LECTURES 5 & 6: FORMAL LANGUAGES AND COMPLEXITY
October 19, 1998

Bud Mishra

1 Context Free Grammar

These are described by four objects as in the following example:

$$G_1 = (\{A\}, \{0,1\}, R, A),$$

The set $\{A\}$ correspond to its *nonterminal symbols* or *variables*; 0 and 1 are its *terminal symbols*; A is its *start symbol*–a distinguished nonterminal. Finally R are its *production rules* each rule of the form

$$V \to (T \cup V)^*,$$

The left hand side is just a variable and the right hand side is a string made of variables and terminals. For instance, we may have

$$\begin{aligned} A &\to 0A1 \\ A &\to \epsilon \end{aligned}$$

Using these rules we can derive a string in the language:

$$\begin{aligned} A &\to 0A1 \\ &\to 00A11 \\ &\to 000A111 \\ &\to 000111 \end{aligned}$$

From this we conclude that the language generated by this grammar is

$$\mathcal{L}(G_1) = \{0^n 1^n | n \geq 0\} = \text{Context Free Language, CFL.}$$

Definition. 1.1 A *context free grammar*, G is a quadruple (4-tuple)

$$G = (V, \Sigma, R, S),$$

where

1. $V =$ A finite set of variables.

2. $\Sigma =$ A finite set of terminals (disjoint from V, $V \cap \Sigma = \emptyset$).

3. $R \subseteq V \times (V \cup \Sigma)^*, =$ Finite set of rules, each rule being a variable and a string of variables and terminals.

4. $S \in V, =$ Distinguished start symbol.

Let u, v, $w \in (V \cup \Sigma)^*$ be a string of variables and terminals. Let $A \to w \in R$ be a rule of the grammar. We say

$$uAv \text{ yields } uwv \quad uAv \Rightarrow uwv.$$

Consider the reflexive transitive closure of yield ($\Rightarrow$).

$$u \Rightarrow^* v,$$

if 1) $u = v$ or 2) $\exists$ a sequence $u_1, u_2, \ldots, u_k$, $k \geq 0$ such that

$$u \Rightarrow u_1 \Rightarrow u_2 \Rightarrow \cdots \Rightarrow u_k \Rightarrow v.$$

Define the language of the grammar G as

$$\mathcal{L}(G) = \{\sigma \in \Sigma^* | S \Rightarrow^* \sigma\}.$$

See Sipser for many examples of CFL's and how to construct a CFG.

2 Facts

1. *Regular languages are context free*

   Let
   $$\mathcal{L} = \mathcal{L}(M), \quad \text{where } M = (Q, \Sigma, \delta, q_0, F) = \text{ DFA }.$$
   Construct a context free grammar (also, called left linear) as follows:
   $$G = (Q, \Sigma, R, q_0)$$
   The production rules R are as follows:
   $$\begin{array}{rll} \langle q_i \rangle \to a\langle q_j \rangle & \text{if} & \delta(q_i, a) = q_j \\ \langle q_i \rangle \to \epsilon & \text{if} & q_i \in F \end{array}$$

2. *The class of context free languages is closed under union, concatenation and Kleene star operations.*

Consider two grammars

$$G_1 = (V_1, \Sigma_1, R_1, S_1) \quad G_2 = (V_2, \Sigma_2, R_2, S_2)$$

where $V_1 \cap V_2 = \emptyset$.

$$\mathcal{L}_1 = \mathcal{L}(G_1) \quad \mathcal{L}_2 = \mathcal{L}(G_2).$$

First construct the grammar G_3 as

$$G_3 = (V_1 \cup V_2 \cup \{S_3\}, \Sigma_1 \cup \Sigma_2, R_3, S_3),$$

where

$$R_3 = \{S_3 \to S_1 \mid S_2\} \cup R_1 \cup R_2.$$

Clearly

$$\mathcal{L}(G_3) = \mathcal{L}_1 \cup \mathcal{L}_2$$

is context free.

First construct the grammar G_4 as

$$G_4 = (V_1 \cup V_2 \cup \{S_4\}, \Sigma_1 \cup \Sigma_2, R_4, S_4),$$

where

$$R_4 = \{S_4 \to S_1 S_2\} \cup R_1 \cup R_2.$$

Clearly

$$\mathcal{L}(G_4) = \mathcal{L}_1 \circ \mathcal{L}_2$$

is context free.

Next construct the grammar G_5 as

$$G_5 = (V_1 \cup \{S_5\}, \Sigma_1, R_5, S_5),$$

where

$$R_5 = \{S_5 \to \epsilon \mid S_1 S_5\} \cup R_1 \cup R_2.$$

Clearly

$$\mathcal{L}(G_5) = \mathcal{L}_1^*$$

is context free.

