Lecture 21

Normal Forms

For many applications, it is often helpful to assume that CFGs are in one or another special restricted form. Two of the most useful such forms are *Chomsky normal form* (CNF) and *Greibach normal form* (GNF).

Definition 21.1 A CFG is in *Chomsky normal form* (CNF) if all productions are of the form

$$A \rightarrow BC \text{ or } A \rightarrow a,$$

where $A, B, C \in N$ and $a \in \Sigma$. A CFG is in *Greibach normal form* (GNF) if all productions are of the form

$$A \rightarrow aB_1B_2 \cdots B_k$$

for some $k \geq 0$, where $A, B_1, \ldots, B_k \in N$ and $a \in \Sigma$. Note that $k = 0$ is allowed, giving productions of the form $A \rightarrow a$. □

For example, the two grammars

$$S \rightarrow AB \mid AC \mid SS, \qquad C \rightarrow SB, \qquad A \rightarrow \texttt{[}, \qquad B \rightarrow \texttt{]}, \tag{21.1}$$

$$S \rightarrow \texttt{[}B \mid \texttt{[}SB \mid \texttt{[}BS \mid \texttt{[}SBS, \qquad B \rightarrow \texttt{]} \tag{21.2}$$

are grammars in Chomsky and Greibach normal form, respectively, for the set of all nonnull strings of balanced parentheses [ ].

No grammar in Chomsky or Greibach form can generate the null string ϵ (Why not?). Apart from this one exception, they are completely general:

D. C. Kozen, *Automata and Computability*

Theorem 21.2 *For any CFG G, there is a CFG G' in Chomsky normal form and a CFG G'' in Greibach normal form such that*

$$L(G'') = L(G') = L(G) - \{\epsilon\}.$$

Getting Rid of ϵ- and Unit Productions

To prove Theorem 21.2, we must first show how to get rid of all *ϵ-productions* $A \rightarrow \epsilon$ and *unit productions* $A \rightarrow B$. These productions are bothersome because they make it hard to determine whether applying a production makes any progress toward deriving a string of terminals. For instance, with unit productions, there can be loops in the derivation, and with ϵ-productions, one can generate very long strings of nonterminals and then erase them all. Without ϵ- or unit productions, every step in the derivation makes demonstrable progress toward the terminal string in the sense that either the sentential form gets strictly longer or a new terminal symbol appears.

We cannot simply throw out the ϵ- and unit productions, because they may be needed to generate some strings in $L(G)$; so before we throw them out, we had better throw in some other productions we can use instead.

Lemma 21.3 *For any CFG $G = (N, \Sigma, P, S)$, there is a CFG G' with no ϵ- or unit productions such that $L(G') = L(G) - \{\epsilon\}$.*

Proof. Let $\widehat{P}$ be the smallest set of productions containing P and closed under the two rules

(a) if $A \rightarrow \alpha B \beta$ and $B \rightarrow \epsilon$ are in $\widehat{P}$, then $A \rightarrow \alpha\beta$ is in $\widehat{P}$; and

(b) if $A \rightarrow B$ and $B \rightarrow \gamma$ are in $\widehat{P}$, then $A \rightarrow \gamma$ is in $\widehat{P}$.

We can construct $\widehat{P}$ inductively from P by adding productions as required to satisfy (a) and (b). Note that only finitely many productions ever get added, since each new right-hand side is a substring of an old right-hand side. Thus $\widehat{P}$ is still finite.

Now let $\widehat{G}$ be the grammar

$$\widehat{G} = (N, \Sigma, \widehat{P}, S).$$

Since $P \subseteq \widehat{P}$, every derivation of G is a derivation of $\widehat{G}$; thus $L(G) \subseteq L(\widehat{G})$. But $L(G) = L(\widehat{G})$, since each new production that was thrown in because of rule (a) or (b) can be simulated in two steps by the two productions that caused it to be thrown in.

Now we show that for nonnull $x \in \Sigma^*$, any derivation $S \xrightarrow[\widehat{G}]{*} x$ of minimum length does not use any ϵ- or unit productions. Thus the ϵ- and unit productions are superfluous and can be deleted from $\widehat{G}$ with impunity.

