chomsky Normal Form has four main steps:

1. Get rid of all ε productions.
2. Get rid of all productions where RHS is one variable.
3. Replace every production that is too long by shorter productions.
4. Move all terminals to productions where RHS is one terminal.

1) Eliminate ε Productions

Determine the nullable variables (those that generate ε) (algorithm given earlier).

Go through all productions, and for each, omit every possible subset of nullable variables.

For example, if $P \to A\mathtt{x}B$ with both A and B nullable, add productions $P \to \mathtt{x}B \mid A\mathtt{x} \mid \mathtt{x}$.

After this, delete all productions with empty RHS.

2) Eliminate Variable Unit Productions

A unit production is where RHS has only one symbol.

Consider production $A \to B$. Then for every production $B \to \alpha$, add the production $A \to \alpha$. Repeat until done (but don't re-create a unit production already deleted).

3) Replace Long Productions by Shorter Ones

For example, if have production $A \to BCD$, then replace it with $A \to BE$ and $E \to CD$.

(In theory this introduces many new variables, but one can re-use variables if careful.)

4) Move Terminals to Unit Productions

For every terminal on the right of a non-unit production, add a substitute variable.

For example, replace production $A \to \mathtt{b}C$ with productions $A \to BC$ and $B \to \mathtt{b}$.

Example

Consider the CFG:

$S \to \mathtt{a}X\mathtt{b}X$

$X \to \mathtt{a}Y \mid \mathtt{b}Y \mid \varepsilon$

$Y \to X \mid \mathtt{c}$

The variable X is nullable; and so therefore is Y. After elimination of ε, we obtain:

$S \to \mathtt{a}X\mathtt{b}X \mid \mathtt{ab}X \mid \mathtt{a}X\mathtt{b} \mid \mathtt{ab}$

$X \to \mathtt{a}Y \mid \mathtt{b}Y \mid \mathtt{a} \mid \mathtt{b}$

$Y \to X \mid \mathtt{c}$

Example: Step 2

After elimination of the unit production $Y \to X$, we obtain:

$S \to \mathtt{a}X\mathtt{b}X \mid \mathtt{ab}X \mid \mathtt{a}X\mathtt{b} \mid \mathtt{ab}$
$X \to \mathtt{a}Y \mid \mathtt{b}Y \mid \mathtt{a} \mid \mathtt{b}$
$Y \to \mathtt{a}Y \mid \mathtt{b}Y \mid \mathtt{a} \mid \mathtt{b} \mid \mathtt{c}$

Example: Steps 3 & 4

Now, break up the RHSs of S; and replace $\mathtt{a}$ by A, $\mathtt{b}$ by B and $\mathtt{c}$ by C wherever not units:

$$S \rightarrow EF \mid AF \mid EB \mid AB$$
$$X \rightarrow AY \mid BY \mid \mathtt{a} \mid \mathtt{b}$$
$$Y \rightarrow AY \mid BY \mid \mathtt{a} \mid \mathtt{b} \mid \mathtt{c}$$
$$E \rightarrow AX$$
$$F \rightarrow BX$$
$$A \rightarrow \mathtt{a}$$
$$B \rightarrow \mathtt{b}$$
$$C \rightarrow \mathtt{c}$$

Practice

Convert the following CFG into Chomsky Normal Form:

$$S \to A\mathtt{b}A$$
$$A \to A\mathtt{a} \mid \varepsilon$$

Solution to Practice

After the first step, one has:

$$S \to A\texttt{b}A \mid \texttt{b}A \mid A\texttt{b} \mid \texttt{b}$$
$$A \to A\texttt{a} \mid \texttt{a}$$

The second step does not apply. After the third step, one has:

$$S \to TA \mid \texttt{b}A \mid A\texttt{b} \mid \texttt{b}$$
$$A \to A\texttt{a} \mid \texttt{a}$$
$$T \to A\texttt{b}$$

Solution Continued

And finally, one has:

$$
\begin{aligned}
S &\to TA \mid BA \mid AB \mid \texttt{b} \\
A &\to AC \mid \texttt{a} \\
T &\to AB \\
B &\to \texttt{b} \\
C &\to \texttt{a}
\end{aligned}
$$

Summary

There are special forms for CFGs such as Chomsky Normal Form, where every production has the form $A \to BC$ or $A \to \mathtt{c}$. The algorithm to convert to this form involves (1) determining all nullable variables and getting rid of all ε-productions, (2) getting rid of all variable unit productions, (3) breaking up long productions, and (4) moving terminals to unit productions.