3. *The class of context free languages is closed under homomorphisms and inverse homomorphisms.*

4. *The class of context free languages is* <u>*not*</u> *closed under intersection or complementation.*

$$\mathcal{L}_2 = \{a^n b^n | n \geq 1\}, \quad \mathcal{L}_3 = \{b^n c^n | n \geq 1\}.$$

   Clearly both $\mathcal{L}_2$ and $\mathcal{L}_3$ are CFL's.

$$\mathcal{L}_2 \circ c^* \cap a^* \circ \mathcal{L}_3 = \{a^n b^n c^n | n \geq 1\} = \mathcal{L}_1.$$

   We will see that $\mathcal{L}_1$ is not context free. But since regular languages are context free and since CFL's are closed under concatenation, it must be the case that CFL's are not closed under intersection.

5. *The class of context free languages is closed under intersection with a regular language.*

$$\mathcal{L} = \text{ CFL } \& R = \text{ regular } \Rightarrow \mathcal{L} \cap R = \text{ CFL }.$$

3 Pumping Lemmas

We will do this in two steps:

1) First we will see (next lecture) that CFG's can be written in special Normal Forms:

- **Chomsky Normal Forms**: All rules are of the following forms:

$$\begin{aligned} A &\rightarrow BC \\ A &\rightarrow a \end{aligned}$$

  $A \in V$; $B, C \in V \setminus \{S\}$; $a \in \Sigma$.

- **Greibach Normal Forms**: All rules are of the following forms:

$$\begin{aligned} A &\rightarrow a\alpha \\ A &\rightarrow a \end{aligned}$$

  $A \in V$; $\alpha \in V^*$; $a \in \Sigma$.

2) Use these facts to show a Pumping Lemma in Chomsky Normal Forms.

Theorem. 3.1 *Let L be a context free language. Then there is a p (depending on $\mathcal{L}$, $p \equiv$ pumping length), such that*
$\forall \sigma \in \mathcal{L}$ *with* $|\sigma| \geq p$
$\exists$ *strings* $u, v, w, x, y \in \Sigma^*$, $\sigma = uvwxy$ *such that*

1. $|vx| > 0$.
2. $|vwx| \leq p$.
3. $\forall i \geq 0, uv^i wx^i y \in \mathcal{L}$.

PROOF. See Sipser.

4 Chomsky Normal Form

Definition. 4.1 A context free grammar is in Chomsky normal form, if every rule is of the form:

$$\begin{aligned} A &\rightarrow BC, \quad B, C \in V \\ A &\rightarrow a, \quad a \in \Sigma \end{aligned}$$

where a is any terminal and A, B and C are any variables, except that B and C may not be the start variable. In addition, we permit the rule $S \rightarrow \epsilon$, where S is the start variable.

Theorem. 4.2 *Any context free language is generated by a context free grammar in Chomsky Normal Form.*
PROOF. See Sipser. Also study the example in Sipser.

5 Greibach Normal Form

Definition. 5.1 A context free grammar is in Greibach normal form, if every rule is of the form:

$$A \rightarrow a\alpha, \quad a \in \Sigma, \alpha \in V^*$$

where a is any terminal and α is a (possibly empty) string of variables. In addition, we permit the rule $S \rightarrow \epsilon$, where S is the start variable.

Theorem. 5.2 *Any context free language is generated by a context free grammar in Greibach Normal Form.*

PROOF.

Lemma 1 Any pair of rules

$$\begin{aligned} A &\rightarrow \alpha_1 B \alpha_2 \\ B &\rightarrow \beta_1 \mid \beta_2 \mid \cdots \mid \beta_r \end{aligned}$$

can be replaced by a rule

$$A \rightarrow \alpha_1\beta_1\alpha_2 \mid \alpha_1\beta_2\alpha_2 \mid \cdots \mid \alpha_1\beta_r\alpha_2$$

Lemma 2 Any pair of rules

$$\begin{aligned} A &\rightarrow A\alpha_1 \mid A\alpha_2 \mid \cdots \mid A\alpha_r \\ A &\rightarrow \beta_1 \mid \beta_2 \mid \cdots \mid \beta_s \end{aligned}$$

can be replaced by a set of rules

$$\begin{aligned} A &\rightarrow \beta_1 \mid \beta_2 \mid \cdots \mid \beta_s \\ A &\rightarrow \beta_1 B \mid \beta_2 B \mid \cdots \mid \beta_s B \\ B &\rightarrow \alpha_1 B \mid \alpha_2 B \mid \cdots \mid \alpha_r B \\ B &\rightarrow \alpha_1 \mid \alpha_2 \mid \cdots \mid \alpha_r. \end{aligned}$$

Note that

$$A \rightarrow (\beta_1 + \cdots + \beta_s)(\alpha_1 + \cdots + \alpha_r)^*.$$

Continuing with the proof: Let $G = (V, \Sigma, R, S)$ be a Chomsky Normal form grammar generating the CFL, $\mathcal{L}$.

