Supplementary Lecture G

The Chomsky–Schützenberger Theorem

Let PAREN_n denote the language consisting of all balanced strings of parentheses of n distinct types. This language is generated by the grammar

$$S \to \underset{1}{[} S \underset{1}{]} \mid \underset{2}{[} S \underset{2}{]} \mid \cdots \mid \underset{n}{[} S \underset{n}{]} \mid SS \mid \epsilon.$$

The languages PAREN_n are sometimes called *Dyck languages* in the literature.

The following theorem shows that the parenthesis languages PAREN_n play a special role in the theory of context-free languages: *every* CFL is essentially a parenthesis language modified in some relatively simple way. In a sense, balanced parentheses capture the essential structure of CFLs that differentiates them from the regular sets.

Theorem G.1 **(Chomsky–Schützenberger)** *Every context-free language is a homomorphic image of the intersection of a parenthesis language and a regular set. In other words, for every CFL A, there is an $n \geq 0$, a regular set R, and a homomorphism h such that*

$$A = h(\text{PAREN}_n \cap R).$$

Recall from Lecture 10 that a *homomorphism* is a map $h : \Gamma^* \to \Sigma^*$ such that $h(xy) = h(x)h(y)$ for all $x, y \in \Gamma^*$. It follows from this property that $h(\epsilon) = \epsilon$ and that h is completely determined by its values on Γ. The

homomorphic image of a set $B \subseteq \Gamma^*$ under h is the set $\{h(x) \mid x \in B\} \subseteq \Sigma^*$, denoted $h(B)$.

Proof. Let $G = (N, \Sigma, P, S)$ be an arbitrary CFG in Chomsky normal form. Denote productions in P by $\pi, \rho, \sigma, \ldots$.

For $\pi \in P$, define

$$\pi' = \begin{cases} A \to \overset{1}{\underset{\pi}{[}} B \overset{12}{\underset{\pi\pi}{]}} \overset{2}{\underset{\pi}{[}} C \overset{2}{\underset{\pi}{]}} & \text{if } \pi = A \to BC, \\ A \to \overset{1}{\underset{\pi}{[}} \overset{12}{\underset{\pi\pi}{]}} \overset{2}{\underset{\pi}{[}} \overset{2}{\underset{\pi}{]}} & \text{if } \pi = A \to a; \end{cases}$$

and define the grammar $G' = (N, \Gamma, P', S)$ with

$$\Gamma = \{\overset{1}{\underset{\pi}{[}}, \overset{1}{\underset{\pi}{]}}, \overset{2}{\underset{\pi}{[}}, \overset{2}{\underset{\pi}{]}} \mid \pi \in P\},$$

$$P' = \{\pi' \mid \pi \in P\}.$$

The idea here is that a balanced string of parentheses generated by G' encodes a corresponding string generated by G along with its parse tree.

Let PAREN_Γ be the parenthesis language over parentheses Γ. Surely $L(G') \subseteq \text{PAREN}_\Gamma$, since the productions of G' generate parentheses in well-nested matched pairs. However, not all strings in PAREN_Γ are generated by G'. Here are some properties satisfied by strings in $L(G')$ that are not satisfied by strings in PAREN_Γ in general:

(i) Every $\overset{1}{\underset{\pi}{]}}$ is immediately followed by a $\overset{2}{\underset{\pi}{[}}$.

(ii) No $\overset{2}{\underset{\pi}{]}}$ is immediately followed by a left parenthesis.

(iii) If $\pi = A \to BC$, then every $\overset{1}{\underset{\pi}{[}}$ is immediately followed by $\overset{1}{\underset{\rho}{[}}$ for some $\rho \in P$ with left-hand side B, and every $\overset{2}{\underset{\pi}{[}}$ is immediately followed by $\overset{1}{\underset{\sigma}{[}}$ for some $\sigma \in P$ with left-hand side C.

