Generalized context-free grammars

By J. GRUSKA

1. *Introduction.* Generalized context-free grammars can be thought of as context-free grammars all rules of which are of the form $A \to \alpha$ where α is a regular expression. Generalized context-free grammars and their representation by a set of finite-state diagrams are a convenient tool to describe context-free languages. In this paper a classification of context-free languages according to the minimal number of non-terminals of generalized context-free grammars is studied and the corresponding decision problems are investigated.

2. *Definitions.* By a generalized context-free grammar we mean a quadruple $G = \langle V, \Sigma, P, \sigma \rangle$ where V, Σ and σ have the same meaning as for context-free grammars (see [2]) and P is a set (maybe infinite) of context-free rules such that for any nonterminal $A \in V - \Sigma$, the set $\{w; A \to w \in P\} \subset V^*$ is regular. The relations $\Rightarrow$ and $\overset{*}{\Rightarrow}$ for a generalized context-free grammar are defined in the same way as for context-free grammars.

It is obvious that a language L is context-free if and ony if $L = L(G)$ for a generalized context-free grammar G.

3. *Representations.* A generalized context-free grammar $G = \langle V, \Sigma, P, \sigma \rangle$ can be represented by a finite set of rules $A \to \alpha$, one for each nonterminal in $V - \Sigma$, where α is a regular expression over V. This in turn means that a generalized context-free grammar can be represented by a finite set of transition diagrams, one for each nonterminal of G, each of which represents a finite-state automaton which is capable of recursively calling other finite state automata [1], or G can be represented by a finite set of the so-called flag diagrams, one for each nonterminal of G [4].

4. *Problems.* As suggested by Kalmár [4], for a context-free language L let $N(L)$ be the minimum of the number of non-terminals of generalized context-free grammars generating L. Since $N(L)$ is also the minimum of transition diagrams for L, $N(L)$ may be thought of as a measure of non-finite state character of L.

5. *Results.* It will be shown now that for any integer n there is a context-free language L_n such that $N(L_n) = n$ and that there is no effective way to calculate $N(L)$.

Theorem 1. For any integer n there is a context-free language $L_n \subset \{a, b\}^*$ such that $N(L_n) = n$.

Proof. The case $n=1$ is trivial. Let now $n>1$ and let L_n be the language generated by the context-free grammar

$$\sigma \to a\sigma b, \qquad \sigma \to aba^2A_2bab$$

$$A_i \to a^iA_ib, \quad A_i \to ba^{i+1}A_{i+1}ba \qquad 2 \leqq i \leqq n-1$$

$$A_n \to a^nA_nb, \quad A_n \to b\sigma a, \quad A_n \to b^2a^2.$$

Let G be a generalized context-free grammar generating L_n and such that no generalized context-free grammar for L_n has fewer nonterminals. It means that from any nonterminal of G an infinite set of terminal words can be derived. All words of L_n posses a very regular structure. It holds

(1) If $x \in L_n$, then $x=ub^2a^2v$, u (v) is uniquelly determined by v (by u) and neither u nor v contains b^2a^2 as a subword.

From (1) it follows

(2) All rules of G are of the form $A \to uBv$ or $A \to ub^2a^2v$ where $u, v \in \Sigma^*$ and $B \in V-\Sigma$.

(3) If $A \to uBv$, $A \to u'Bv$ or $A \to ub^2a^2v$ and $A \to u'b^2a^2v$ are rules of G, then $u=u'$. If $A \to uBv$, $A \to uBv'$ or $A \to ub^2a^2v$, $A \to ub^2a^2v'$ are rules of G, then $v=v'$.

Since for any nonterminal A of G, the set $\{w;\ A \to w \in P\}$ is regular, it follows easily from (1) to (3) that the set P must be finite and therefore G is a "normal" context-free grammar. It was shown in [3], that the language L_n can not be generated by a context-free grammar having less than n nonterminals and therefore $N(L_n) \geqq n$. Since $N(L_n) \leqq n$ is obviously true we get the theorem.

Theorem 2. Let $n \geqq 1$ be an integer. It is undecidable for an arbitrary context-free grammar G whether or not $N(L(G))=n$.

