

HW 4: Context Free Languages

Assigned: February 14, 2008

Due: February 21, 2008

Note: Take time to write clear and concise solutions. Confused and long-winded answers may be penalized. Consult the course webpage for course policies on collaboration.

1. (8 points) Let $\Sigma = \{0, 1\}$.
Write down a context free grammar for the complement of the language $\{0^n 1^n : n \geq 0\}$, that is, a grammar generating $L = \{w \in \{0, 1\}^* : w \neq 0^n 1^n, \forall n \geq 0\}$. Explain why your grammar is correct. [Hint: Try breaking L into simpler languages.]
Also, give a PDA that recognizes the same language. (No need to prove that your PDA is correct.)
2. (4 points) Define the *size* of a context-free grammar to be the total number of characters used in writing the rules of the grammar down (including variables, terminals, $|$ and \rightarrow). For example, the one-rule grammar $A \rightarrow A1 \mid \varepsilon$ has size six because it uses six characters.
Consider a grammar that generates *only* the string “a rose is a rose is a rose” and no other strings. Here the set of terminals is the set of small letters in the English alphabet and the whitespace character (denoted explicitly here by \square), i.e., it is the set $\{a, b, c, \dots, z, \square\}$. The smallest context-free grammar that generates only this string has size *twenty*. Write the rules for this grammar and prove that it recognizes only this string.
3. (12 points) Let G be a CFG in Chomsky Normal Form.
 - (a) (4 points) Prove that, if w is a string in $L(G)$ of length n where $n \geq 1$, then any derivation of w in G takes exactly $2n - 1$ steps.
 - (b) (8 points) Suppose G contains less than m variables. Prove that, if G generates a string with a derivation having at least 2^m steps, $L(G)$ is infinite.
[Hint: Use part (a) and the statement and proof idea of the CFL pumping lemma.]
4. (6 points) Recall the discussion we had in class about families of computer viruses that can be characterized as languages. A computer program is viewed as a string of instructions drawn from a finite alphabet Σ . A language characterizes a subset of virus programs.
Suppose you work at anti-virus software company ImmuneSystems. You have already characterized the members of two virus families with *context-free grammars* A and B . However, to foil your anti-virus software, the virus creators have released a new “mutation” tool that randomly combines virus programs from these two virus families using the $*$ (star) and \circ (concatenation) operators you know from class.

You are tasked with updating your software to recognize the new virus language. However, your co-worker Ignoramus complains that there is no context-free grammar that can recognize the new language (and hence the detection problem is “too hard”).

Prove Ignoramus wrong; i.e., show that the set of new viruses *is* a context-free language.

[Hint: Show how to construct a CFG generating the set of new viruses from A and B]