(iv) If $\pi = A \to a$, then every $\overset{1}{\underset{\pi}{[}}$ is immediately followed by $\overset{1}{\underset{\pi}{]}}$ and every $\overset{2}{\underset{\pi}{[}}$ is immediately followed by $\overset{2}{\underset{\pi}{]}}$.

In addition, all strings x such that $A \xrightarrow[G']{*} x$ satisfy the property

(v_A) The string x begins with $\overset{1}{\underset{\pi}{[}}$ for some $\pi \in P$ with left-hand side A.

Each of the properties (i) though (v_A) can be described by a regular expression; thus the sets

$$R_A = \{x \in \Gamma^* \mid x \text{ satisfies (i) through } (\text{v}_A)\}$$

are regular. We claim

Lemma G.2 $A \xrightarrow[G']{*} x \iff x \in \mathrm{PAREN}_\Gamma \cap R_A$.

Proof. The direction ($\Rightarrow$) is a straightforward proof by induction on the length of the derivation. For the direction ($\Leftarrow$), suppose $x \in \mathrm{PAREN}_\Gamma \cap R_A$. We proceed by induction on the length of x. It follows from properties (i) through (v_A) and the fact that x is a string of balanced parentheses that x is of the form

$$x = \underset{\pi}{\overset{1}{[}} y \underset{\pi}{\overset{1}{]}} \underset{\pi\pi}{\overset{2}{[}} z \underset{\pi}{\overset{2}{]}}$$

for some $y, z \in \Gamma^*$ and π with left-hand side A. If $\pi = A \to BC$, then from property (iii), y satisfies (v_B) and z satisfies (v_C). Also, y and z satisfy (i) through (iv) and are balanced. Thus $y \in \mathrm{PAREN}_\Gamma \cap R_B$ and $z \in \mathrm{PAREN}_\Gamma \cap R_C$. By the induction hypothesis, $B \xrightarrow[G']{*} y$ and $C \xrightarrow[G']{*} z$; therefore,

$$A \xrightarrow[G']{1} \underset{\pi}{\overset{1}{[}} B \underset{\pi}{\overset{1}{]}} \underset{\pi\pi}{\overset{2}{[}} C \underset{\pi}{\overset{2}{]}} \xrightarrow[G']{*} \underset{\pi}{\overset{1}{[}} y \underset{\pi}{\overset{1}{]}} \underset{\pi\pi}{\overset{2}{[}} z \underset{\pi}{\overset{2}{]}} = x.$$

If $\pi = A \to a$, then from property (iv), $y = z = \epsilon$, and

$$A \xrightarrow[G']{1} \underset{\pi}{\overset{1}{[}} \underset{\pi\pi}{\overset{1}{]}} \underset{\pi\pi}{\overset{2}{[}} \underset{\pi}{\overset{2}{]}} = x. \qquad \square$$

It follows from Lemma G.2 that $L(G') = \mathrm{PAREN}_\Gamma \cap R_S$. Now define the homomorphism $h : \Gamma^* \to \Sigma^*$ as follows. For π of the form $A \to BC$, take

$$h(\underset{\pi}{\overset{1}{[}}) = h(\underset{\pi}{\overset{1}{]}}) = h(\underset{\pi}{\overset{2}{[}}) = h(\underset{\pi}{\overset{2}{]}}) = \epsilon.$$

For π of the form $A \to a$, take

$$h(\underset{\pi}{\overset{1}{]}}) = h(\underset{\pi}{\overset{2}{[}}) = h(\underset{\pi}{\overset{2}{]}}) = \epsilon,$$

$$h(\underset{\pi}{\overset{1}{[}}) = a.$$

Applying h to the production π' of P' gives the production π of P; thus $L(G) = h(L(G')) = h(\mathrm{PAREN}_\Gamma \cap R_S)$. This completes the proof of the Chomsky–Schützenberger theorem. $\square$

Historical Notes

The pivotal importance of balanced parentheses in the theory of context-free languages was recognized quite early on. The Chomsky–Schützenberger theorem is due to Chomsky and Schützenberger [19, 22].