ON THE SIZE OF DERIVATION TREE

by

Y. Eric Cho

Memorandum No. ERL-M380

2 April 1973

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

ON THE SIZE OF DERIVATION TREE

by

Y. Eric Cho

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley, California 94720

ABSTRACT

The number of nodes in the derivation tree for a string w is shown to be linearly bounded by the length of w. This bound is also very convenient for proving the time bounds for deterministic parsing algorithms.

Research sponsored by the National Science Foundation Grant GK-10656x2

"Derivation tree" is a useful concept in formal language theory and the design of compilers [1],[2]. The problem that we are interested in is as follows. For a string w of length n, in the language of a context-free grammar G, what is the bound on the number of nodes in w's derivation tree with respect to G? If G is ambiguous, what is the bound for the smallest derivation tree for w? We prove that the bound is linear in n. As we shall see later, this bound is also very useful in proving the time and space bounds of deterministic parsers. An equivalent problem is the bound on the number of steps required to derive w. For the case that G is non-leftrecursive, it is proved in [3] that the bound is linear.

The basic definitions used in [1] are also used here. In addition, we employ the following:

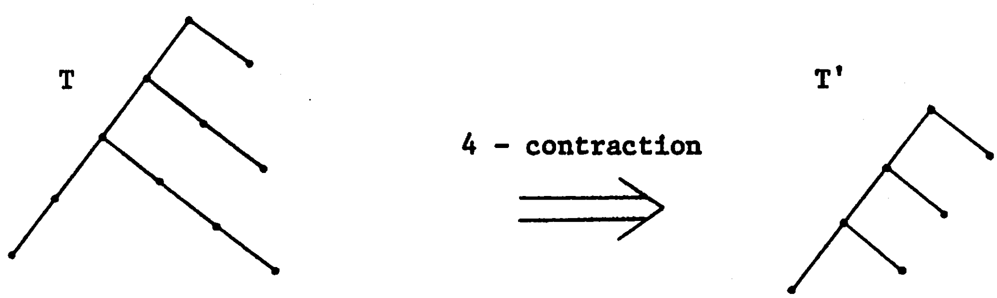
$\ell g(\alpha)$ is the length of the string α.

The terminal nodes of a tree are those nodes that do not have descendants (The concept of immediate descendant, and descendant have been defined in [1]). The other nodes are internal nodes.

A sequence of nodes $\{n_1, \ldots, n_k\}$ $K \geq 2$ is a chain if n_i is the only direct descendant of node n_{i-1} for $2 \leq i \leq K$. A chain is a maximum chain if n_1 is not the only direct descendant of another node and n_k does not have exactly one direct descendant. A tree is chain-free if it does not contain any chains.

For a tree T if all of its chains are shorter than K, then the K-contracted tree of T is obtained from T by replacing all the maximum chains of T by single nodes. More precisely, if $\{n_1, \ldots, n_k\}$ is a maximum chain, then the contraction of this chain is done by deleting nodes $\{n_2, \ldots, n_k\}$ and connecting n_1 to all of n_k's descendants.

Example:



For practical reasons, the bound will be derived for unambiguous grammars, since all the practical grammars are unambiguous. Later on, we will show that this bound is still true for the smallest derivation tree of w, if G is ambiguous.

Lemma 1. For a string $w \in L(G)$, if G is unambiguous, then w's derivation tree has no chain longer than ℓ where ℓ equals to the number of variables plus one.

Proof. If there is a chain longer than ℓ, then part of this chain would look like ..-A-B-D- . . . -A- . . There are infinitely many ways to derive A, then G is ambiguous which is a contradiction. Therefore, all the chains must be shorter than ℓ.

Lemma 2. For a tree T, if all the chains are shorter than ℓ, then the number of nodes in T is not more than ℓ times the number of nodes in T's ℓ-contracted tree T'.

