KYBERNETIKA — VOLUME 8 (1972), NUMBER 3

On the Size of Context-free Grammars

JOZEF GRUSKA

Two criteria of complexity of context-free grammars and languages are considered — the number of rules and the number of symbols — and hierarchy of complexity classes, undecidability of basic complexity problems and relation between complexity and unambiguity are established.

1. INTRODUCTION

In papers [2] and [3] four criteria of complexity of context-free grammars (CFG's), denoted by Var, Lev, Lev$_n$, and Depth, have been studied. These criteria reflect the intrinsic complexity of CFG's and they induce the criteria of complexity of context-free languages (CFL's) which reflect the intrinsic complexity of the description of CFL's by CFG's. The criterion Prod (G) = the number of rules of a CFG G, studied in [3] represents the size of CFG's.

In the present paper one more criterion of complexity of CFG's, namely Symb $(G) =$ = the number of all occurrences of all symbols in the rules of G, is defined and some results concerning the criteria Prod and Symb are derived.

2. PRELIMINARIES

A CFG G is quadruple $G = \langle V, \Sigma, P, \sigma \rangle$ where V is a finite set of symbols, $\Sigma \subset V$ and elements of Σ (of $V - \Sigma$) are called terminal symbols or terminals (nonterminal symbols or nonterminals); P is a finite set of rules of the form $A \to \alpha$ where $A \in V - \Sigma$, $\alpha \in V^*$; $\sigma \in V - \Sigma$ is called the initial symbol of G. If $A \to \alpha$ is in P and ω_1, ω_2 are in V^*, then we write $\omega_1 A \omega_2 \Rightarrow \omega_1 \alpha \omega_2$. Let $\overset{*}{\Rightarrow}$ be the transitive and reflexive closure of $\Rightarrow$ and let $L(G) = \{w;\ \sigma \overset{*}{\Rightarrow} w \in \Sigma^*\}$. A language L is said to be context-free if $L = L(G)$ for a CFG G. The symbol ε will denote the empty word.

For a CFG $G = \langle V, \Sigma, P, \sigma \rangle$ let Prod (G) be the number of rules of G and Symb $(G) = \sum_{p \in P}$ Symb (p) where Symb (p) is the length of the right side of p increased by 2. For a CFL L and K = Symb or Prod let

$$K(L) = \{\min K(G); L(G) = L\},$$

be the complexity of L with respect to K.

3. HIERARCHY OF COMPLEXITY CLASSES

The criteria Prod and Symb induce infinite hierarchies of CFL's and as the following theorem shows there are no gaps in these hierarchies.

Theorem 1. *For any integer n $(n \geqq 2)$ there is a* CFL $L_n \subset \{a\}^*$, $(L'_n \subset \{a\}^*)$ *such that* Prod $(L_n) = n$ (Symb $(L'_n) = n$).

Proof. The existence of a language $L_n \subset \{a\}^*$ with Prod $(L_n) = n$ was shown in [3] for any integer n and the existence of $L'_n \subset \{a\}^*$, $n > 2$, follows immediately from (i) to (iii):

(i) $$\text{Symb}(\{\epsilon\}) = 2 .$$

(ii) $$\text{Symb}(\{a^{j+1}\}) \leqq \text{Symb}(\{a^j\}) + 1 \text{ for any } j \geqq 0 .$$

(iii) For any k there are only a finite number of j's such that Symb $(\{a^j\}) \leqq k$.

Remark. If can be shown that Symb $(\{a^{2^i}\}) = 3i$ for i even and $3i + 1$ for i odd.

4. UNDECIDABILITY OF SOME COMPLEXITY PROBLEMS

One can effectively determine Prod (G) and Symb (G), given an arbitrary CFG G. However, can one effectively determine Prod $(L(G))$ and Symb $(L(G))$? The negative answer to this question and the undecidability of some other complexity problems concerning the criteria Prod and Symb is shown in this section.

Theorem 2. *If $n \geqq 2$, then it is undecidable for an arbitrary* CFG G *whether or not* Prod $(L(G)) = n$.

Proof. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be n-tuples of nonempty words over $\{a, b\}$ and $L(x)$, $L(y)$, $L(x, y)$, L_s and $L_{a,b}$ be languages defined by*

$$L(x) = \{ba^{i_k} \ldots ba^{i_1}c\, x_{i_1} \ldots x_{i_k};\ 1 \leqq i_j \leqq n,\ 1 \leqq j \leqq k\},$$

$$L(y) = \{ba^{i_k} \ldots ba^{i_1}c\, y_{i_1} \ldots y_{i_k};\ 1 \leqq i_j \leqq n,\ 1 \leqq j \leqq k\},$$

$$L(x, y) = L(x)\, c\, L^R(y),\quad L_s = \{w_1cw_2cw_2^Rcw_1^R;\ w_1w_2 \in \{a, b\}\}$$

* If w is a word, then w^R is the reverse of w and for a language L, $L^R = \{w^R; w \in L\}$.

and $L_{a,b}$ be the language generated by the grammar with two rules $\sigma \rightarrow \sigma a \sigma b$, $\sigma \rightarrow \epsilon$.

