KYBERNETIKA — VOLUME 11 (1975), NUMBER 1

A Note on ε-rules in Context-Free Grammars

JOZEF GRUSKA

The importance of ε-rules in context-free grammars (CFG's) is investigated. It is shown how much can ε-rules simplify the description of a context-free language (CFL) and that one can not effectively construct the simplest ε-free CFG for a given CFL.

1. INTRODUCTION

In general, context-free grammars (CFG's) contain ε-rules. The purpose of this note is to explore both theoretical and "practical" importance of these rules for the description of context-free languages (CFL's). The main questions to be answered here are: How much can ε-free rules simplify the description of CFL's? Can one determine the simplest ε-free CFG to a given CFL?

2. BASIC DEFINITIONS

A CFG G is a quadruple $G = \langle V, \Sigma, P, \sigma \rangle$ where V is a finite set of symbols, $\Sigma \subset V$ and the elements of Σ (of $V - \Sigma$) are called terminals (nonterminals), P is a finite set of rules of the form $A \to \alpha$ where $A \in V - \Sigma$ and $\alpha \in V^*$, $\sigma \in V - \Sigma$ is called the initial symbol of G. If $A \to \alpha$ is in P and w_1, w_2 are in V^*, then we write $w_1 A w_2 \Rightarrow w_1 \alpha w_2$. The relation $\overset{*}{\Rightarrow}$ is the transitive and reflexive closure of $\Rightarrow$ and we define $L(G) = \{w, \sigma \overset{*}{\Rightarrow} w \in \Sigma^*\}$. A language L is termed context-free if $L = L(G)$ for a CFG G. The symbol ε will denote the empty word. A CFG G is said to be ε-free if G does not contain an ε-rule, i.e. a rule of the form $A \to \varepsilon$.

3. ARE ε-RULES NECESSARY?

The answer to this question is well-known. It holds:

Theorem 1. [1]. *There is an effective method how to construct to a given CFG G an ε-free CFG G' such that $L(G') = L(G) - \{\varepsilon\}$.*

This theorem implies.

Corollary. ε-rules do not increase the generative power of CFG's and therefore, they are not necessary.

Remark. Theorem 1 can be even strengthened by saying that if G is unambiguous, then G' can be constructed to be unambiguous, too [1].

4. DO ε-RULES SIMPLIFY THE SIZE AND THE "UNDERSTANDING" OF THE DESCRIPTION OF CONTEXT-FREE LANGUAGES?

In order to answer this question we have to introduce some criteria of complexity of CFG's.

The criteria Var, Depth, Lev, Lev_n [2] and Ind [3] characterize in a way the intrinsic complexity of CFG's. For any of them, let us call it K, $K(G)$ is an integer. The criterion K induces a criterion of complexity of CFL's which is also called K and defined as follows: $K(L) = \min \{K(G);\ L(G) = L\}$ for any CFL L. Thus, $K(L)$ represents the intrinsic complexity of the description of L by CFG's or the difficulty of the understanding of L. Similarly we can define for a CFL L not containing ε $K_\varepsilon(L) = \min \{K(G);\ G$ is ε-free and $L(G) = L\}$. As far as the above criteria are concerned we have a result which is easy to verify going through a standard procedure of constructing an ε-free grammar for $L(G) - \{\varepsilon\}$, given a CFG G.

Theorem 2. *If K is one of the criteria* Var, Lev, Lev_n, Depth *or* Ind, *then* $K(L) = $ $= K_\varepsilon(L - \{\varepsilon\})$.

This result may be interpreted as follows.

Corollary. ε-rules do not simplify the intrinsic complexity of the description of CFL's.

Two more criteria of complexity of CFG's $G = \langle V, \Sigma, P, \sigma\rangle$ are studied in [2] and [4]. They are Prod (G) = the number of rules of G and Symb $(G) = \sum_{p\in P}$ Symb (p) where Symb (p) is the lenght of the right side of p increased by 2. These two criteria represent "the size of CFG's". As the folloving theorem indicates, with regard to the criterion Prod the use of ε-rules can substantially simplify the description of CFL's.

Theorem 3. *For any integer n, there exists a CFL L_n such that* Prod $(L_n) = 2$ *and* $\mathrm{Prod}_\varepsilon$ $(L_n - \{\varepsilon\}) \geqq n$.