Order the variables of G as

$$V = \{A_1, A_2, \ldots, A_m\}.$$

Next the modify the rules so that

$$\forall 1 \leq i \leq m \; A_i \rightarrow A_j\gamma$$

is a rule with $j > i$. This can be done in two steps:

1. If $A_k \to A_j\alpha$ is a rule with $j < k$ and $A_j \to \beta$ is another rule, replace them by $A_k \to \beta\alpha$.
2. If $A_k \to A_k\alpha$ is a rule and $A_k \to \beta$ is another rule, replace them by a set of rules

$$\begin{aligned} B_k &\to \alpha \mid \alpha B_k \\ A_k &\to \beta \mid \beta B_k. \end{aligned}$$

At the end of this process we have following rules:

$$\begin{aligned} A_i &\to A_j\gamma, \quad j > i \\ A_i &\to a\gamma, \quad a \in \Sigma \\ B_i &\to \gamma, \quad \gamma \in (V \cup \{B_1, \ldots, B_{i-1}\})^* \end{aligned}$$

Note that $A_m \to a\gamma$ is a rule then the first symbol $a \in \Sigma$. Any rule with A_{m-1} in the left hand side has a right hand side starting with a terminal or A_m; by substitution, this can be put in the correct form. That is replace $A_{m-1} \to A_m\beta$ by $A_{m-1} \to a\gamma\beta$. By successively modifying the right hand sides of the rules in the order A_{m-2}, A_{m-3}, ..., A_2 and A_1, we get the correct form.

5.1 Example

Start with the following grammar, G:

$$\begin{aligned} A_1 &\to A_2A_3 \\ A_2 &\to A_3A_1 \mid b \\ A_3 &\to A_1A_2 \mid a \end{aligned}$$

Modify the last rule $A_3 \to A_1A_2$:

$$\begin{aligned} A_1 &\to A_2A_3 \\ A_2 &\to A_3A_1 \mid b \\ A_3 &\to A_2A_3A_2 \mid a \end{aligned}$$

Modify the last rule $A_3 \rightarrow A_2A_3A_2$:

$$\begin{aligned}
A_1 &\rightarrow A_2A_3 \\
A_2 &\rightarrow A_3A_1 \mid b \\
A_3 &\rightarrow A_3A_1A_3A_2 \mid bA_3A_2 \mid a
\end{aligned}$$

Modify the last rule $A_3 \rightarrow A_3A_1A_3A_2$:

$$\begin{aligned}
A_1 &\rightarrow A_2A_3 \\
A_2 &\rightarrow A_3A_1 \mid b \\
A_3 &\rightarrow bA_3A_2 \mid a \mid bA_3A_2B_3 \mid aB_3 \\
B_3 &\rightarrow A_1A_3A_2 \mid A_1A_3A_2B_3
\end{aligned}$$

Finally, after rewriting:

$$\begin{aligned}
A_3 &\rightarrow bA_3A_2B_3 \mid bA_3A_2 \mid aB_3 \mid a \\
A_2 &\rightarrow bA_3A_2B_3A_1 \mid bA_3A_2A_1 \mid aB_3A_1 \mid aA_1 \mid b \\
A_1 &\rightarrow bA_3A_2B_3A_1A_3 \mid bA_3A_2A_1A_3 \mid aB_3A_1A_3 \mid aA_1A_3 \mid bA_3 \\
B_3 &\rightarrow bA_3A_2B_3A_1A_3A_3A_2 \mid bA_3A_2B_3A_1A_3A_3A_2B_3 \mid bA_3A_2A_1A_3A_3A_2 \\
&\quad \mid bA_3A_2A_1A_3A_3A_2B_3 \mid aB_3A_1A_3A_3A_2 \mid aB_3A_1A_3A_3A_2B_3 \mid aA_1A_3A_3A_2 \\
&\quad \mid aA_1A_3A_3A_2B_3 \mid bA_3A_3A_2 \mid bA_3A_3A_2B_3
\end{aligned}$$