Let $x \neq \epsilon$ and consider a minimum-length derivation $S \xrightarrow[\widehat{G}]{*} x$. Suppose for a contradiction that an ϵ-production $B \to \epsilon$ is used at some point in the derivation, say

$$S \xrightarrow[\widehat{G}]{*} \gamma B \delta \xrightarrow[\widehat{G}]{1} \gamma\delta \xrightarrow[\widehat{G}]{*} x.$$

One of γ, δ is nonnull, otherwise x would be null, contradicting the assumption that it isn't. Thus that occurrence of B must first have appeared earlier in the derivation when a production of the form $A \to \alpha B \beta$ was applied:

$$S \xrightarrow[\widehat{G}]{m} \eta A \theta \xrightarrow[\widehat{G}]{1} \eta\alpha B\beta\theta \xrightarrow[\widehat{G}]{n} \gamma B \delta \xrightarrow[\widehat{G}]{1} \gamma\delta \xrightarrow[\widehat{G}]{k} x$$

for some $m, n, k \geq 0$. But by rule (a), $A \to \alpha\beta$ is also in $\widehat{P}$, and this production could have been applied at that point instead, giving a strictly shorter derivation of x:

$$S \xrightarrow[\widehat{G}]{m} \eta A \theta \xrightarrow[\widehat{G}]{1} \eta\alpha\beta\theta \xrightarrow[\widehat{G}]{n} \gamma\delta \xrightarrow[\widehat{G}]{k} x.$$

This contradicts our assumption that the derivation was of minimum length.

A similar argument shows that unit productions do not appear in minimum-length derivations in $\widehat{G}$. Let $x \neq \epsilon$ and consider a derivation $S \xrightarrow[\widehat{G}]{*} x$ of minimum length. Suppose a unit production $A \to B$ is used at some point, say

$$S \xrightarrow[\widehat{G}]{*} \alpha A \beta \xrightarrow[\widehat{G}]{1} \alpha B \beta \xrightarrow[\widehat{G}]{*} x.$$

We must eventually dispose of that occurrence of B, say by applying a production $B \to \gamma$ later on.

$$S \xrightarrow[\widehat{G}]{m} \alpha A \beta \xrightarrow[\widehat{G}]{1} \alpha B \beta \xrightarrow[\widehat{G}]{n} \eta B \theta \xrightarrow[\widehat{G}]{1} \eta\gamma\theta \xrightarrow[\widehat{G}]{k} x.$$

But by rule (b), $A \to \gamma$ is also in $\widehat{P}$, and this could have been applied instead, giving a strictly shorter derivation of x:

$$S \xrightarrow[\widehat{G}]{m} \alpha A \beta \xrightarrow[\widehat{G}]{1} \alpha\gamma\beta \xrightarrow[\widehat{G}]{n} \eta\gamma\theta \xrightarrow[\widehat{G}]{k} x.$$

Again, this contradicts the minimality of the length of the derivation.

Thus we do not need ϵ-productions or unit productions to generate nonnull strings. If we discard them from $\widehat{G}$, we obtain a grammar G' generating $L(G) - \{\epsilon\}$. □

Chomsky Normal Form

Once we are rid of ϵ- and unit productions, it is a simple matter to put the resulting grammar into Chomsky normal form. For each terminal $a \in \Sigma$, introduce a new nonterminal A_a and production $A_a \rightarrow a$, and replace all occurrences of a on the right-hand sides of old productions (except productions of the form $B \rightarrow a$) with A_a. Then all productions are of the form

$$A \rightarrow a \text{ or } A \rightarrow B_1 B_2 \cdots B_k, \quad k \geq 2,$$

where the B_i are nonterminals. The set of terminal strings generated is not changed; it just takes one more step than before to generate a terminal symbol. For any production

$$A \rightarrow B_1 B_2 \cdots B_k$$

with $k \geq 3$, introduce a new nonterminal C and replace this production with the two productions

$$A \rightarrow B_1 C \quad \text{and} \quad C \rightarrow B_2 B_3 \cdots B_k.$$

Keep doing this until all right-hand sides are of length at most 2.

Example 21.4 Let's derive a CNF grammar for the set

$$\{a^n b^n \mid n \geq 0\} - \{\epsilon\} = \{a^n b^n \mid n \geq 1\}.$$

Starting with the grammar

$$S \rightarrow aSb \mid \epsilon$$

for $\{a^n b^n \mid n \geq 0\}$, we remove the ϵ-production as described in Lemma 21.3 to get

$$S \rightarrow aSb \mid ab,$$

which generates $\{a^n b^n \mid n \geq 1\}$. Then we add nonterminals A, B and replace these productions with

$$S \rightarrow ASB \mid AB, \qquad A \rightarrow a, \qquad B \rightarrow b.$$

Finally, we add a nonterminal C and replace $S \rightarrow ASB$ with

$$S \rightarrow AC \quad \text{and} \quad C \rightarrow SB.$$

The final grammar in Chomsky form is

$$S \rightarrow AB \mid AC, \qquad C \rightarrow SB, \qquad A \rightarrow a, \qquad B \rightarrow b.$$