Let φ be a homomorphism on $\{a, b, c\}^*$ defined by $\varphi(a) = ab$, $\varphi(b) = aabb$ and $\varphi(c) = aaabbb$. By [1] given x and y, a CFG $G_{x,y}$ generating the language $L_{x,y} = \{a, b, c\}^* - L(x, y) \wedge L_s$ can be effectively constructed. From that it follows that for given x and y also CFG's $G'_{x,y}$ and $G''_{x,y}$ such that $L(G'_{x,y}) = L_{a,b} - \varphi(L(x, y) \wedge L_s) = (L_{a,b} - \{ab, aabb, aaabbb\}^*) \cup \varphi(L_{x,y})$, $L(G''_{x,y}) = \{a, b\}^* - \varphi(L(x, y) \wedge L_s)$ can be effectively constructed. It is easy to see that Prod $(L(G'_{x,y})) = 2$ (Prod $(L(G''_{x,y})) = 3$) if and only if $L(x, y) \wedge L_s = \theta$. On the other hand $L(x, y) \wedge L_s = \theta$ if and only if the Post correspondence problem for x and y has a solution and therefore the undecidability of Post correspondence problem implies the Theorem for $n = 2$ and $n = 3$.

For $n > 3$ we proceed as follows. By [3] for $m = n - 3$ a CFG G_m can be effectively constructed such that $L(G_m)$ is a finite subset of $\{d\}^*$ and Prod $(L(G_m)) = m$. Combining G_m and $G'_{x,y}$ we get a grammar for $L(G_m) \cup L(G'_{x,y})$. Clearly, Prod $(L(G_m) \cup L(G'_{x,y})) = n$ if and only if the Post correspondence problem for x and y has a solution and therefore also for $n > 3$ the Theorem follows from undecidability of the Post correspondence problem.

Corollary 3. *There is no effective method to determine* Prod $(L(G))$ *for an arbitrary* CFG G.

Theorem 4. *If* $n \leqq 7$ $(n \geqq 8)$, *then it is decidable (undecidable) for an arbitrary* CFG G *whether or not* Symb $(L(G)) = n$.

Proof. Symb $(L(G)) \leqq 7$ if and only if the language $L(G)$ has one of the following forms: $\{x\}$, $|x| \leqq 5$; $\{x_1, x_2\}$, $|x_1| + |x_2| \leqq 3$; $\{a\}^*$; $\{a\}^* b$; $b\{a\}^*$; $\{a^i b^i, i \geqq 0\}$; $\{ab\}^*$; where a and b are symbols or $L(G)$ is empty. Since any of these language is bounded and, moreover, see [1], given any bounded language L_0 it is decidable for an arbitrary CFG G whether of not $L(G) = L_0$, the theorem holds for $n \leqq 7$.

We will use notation of the proof of Theorem 2 in order to prove Theorem for $n \geqq 8$ and the proof will be again based on the undecidability of the Post correspondence problem from what it follows that it is undecidable for arbitrary x and y whether or not $L(x, y) \wedge L_s = \theta$. Now the proof can be reduce to determine Symb (L) for several simple languages and in all cases this can be done very easily.

First, we can see that Symb $(L(G'_{x,y})) = 8$ if and only if $L(x, y) \wedge L_s = 0$ and therefore the Theorem holds for $n = 8$. If now φ_1 and φ_2 are homomorphisms on $\{a, b\}^*$ defined by $\varphi_1(a) = \varphi_2(a) = a$, $\varphi_1(b) = b^2$, $\varphi_2(b) = b^3$, then for $i = 1, 2$, Symb $(\varphi_i(L_{a,b} - \varphi(L(x, y) \wedge L_s))) = 8 + i$ if and only if $L(x, y) \wedge L_s = \theta$ and the Theorem follows for $n = 9$ and $n = 10$. Moreover, Symb $(\{a, b, c\}^* - L(x, y) \wedge L_s) = 11$ if and only if $L(x, y) \wedge L_s = \theta$ and we have the Theorem for $n = 11$.

In order to prove Theorem for $n > 11$ we proceed as follows. By Theorem 1, there is a language $L_{n-9} = \{d^{i_n}\}$, i_n is an integer, such that Symb $(L_{n-9}) = n - 9$. Now it is easy to show that Symb $(L_{n-9} \, . \, L(G'_{x,z})) = n$ if and only if $L(x, y) \wedge L_s = \theta$ and this completes the proof of Theorem for $n \geqq 8$.

Corollary 5. *There is no algorithm to determine* Symb $(L(G))$ *for an arbitrary* CFG G.

Another question which is naturally to ask is whether or not one can effectively determine the simplest grammar for the language generated by a CFG. The answer follows immediately from Corollaries 3 and 5. (See also [5] for the first part of the corollary.)