Proof. Let us first consider the case $n=1$. Let x and y be arbitrary m-tuples of non-empty words over the alphabet $\{a, b\}$. Let $L(x)$, $L(x, y)$ and L_s be the languages defined by

$$L(x) = \{ba^{i_1}ba^{i_2} \dots ba^{i_k}cx_{i_k} \dots x_{i_2}x_{i_1};\ 1 \leqq i_j \leqq m\}$$

$$L(x, y) = L(x)cL^R(y)$$

$$L_s = \{w_1cw_2cw_2^Rcw_1^R;\ w_1w_2 \in \{a, b\}^*\}$$

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

By [2], given x and y, one can effectively construct a context-free grammar $G_{x,y}$ generating the language

$$L_{x,y} = \{a, b, c\}^* - L(x, y) \cap L_s.$$

If $L(x, y) \wedge L_s = \emptyset$, then obviously $N(L_{x,y})=1$. Let us now consider the case $L(x, y) \wedge L_s \neq \emptyset$ and let us assume that again $N(L_{x,y})=1$. Then there is a generalized context-free grammar $G=\langle V, \Sigma, P, \sigma\rangle$ with only one nonterminal σ which generates the language $L_{x,y}$.

Since $L(x, y) \wedge L_s \neq \emptyset$, there are indices $i_1, \dots, i_k$ such that if we denote

$$I=ba^{i_1}\dots ba^{i_k}, \quad X=x_{i_k}\dots x_{i_1}, \quad j=I^R, \quad Y=X^R$$

then $I^rcX^rcY^rcJ^r \in L_{x,y}$ for no integer $r \geqq 1$.

Since the set $R=\{\alpha;\ \sigma \to \alpha \in P\}$ is regular, there must exists an integer N such that if $i>N$, then $z_i=I^icX^{i+1}cY^{i+1}cJ^{i+1} \notin R$ and, moreover, if $u_i\sigma v_i \in R$, $u_iv_i \neq \varepsilon$,

$u_i \in \{a, b, c\}^*$, $u_i \sigma v_i \overset{*}{\Rightarrow} z_i$, then u_i does not contain the symbol c. Hence there exists a word $\bar{u}_i c \bar{v}_i \in L(G)$ such that $\bar{u}_i \in \{a, b\}^*$ and $u_i \bar{u}_i c \bar{v}_i v_i = z_i$. But then the word $\bar{u}_i I c \bar{v}_i$ is also in $L_{x,y}$ and therefore $L(G)$ generates the word $u_i \bar{u}_i I c \bar{v}_i v_i = I^{i+1} c X^{i+1} c Y^{i+1} c J^{i+1} \notin$ $\notin L_{x,y}$ what is a contradiction. Thus $N(L_{x,y}) = 1$ if and only if $L(x, y) \wedge L_s = \emptyset$. Since it is undecidable for arbitrary x and y whether or not $L(x, y) \wedge L_s = \emptyset$ [2], we get the theorem for the case $n = 1$.

For $n > 1$ we proceed as follows. By Theorem 2, for $n > 2$ there is a context-free language $L_{n-2} \subset \{d, e\}^*$ such that $N(L_{n-2}) = n - 2$. For $n = 2$ let us consider the language $L_{x,y,2} = \{a, b, c\}^* - L(x, y) \wedge L_s \cup \{f\}$ and for $n > 2$ let $L_{x,y,n} = L_{x,y} \cup \{f\} \cup L_{n-2}$ where f, d, e are new symbols. It is easy to verify that $N(L_{x,y,n}) = n$ if and only if $L(x, y) \wedge L_s = \emptyset$ and now the theorem for the case $n > 1$ follows in the same way as for $n = 1$.

Corollary. There is no effective way to construct for an arbitrary context-free grammar G a generalized context-free grammar with fewest states and generating the language $L(G)$.

It follows from this corollary that there is no effective way to determine for an arbitrary context-free grammar G the minimum of transition diagrams for the language $L(G)$. Can we, however, at least to minimize effectively the overall number of states of transition diagrams for $L(G)$? It was shown implicitly in the course of the proof of Theorem 2 that the answer is again in negative.

Обовщенные контекстно-свободные грамматики

Обобщенные контекстно-свободные грамматики — это грамматики имеющие правила вида $A \to \alpha$, где A вспомогательный символ и α регулярное выражение над основыми и вспомогательными символами. В работе установлена классификация контекстно-свободных языков в зависимости от минимального числа вспомогательных символов обощенных контекстно-свободных грамматик, которые порождают данный контекстно-свободный язык. Доказана алгоритмическая неразрешимость основных проблем связанных с этой классификацией, как напр. проблема построить минимальную грамматику для данного языка.

References

[1] Conway, M. E., Design of a separable transition-diagrams compiler, *Comm. ACM.* v. 6, 1963, pp. 396—408.

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

[3] Gruska, J., Some classifications of context-free languages, *Information and Control*, v. 14, 1969, pp. 152—173.

[4] Kalmár, L., An intuitive representation of context-free languages, COLING, *The proceedings of the International Conference on Computational Linguistics*, Sånga—Säby, 1969.

(Received April 18, 1972)