Proof. Let $a_1, a_2, \ldots$ be an infinite sequence of distinct symbols. For any integer m let G_m be the grammar with two rules

$$\sigma \to \sigma a_1 \sigma a_2 \ldots a_m \sigma\,, \quad \sigma \to \varepsilon$$

and let $L_m = L(G_m)$. In order to prove the theorem it is sufficient to show that there is no integer K such that for any m $\mathrm{Prod}_\varepsilon\,(L_m - \{\varepsilon\}) \leqq K$. The proof will be by contradiction but first we have to define some mappings. For any integer m, let φ_m and φ_m^* be mappings on $A_m = \{a_1, \ldots, a_m\}^*$ defined as follows. If $x \in A_m$, then $\varphi_m(x)$ is the word obtained from x by deleting the leftmost occurence (if any) of the subword $a_1 a_2 \ldots a_m$ and $\varphi_m^*(x) = \varphi_m^{|x|}(x)$.*

Let us now assume that there exists an integer K such that for any m there exists an ε-free CFG G_m generating $L_m - \{\varepsilon\}$ with no useless nonterminals and $\mathrm{Prod}\,(G_m) \leqq \leqq K$. Since $L_m = \{x\colon x \in A_m, \varphi_m^*(x) = \varepsilon\}$, the following assertion holds for any nonterminal A of G_m

(*) $\quad$ if $A \overset{*}{\Rightarrow} x_1\,,\quad A \overset{*}{\Rightarrow} x_2\,,\quad x_1 x_2 \in A_m\,,\quad$ then $\varphi_m^*(x_1) = \varphi_m^*(x_2)\,.$

(In the rest of this proof we will make an implicit use of this fact several times.)

Now let $C_m = \{x;\, x \in L_m$ and $|x| \leqq 2m\}$. In what follows we will modify in several steps the grammar G_m in such a way that at any stage the resulting grammar will generate a subset of L_m which contains C_m.

Step A. Remove from G_m all rules which contain a nonterminal more then twice. By (*) such rules cannot be used in any derivation of words of C_m. The resulting grammar, say G_m', has at most K rules and at most $2K$ nonterminals in any rule.

The step B will be carried out for every nonterminal but the initial symbol of G_m' and therefore less than K times.

Step B. (i) If the chosen nonterminal, say C, has no rule of the form $C \to uCv$, then remove all rules with C on left side and in the remaining rules make all possible replacements of C's by its right sides.

(ii) Let C have a rule of the form $C \to uCv$. Let $B_C = \{x;\, C \overset{*}{\Rightarrow} x$ in the grammar under consideration, and x can be derived from C in at most two steps and in each step a "C-rule" is used i.e. a rule $C \to \gamma\}$. Remove all rules with C on left side and in the remaining rules make all possible replacements of C's by words from B_C.

After finishing the step B we get a grammar, say G_m'', with the only one nonterminal, say σ. Since the grammar G_m' has at most K rules and each rule has at most $2K$ nonterminals, there exists an increasing function f_1 such that G_m'' has at most $f_1(K)$ rules. If $x \in C_m$, then in G_m'' either $\sigma \to x$ or $\sigma \Rightarrow \alpha \Rightarrow x$ for some α. Thus, for any m, G_m'' derives at most $f_1^2(K)$ words of the length less or equal to $2m$ and therefore

* $|x|$ denotes the length of the word x and $\varphi_m^1(x) = \varphi_m(x)$, $\varphi_m^{i+1}(x) = \varphi_m(\varphi_m^i(x))$ for $i \geqq 1$.

for sufficiently large m $C_m \not\subset L(G''_m)$ what contradicts the construction of G''_m. Hence, the K with the assumed property cannot exist and thereby the theorem is proved.

On the other hand a quite different situation is with the criterion Symb and, as the following theorem indicates, with respect to this criterion ε-rules do not simplify the description of CFL's too much.

Theorem 4. $\mathrm{Symb}_\varepsilon (L - \{\varepsilon\}) \leqq 10\, \mathrm{Symb}\, (L)$ *for any CFL L.*

Proof. Let $G = \langle V, \Sigma, P, \sigma \rangle$ be a minimal grammar for L with respect to the criterion Symb. Let $E = \{A : A \in V - \Sigma, A \overset{*}{\Rightarrow} \varepsilon\}$. Let us remove all ε-rules from G and let G' be the resulting grammar. In the next step each rule $A \to \alpha$ of G' will be replaced by a set $\varphi(A \to \alpha)$ of new rules and the resulting grammar will be termed G''. The sets $\varphi(A \to \alpha)$ are determined as follows

(i) Let α have no occurence of a symbol in E. Then

$$\varphi(A \to \alpha) = \{A \to \alpha\} .$$

(ii) Let α contain a symbol in E and also a symbol not in E. In this case α can be expressed in the form

(†) $\alpha = u_1 \alpha_1 u_2 \alpha_2 \dots u_k \alpha_k u_{k+1}$ where $k \geqq 1$, $u_i \in (V - E)^*$ if $1 \leqq i \leqq k + 1$, $u_j \neq \varepsilon$ if $2 \leqq j \leqq k$ and, moreover, for $1 \leqq i \leqq k$, α_i has one of the forms aF_1, $F_1 a$ or $F_1 a F_2$ where $a \in V - E$ and F_1, F_2 are in E^*.

