Solutions

October 29, 2014

1 Problem 6

Assume a grammar in Chomsky normal form has n non-terminals. Show that if the grammar can generate a word with a derivation having at least 2^n steps, then the recognized language should be infinite.

1.1 Solution

Let G be such a grammar in Chomsky normal form with n non-terminals. Let the parse tree for w contain at least 2^n derivation steps. We now show that in the parse tree of w there is a path from the start symbol to one of the terminals that contains at least $n+1$ non-terminals.

How many derivation do we have in a complete CNF parse tree of height n (without counting the terminals). There is exactly one derivation at level 1 that is of the form $S \to N_1 N_2$ and two children (N_1 and N_2). There are two derivations at level 2 of the form $N_1 \to N_3 N_4$ and $N_2 \to N_5 N_6$ and 4 children. There are 4 derivations at level 3 and 8 children so on. At level n-1, we have 2^{n-2} derivations and 2^{n-1} children. At level n all the 2^{n-1} children will derive terminals resulting in 2^{n-1} derivations. Therefore, the number of derivations of complete CNF parse tree of height n (excluding the leaves) is

$$1 + 2 + 4 + 8 + ... + 2^{n-2} + 2^{n-1} = 2^n - 1$$

Since, it is given that the number of derivations in parse of w is at least 2^n, it should have height at least $n + 1$ (excluding the leaves).

Therefore, there exists a path from the start symbol to a terminal along which at least one non-terminal repeats. For example, the path may look like: $N_1 \to \ldots \to M \to \ldots \to M \to \ldots N_{i_{n+1}} \to w_k$.

Therefore we can split the word w into: $w_I w_L w_M w_R w_F$

where w_M is the word derived by the sub-tree root at the second M and $w_L w_M w_R$ the word derived by the sub-tree rooted at the first M.

Because the first M does not immediately rewrite to the second M (there are no unit productions in a CNF grammar), at least one of the two w_L and w_R is not empty (there are also no epsilon productions).

Therefore, by replacing the second derivation tree starting from the second M by the one starting by the first M, we can generate the word:

$w_I w_L^2 w_M w_R^2 w_F$ wich is in the recognized language.

By reccurrence, we show that $w_I w_L^i w_M w_R^i w_F$ is in the recognized language. Therefore the recognized language is infinite. QED.