Corollary 6. *There is no effective method to construct to an arbitrary* CFG G *a new* CFG G' *such that* $L(G) = L(G')$ *and* Prod $(G') =$ Prod $(L(G))$ (Symb $(G') =$ = Symb $(L(G))$).

We know now that there is no effective way to find the simplest grammar but can we at least to decide whether a given CFG is the simplest one?

Theorem 7. *It is undecidable for an arbitrary* CFG G *whether or not* Symb $(G) =$ = Symb $(L(G))$.

Proof. Would it be decidable, the following procedure would determine Symb $(L(G))$ for an arbitrary CFG G.

(i) Decide if G is the simplest grammar. If yes Symb $(L(G)) =$ Symb (G). If not go to step (ii).

(ii) Construct all CFG's which are simpler than G with respect to the criterion Symb. (There is only a finite number of such grammars

$$(*) \qquad G_1, G_2, \ldots, G_k$$

if we do not distinguish grammars which differ only in names of nonterminals.)

(iii) Remove from (*) all CFG's which are not the simplest CFG's with respect to Symb. Let

$$(**) \qquad G'_1, \ldots, G'_e$$

be the resulting sequence of CFG's.

(iv) Starting with (**), do for $n = 1, 2, \ldots$ step n by which the sequence (**) is subsequently reduced until Symb $(G_1) =$ Symb (G_2) for any two remaining CFG's G_1 and G_2. Then Symb $(L(G)) =$ Symb (G_1).

(n) For each grammar, say G_0, currently in (**) compare $\{x; x \in L(G_0), |x| \leqq n\}$. and $\{x; x \in L(G), |x| \leqq n\}$. If this two sets differ remove G_0 from (**); otherwise let G_0 in (**).

Now the Theorem follows from Corollary 6.

In the preceding Theorem only the criterion Symb is considered. We are convinced that the same is true for the criterion Prod but have no proof.

Open problem 1. Is it decidable for an arbitrary CFG G whether or not Prod $(G) = $ $=$ Prod $(L(G))$?

Open problem 2. Are the undecidability results of this section valid if only bounded CFG's and CFL's are considered?

5. COMPLEXITY AND UNAMBIGUITY

If was shown in [4], for the criteria Var, Lev, Lev$_n$ and Depth that the complexity and unambiguity are, in general, in conflict. The same is true for the criteria Prod and Symb. Indeed, let L_k be the language generated by the grammar.

$$\sigma \to a\sigma b\sigma \,, \quad \sigma \to b\sigma a\sigma \,, \quad \sigma \to \varepsilon \,.$$

By [4], any unambiguous CFG for L_k has at least two nonterminals and from that it follows easily that Prod $(G) > 3$ and Symb $(G) > 14$ for any unambiguous grammar G for L_k. By using the technique of the proofs of the foregoing Section we can show even more.

Theorem 8. *For any $n \geqq 3$ $(n \geqq 14)$, there is an unambiguous* CFL L_n (L'_n) *such that* Prod $(L_n) = n$ (Symb $(L'_n) = n$) *and* Prod $(G) > n$ (Symb $(G) > n$ *for any unambigous* CFG *for* L_n (*for* L'_n).

Remark. The only case which makes a little trouble is the case $n = 15$ for the criterion Symb. In this case the language generated by the grammar $\sigma \to a^2\sigma b$, $\sigma \to a^3\sigma b$, $\sigma \to \varepsilon$ should be considered.

(Received August 2, 1971.)

REFERENCES

[1] Ginsburg, S.: The mathematical theory of context-free languages. McGraw-Hill, New York 1966.

[2] Gruska, J.: On a classification of context-free grammars. Kybernetika *3* (1967), 1, 22–29.

[3] Gruska, J.: Some classifications of context-free languages. Information and Control *14* (1969), 152–179.

[4] Gruska, J.: Complexity and unambiguity of context-free grammars and languages. Information and Control *18* (1971), 502–519.

[5] Taniguchi, K , Kasami, T.: Reduction of Context-Free Grammars. Information and Control *17* (1970), 92–108.

O veľkosti bezkontextových gramatík

JOZEF GRUSKA

V práci sa vyšetrujú dve kritéria zložitosti bezkontextových gramatík a jazykov — počet pravidiel a počet symbolov. Ukazuje sa, že obe kritériá indukujú nekonečné hierarchie bezkontextových jazykov. Dokazuje sa nerozhodnuteľnosť základných problémov, týkajúcich sa vyšetrovaných kritérií zložitosti. V závere práce sa ukazuje, že pre niektoré jednoznačné bezkontextové jazyky sú jednoznačné gramatiky nutne zložitejšie, než viacznačné.

RNDr. Jozef Gruska CSc., Matematický ústav SAV (Mathematical Institute — Slovak Academy of Sciences), Štefánikova 41, Bratislava.