4CS4-6: Theory of Computation (Pumping Lemma for Context-free languages)

Lecture 14: March, 17, 2020

Prof. K.R. Chowdhary *: Professor of CS*

14.1 Pumping Lemma for Context-free languages

Here we present a some what similar to the regular languages pumping lemma, but more complex one, for context-free languages. It is used to show that, for example, the language $\{a^n b^n c^n \mid n \geq 0\}$ cannot be generated by context-free grammars. Roughly speaking, the pumping lemma for CFLs claim that if we represent some part of the string in context-free language, then also the resulting string remains in the language.

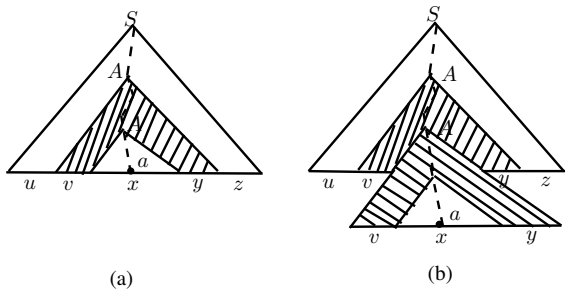

Figure 14.1: (a) $w = uvxyz$, (b) $w = uv^2xy^2z$

The Fig. 14.1 shows how this happens, since a string w has long path, from root with start symbol S, to a leaf node with a terminal, say a. In the parse-tree, there exists a non-terminal, say A, that appears at least twice on the path. This divides the string into five parts, u, v, x, y, z, so that the string obtained by repeating 2nd and 4th part still remains in the language (see Fig. 14.1(a), and in (b)), these parts are represented twice. Due to the nature of productions, which cause recursions, even if these parts (v, and y) are repeated arbitrary number of times, the string w should still belong to the language. More precisely,

$$S \Rightarrow^* uv^i xy^i z \Rightarrow^* w, \text{ for any } i \geq 0.$$

Definition 14.1 *For every context-free language L there is an integer m such that for every string $w \in L$, and $|w| \geq m$,*

1. *there exists u, v, x, y, z such that $w = uvxyz$,*

2. $vy \neq \varepsilon$,

3. $|vxy| \leq m$, *and*

4. *for all* $i, uv^ixy^iz \in L$.

Theorem 14.2 *Proof of Pumping Lemma for CFLs.*

Proof: Assume that the grammar $G = (V, \Sigma, S, P)$ is a Chomsky Normal Form (CNF). In case, the grammar is not in CNF, it can always be transformed into CNF format[1].

Let us examine the relationship between height of a parse-tree and the length of the string generated by that tree. For this, we try to analyze the parse-trees of heights 1 through 3, for productions $A \to AA \mid a$

Parse-tree of height 1: $A \Rightarrow a$.

Parse-tree of height 2: $A \Rightarrow AA \Rightarrow^* aa$.

Parse-tree of height 3: $A \Rightarrow AA \Rightarrow AAAA \Rightarrow^* aaaa$.

In CNF, the length of a string w generated by a parse-tree of height h is at most 2^{h-1}. In other words, we claim that generating a string $|w| = 2^{h-1}$ requires a parse-tree of height at least h.

Considering that $|V|$ is number of variables. Let height of parse-tree is, $h = |V| + 1$. This parse-tree generates a string w of at least $2^{h-1} = 2^{|V|+1-1} = 2^{|V|}$. Hence, length of string w is $|w| = m = 2^{|V|}$, where m is some integer. Then we can conclude that any parse-tree that generates a string w of length at least m, has path from root to a leaf, of length at least $|V| + 1$. Note that such a path consists of one node as a terminal at leaf, and at least $|V| + 1$ nodes including one non-terminal. Since the grammar G has only $|V|$ non-terminals, there can be at the most $|V|$ number of productions. Hence, one non-terminal, say, A (see Fig. 14.1(a) and (b)) appears at least twice on this path. For condition $|vxy| \leq m$, in the theorem to hold, let A is chosen from lower $|V| + 1$ non-terminals in the path.