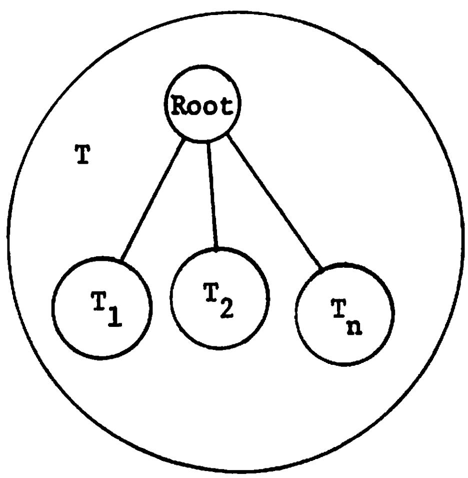
Proof. This can be proved by expanding T', replacing every node of T' by a chain of length ℓ; then we get a new tree T". It is obvious that T" has at least as many nodes as T, and the number of nodes in T" is less than the number of nodes in T'multiplied by ℓ.

Lemma 3. If a tree is chain-free, then it has more terminal nodes than internal nodes.

Proof. We are going to prove this by induction on the number of nodes in the tree -i.

(i) i=1. The tree consists of only one single node only. By our definition, it is a terminal node so this is true for i=1.

(ii) Suppose this lemma is true for trees with less than i nodes, we want to show that it is also true for trees with i nodes. It is worth mentioning that there is no chain-free tree or subtree with 2 nodes. For a tree with i nodes (i>2) we can partition the tree from the root as follows:



The tree T is formed by adding the "ROOT" to the subtrees $T_1, T_2, \ldots, T_n$ ($n \geq 2$). By induction, each subtree has at least one more terminal node than internal nodes, so after putting them together and $n \geq 2$, we still have more terminal nodes than internal nodes.

An interesting feature of context-free grammars is λ-rules. As a convention, we will count each λ in the derivation tree as one node.

Although Theorem 1 is proved for the case that G is unambiguous, actually, the only requirement is that the chains be shorter than ℓ. Therefore, if the grammar is ambiguous, in the derivation trees for w there will be one which does not have chains longer than ℓ. So we have:

<u>Corollary 2</u>. If $w \in L(G)$, then w has a derivation tree with less than $K \cdot \ell g(w)$ nodes.

In [2] it is proved that if G is nonleftrecursive, then for all $w \in L(G)$, the number of steps in deriving w is less than $K \cdot \ell g(w)$. Since the length of the derivation is not more than the number of nodes in the corresponding derivation tree, and by Corollary 2, we have the following more general result.

<u>Corollary 3</u>. For a string w in L(G), there exists a derivation whose length is less than $K \cdot \ell g(w)$.

ACKNOWLEDGEMENT

The author wishes to thank his research supervisor, Professor L. A. Zadeh, for the encouragement and guidance throughout the preparation of this paper; he also thanks Professor M. A. Harrison for the stimulating lectures on Parsing Theory and discussions, [illegible]

REFERENCES

1. Hopcroft, J. E. and Ullman, J. D., "Formal Languages and Their Relation to Automata," Addison-Wesley, 1969.
2. Gries, D., "Compiler Construction for Digital Computers," Wiley, 1971.
3. Aho, A. V. and Ullman, J. D., "The Theory of Parsing, Translation and Compiling," Vol. I, Prentice Hall, 1972
4. Gray, J. N. and Harrison, M. A., "On the Covering and Reduction Problems for Context-Free Grammars," JACM, October 1972, pp. 675-698.
5. Knuth, D. E., "On the Translation of Languages from Left to Right," Information and Control, 1965, pp. 607-639.
6. Wirth, N. and Weber, H., "EULER: A Generalization of ALGOL and its Formal Definitions, Parts I, II," CACM, January, February 1966, pp. 11-23, 89-99.
7. Gray, J. N. and Harrison, M. A., "Canonical Precedence Schemes," JACM (to appear).
8. Lewis, P. M. II and Stearns, R. E., "Syntax-Directed Transductions," JACM, 1968, pp. 465-488.
9. Rosencrantz, D. J. and Stearns, R. E., "Properties of Deterministic Top-Down Grammars," Information and Control, 1970, pp. 226-256.
10. Y. E. Cho, "A Fast Left-Corner Parsing Algorithm," (under preparation).