□

Example 21.5 We derive a CNF grammar for the set of nonnull strings of balanced parentheses [ ]. Start with the grammar

$$S \rightarrow [S] \mid SS \mid \epsilon$$

for all balanced strings of parentheses. Applying the construction of Lemma 21.3 to get rid of the ϵ- and unit productions, we get

$$S \rightarrow [S] \mid SS \mid [\,].$$

Next we add new nonterminals A, B and replace these productions with

$$S \rightarrow ASB \mid SS \mid AB, \qquad A \rightarrow [, \qquad B \rightarrow].$$

Finally, we add a new nonterminal C and replace $S \rightarrow ASB$ with

$$S \rightarrow AC \quad \text{and} \quad C \rightarrow SB.$$

The resulting grammar in Chomsky form is exactly (21.1). □

Greibach Normal Form

Now we show how to convert an arbitrary grammar to an equivalent one (except possibly for ϵ) in Greibach normal form.

We start with a grammar $G = (N, \Sigma, P, S)$ in Chomsky normal form. This assumption is mainly for ease of presentation; we could easily modify the construction to apply more generally. The construction as given here produces a Greibach grammar with at most two nonterminals on the right-hand side (cf. [60, Exercise 4.16, p. 104]).

For α, $\beta \in (N \cup \Sigma)^*$, write

$$\alpha \xrightarrow[G]{L} \beta$$

if β can be derived from α by a sequence of steps in which productions are applied only to the leftmost symbol in the sentential form (which must therefore be a nonterminal). For $A \in N$ and $a \in \Sigma$, define

$$R_{A,a} = \{\beta \in N^* \mid A \xrightarrow[G]{L} a\beta\}.$$

For example, in the CNF grammar (21.1), we would have

$$C \xrightarrow[G]{L} SB \xrightarrow[G]{L} SSB \xrightarrow[G]{L} SSSB \xrightarrow[G]{L} ACSSB \xrightarrow[G]{L} [CSSB,$$

so $CSSB \in R_{C,[}$.

The set $R_{A,a}$ is a regular set over the alphabet N, because the grammar with nonterminals $\{A' \mid A \in N\}$, terminals N, start symbol S', and productions

$$\{A' \rightarrow B'C \mid A \rightarrow BC \in P\} \cup \{A' \rightarrow \epsilon \mid A \rightarrow a \in P\}$$

is a strongly left-linear grammar for it.[1] This may seem slightly bizarre, since the terminals of this grammar are the nonterminals N of G, but a moment's thought will convince you that it makes perfect sense.

Since $R_{A,a}$ is regular, by Homework 5, Exercise 1 it also has a strongly right-linear grammar $G_{A,a}$; that is, one in which all productions are of the form $X \to BY$ or $X \to \epsilon$, where X, Y are nonterminals of $G_{A,a}$ and $B \in N$. Let $T_{A,a}$ be the start symbol of $G_{A,a}$.

Assume without loss of generality that the sets of nonterminals of the grammars $G_{A,a}$ and G are pairwise disjoint. This assumption can be enforced by renaming if necessary. Form the grammar G_1 by adding all the nonterminals and productions of all the $G_{A,a}$ to G. Take the start symbol of G_1 to be S. Productions of G_1 are of the following three forms:

$$X \to b, \qquad X \to \epsilon, \qquad X \to BY,$$

where $b \in \Sigma$ and $B \in N$. Note that G_1 is trivially equivalent to G, since none of the new nonterminals can be derived from S.

Now let G_2 be the grammar obtained from G_1 by removing any production of the form

$$X \to BY$$

and replacing it with the productions

$$X \to bT_{B,b}Y$$

for all $b \in \Sigma$. Productions of G_2 are of the form

$$X \to b, \qquad X \to \epsilon, \qquad X \to bT_{B,b}Y,$$

where $b \in \Sigma$.

Finally, get rid of the ϵ-productions in G_2 using the construction of Lemma 21.3. This construction does not introduce any unit productions, since every non-ϵ-production has a terminal symbol on the right-hand side. Thus the resulting grammar G_3 is in Greibach form with at most two nonterminals on the right-hand side of any production.

Before we prove that $L(G_3) = L(G)$, let's pause and illustrate the construction with an example.