As per Fig. 14.1(a) we have:

$$\begin{aligned} A &\Rightarrow^* \alpha A \beta \\ &\Rightarrow uvAyz \Rightarrow^* uvxyz \end{aligned}$$

where, $\alpha \Rightarrow^* u$, $\beta \Rightarrow^* z$, $A \Rightarrow^* uAy$, and $A \to x$.

But using these derivations, we can form the derivation (see Fig. 14.1 (b)),

[1]Chomsky normal form is $A \to BC \mid a$, where $A, B, C \in V$, and $a \in \Sigma$. Every grammar with no null production, has an equivalent CNF grammar, that generates the same language as the original.

$$S \Rightarrow uvxyz \Rightarrow^* uvvAyyz \Rightarrow^* uv^n xy^n z$$

for any positive integer n.

Now, we take individual conditions in the definition 14.1 of the pumping lemma for CFLs.

Condition 1: The height of upper A in Fig. 14.1(a) is at most $|V|+1$, hence length of vxy that upper A generates is at most $2^{|V|+1-1} = 2^{|V|} = m$. This proves the condition 1.

Condition 2: $|vy| \neq \varepsilon$, i.e., $|vy| \geq 1$.

The only possible ε-rule is $S \rightarrow \varepsilon$, and S does not appear in any right hand side of a production rule. Hence, S does not appear in the tree except at root[2].

Let $A \rightarrow BC$ be the rule that is applied to the upper A in the left side of the tree (see Fig. 14.1(a)). Then at least one of the v and y contain substring s such that $X \Rightarrow^* s$ for $X \in \{B, C\}$, $s \in \Sigma^*$ such that $|s| \geq 1$. This completes the proof for condition (2), i.e., $|vy| \geq 1$.

Condition 3: For $i \geq 0$, $w = uv^i xy^i z \in L$.

For any arbitrary number i, we can construct a parse-tree that generates $uv^i xy^i z$ by patching the region corresponding to v and y, i times. For $i = 0$, we have to eliminate the entire hashed region in Fig. 14.1(a) and (b). ■

Theorem 14.3 *The set of strings of the form $0^n 1^n 0^n$ is not context-free.*

Proof: We shall use the similar argument we used to show that $a^n b^n$ is not regular in earlier chapters. Assume that strings of the form $0^n 1^n 0^n$ are context-free. Next, we apply the pumping lemma for CFL and show that there is some integer m such that for any string of the length at least m, and of the form $0^n 1^n 0^n$ satisfies the conditions of the pumping lemma. Let this string be $0^m 1^m 0^m$. We know that pumping lemma assumes that:

a) $\exists u, v, x, y, z$ such that $w = uvxyz = 0^m 1^m 0^m$,

b) $vy \neq \varepsilon$, and

c) $\forall i, uv^i xy^i z$ is of the form $0^n 1^n 01^n$.

We now observe that either v and y each contain only one kind of symbol (0 or 1), or at least one of them contains some 0s and some 1s. In either case the string uv^2xy^2z cannot be of the form $0^n 1^n 0^n$[3]. This is because, either the runs of 0s and 1s will not be of the same length, or there will be more than three sequences. This contradicts the assumption $\forall i, uv^i xy^i z \in 0^n 1^n 01^n$. This proves the theorem. ■

[2]In arriving to CNF $A \rightarrow BC \mid a$, the ε-productions are removed, and when $w \neq \varepsilon$, $S \rightarrow \varepsilon$ is also removed.

[3]For example, if $w = 0^3 1^3 0^3$, let us pump $v = 00$, and $y = 00$, both to be pumped twice. Then, $w{=}0^5 1^3 0^5 \notin L$.

Theorem 14.4 *The family of context-free languages is not closed under the operation of intersection.*

Proof: Let there be two context-free grammars:

G_1	G_2
$S \to AB$	$S \to BA$
$S \to A$	$S \to A$
$A \to 0A1$	$A \to 1A0$
$A \to 01$	$A \to 10$
$A \to 0B$	$B \to 0B$
$B \to 0$	$B \to 0$

It is evident that the languages generated by these grammars are the sets of strings of the form $0^n1^no^m$ and $0^m1^n0^n$ for $n \geq 1$. The intersection of these languages is $0^n1^n0^n$, which is not context-free, as we have already proved in theorem 14.3. ■