Example 21.6 Consider the balanced parentheses of Example 21.5. Starting with the Chomsky grammar

$$S \to AB \mid AC \mid SS, \qquad C \to SB, \qquad A \to [, \qquad B \to],$$

[1] See Homework 5, Exercise 1.

first compute the regular sets $R_{D,d}$:

$$\begin{aligned}
R_{S,[} &= (B+C)S^*, \\
R_{C,[} &= (B+C)S^*B, \\
R_{A,[} &= R_{B,]} = \{\epsilon\},
\end{aligned}$$

and all others are $\varnothing$. Here are strongly right-linear grammars for these sets:

$$\begin{array}{lll}
T_{S,[} \rightarrow BX \mid CX, & X \rightarrow SX \mid \epsilon, & \\
T_{C,[} \rightarrow BY \mid CY, & Y \rightarrow SY \mid BZ, & Z \rightarrow \epsilon, \\
T_{A,[} \rightarrow \epsilon, & & \\
T_{B,]} \rightarrow \epsilon. & &
\end{array}$$

Combining these grammars with G and making the substitutions as described above, we obtain the grammar G_2:

$$\begin{array}{lll}
S \rightarrow [T_{A,[}B \mid [T_{A,[}C \mid [T_{S,[}S, & C \rightarrow [T_{S,[}B, & A \rightarrow [, \\
T_{S,[} \rightarrow]T_{B,]}X \mid [T_{C,[}X, & X \rightarrow [T_{S,[}X \mid \epsilon, & B \rightarrow], \\
T_{C,[} \rightarrow]T_{B,]}Y \mid [T_{C,[}Y, & Y \rightarrow [T_{S,[}Y \mid]T_{B,]}Z, & Z \rightarrow \epsilon, \\
T_{A,[} \rightarrow \epsilon, & T_{B,]} \rightarrow \epsilon. &
\end{array}$$

Removing ϵ-transitions, we get the Greibach grammar G_3:

$$\begin{array}{lll}
S \rightarrow [B \mid [C \mid [T_{S,[}S, & C \rightarrow [T_{S,[}B, & A \rightarrow [, \\
T_{S,[} \rightarrow]X \mid [T_{C,[}X \mid b \mid [T_{C,[}, & X \rightarrow [T_{S,[}X \mid [T_{S,[}, & B \rightarrow], \\
T_{C,[} \rightarrow]Y \mid [T_{C,[}Y, & Y \rightarrow [T_{S,[}Y \mid]. &
\end{array}$$

The Greibach grammar produced by this construction is by no means the simplest possible, as can be seen by comparing it to the somewhat simpler (21.2). □

Now we prove that $L(G) = L(G_3)$. Surely $L(G) = L(G_1)$, since none of the new nonterminals added in the construction of G_1 can be derived from any nonterminal of G, including the start symbol S of G_1. Also, $L(G_2) = L(G_3)$ by Lemma 21.3. Thus the heart of the proof is showing that $L(G_1) = L(G_2)$.

Lemma 21.7 *For any nonterminal X and $x \in \Sigma^*$,*

$$X \xrightarrow[G_1]{*} x \iff X \xrightarrow[G_2]{*} x.$$

Proof. The proof is by induction on the length of derivations. If x can be derived in one step from X in either grammar, then it must be by a production of the form $X \rightarrow b$ or $X \rightarrow \epsilon$, and these productions are the same in both grammars.

For the induction step, we show that

$$X \xrightarrow[G_1]{*} x \text{ starting with the production } X \rightarrow BY$$

if and only if

$$X \xrightarrow[G_2]{*} x \text{ starting with the production } X \rightarrow bT_{B,b}Y,$$

where b is the first symbol of x. Note that x must have a first symbol, since derivations in G_1 starting with $X \rightarrow BY$ cannot generate ϵ, because B is a nonterminal of the original CNF grammar G, therefore can generate only nonnull strings.

Any leftmost derivation

$$X \xrightarrow[G_1]{1} BY \xrightarrow[G_1]{*} bz$$

is of the form

$$X \xrightarrow[G_1]{1} BY \xrightarrow[G_1]{k+1} bB_1B_2 \cdots B_kY \xrightarrow[G_1]{m} bz,$$

where $bB_1B_2 \cdots B_kY$ is the first sentential form in the sequence in which the terminal b appears, and $B_1B_2 \cdots B_k \in R_{B,b}$. By definition of the grammar $G_{B,b}$, this occurs if and only if

$$X \xrightarrow[G_2]{1} bT_{B,b}Y \xrightarrow[G_1]{k+1} bB_1B_2 \cdots B_kY \xrightarrow[G_1]{m} bz,$$

where the subderivation

$$T_{B,b} \xrightarrow[G_1]{k+1} B_1B_2 \cdots B_k$$

is a leftmost derivation in $G_{B,b}$. By the induction hypothesis, this occurs iff

$$X \xrightarrow[G_2]{1} bT_{B,b}Y \xrightarrow[G_2]{*} bz. \qquad \square$$

It follows from Lemma 21.7 by taking $X = S$ that $L(G_1) = L(G_2)$, therefore $L(G) = L(G_3)$.

Historical Notes

Bar-Hillel, Perles, and Shamir [8] showed how to get rid of ϵ- and unit productions. Chomsky and Greibach normal forms are due to Chomsky [18] and Greibach [53], respectively.