(The decomposition (†) is not unique but it does not matter in what follows.) On the base of the decomposition (†), the set $\varphi(A \to \alpha)$ is determined as follows. It contains

(1) A rule $A \to u_1 A_1 u_2 A_2 \dots A_k u_{k+1}$ where A_i are new distinct symbols not in V and not used in the construction of $\varphi(B \to \beta)$ for other rules $B \to \beta$ in P;

(2) For each i, $1 \leqq i \leqq k$, the set S_i of rules which is determined as follows:

If $\alpha_i = aF$, $F = F_1 \dots F_l$, $F_k \in E$, $1 \leqq k \leqq l$, then S_i contains the following rules ($R_1, \dots, R_{l-1}$ are again new distinct symbols not used outside the set S_i)

$$\begin{aligned}
&R_1 \to a,\ R_1 \to aF_1\,,\\
&R_2 \to R_1,\ R_2 \to R_1 F_2\,.\\
&\vdots\\
&R_{l-1} \to R_{l-2},\ R_{l-1} \to R_{l-2} F_{l-1}\\
&A_i \to R_{l-1},\ A_i \to R_{l-1} F_l\,.
\end{aligned}$$

If $\alpha_i = Fa$ or $\alpha_i = FaF'$, F, F' are in E^*, then the set S_i is constructed in a similar way. In any case it holds

$$\mathrm{Symb}\,\{\varphi(A \to \alpha)\} \leqq 7\, \mathrm{Symb}\,(A \to \alpha)$$

(iii) Let $\alpha \in E^*$ and $\alpha = E_1E_2 \ldots E_k$. In this case the set $\varphi(A \to \alpha)$ will contain the rules

$$A \to E_1,\ A \to E_1R_2,\ A \to R_2\,,$$
$$R_2 \to E_2,\ R_2 \to E_2R_3,\ R_2 \to R_3$$
$$R_{k-1} \to E_{k-1},\ R_{k-1} \to E_{k-1}E_k,\ R_{k-1} \to E_k$$

where again $R_2, \ldots, R_{k-1}$ are new nonterminals not used in other parts of the construction of the sets S_i. From this construction it follows immediately that

$$\text{Symb}\,\{\varphi(A \to \alpha)\} \leqq 10 \quad \text{Symb}\,(A \to \alpha)\,.$$

Summarizing (i) to (ii) we get the inequality Symb $(G'') \leqq 10$ Symb (G). However, the grammar G'' generates the language $L(G) - \{\varepsilon\}$ as it is easy to see from the above constructions and therefore Symb ε $(L - \{\varepsilon\}) \leqq 10$ Symb (L) completing the proof.

Corollary. The use of ε-rules can essentially decrease the number of rules but not too much the total number of symbols in the rules.

Example. In order to illustrate the above technique of removing of ε rules, let us consider the grammar with two rules $\sigma \to \sigma a_1\sigma \ldots \sigma a_n\sigma$, $\sigma \to \varepsilon$ with $a_1, \ldots, a_n$ being distinct symbols. The use of standard technique for removing of ε-rules yields a grammar with 2^{n+1} rules. On the other hand the use of the technique of the preceding theorem results in a grammar with $2n + 1$ rules.

5. UNDECIDABILITY

By Theorem 1 to a given CFG G one can effectively construct an ε-free grammar generating the language $L(G) - \{\varepsilon\}$. Can we effectively find the simplest grammar with this property? The two theorems of this section show that the answer is negative if Prod and Symb are considered to be the criteria of complexity.

Theorem 5. *There is no effective method to construct to a given CFG G, an ε-free CFG G' generating the language $L(G) - \{\varepsilon\}$ and such that* Prod $(G') =$ $=$ Prod$_\varepsilon$ $(L(G) - \{\varepsilon\})$.

Proof. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be arbitrary n-tuples of nonempty words over the alphabet $\{a, b\}$. By [1], one can effectively construct, given x and y, a CFG generating the language

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

where

$$L(x) = \{ba^{i_k} \ldots ba^{i_1}cx_{i_1} \ldots x_{i_k}\,;\quad 1 \leqq i_j \leqq n, 1 \leqq j \leqq k\}$$
$$L(y) = \{ba^{i_k} \ldots ba^{i_1}cy_{i_1} \ldots y_{i_k}\,;\quad 1 \leqq i_j \leqq n, 1 \leqq j \leqq k\}\,.$$

For any x and y $(\{a, b, c\} \cup \{a, b, c\} . \{a, b, c\}) \subset L_{x,y}$ and therefore any ε-free grammar generating $L_{x,y}$ has to contain the rules

$$(\dagger) \qquad A \to a\,, \quad B \to b\,, \quad C \to c$$

for some nonterminals A, B, C and at least one nonterminal rule. Hence $\text{Prod}_\varepsilon (L_{x,y}) \geqq 4$. If $L(x) \cap L(y) = \emptyset$, then the language $L_{x,y}$ is generated by the grammar with the rules

$$(\dagger\dagger) \qquad \sigma \to \sigma\sigma\,, \quad \sigma \to a\,, \quad \sigma \to b, \quad \sigma \to c$$

and therefore $\text{Prod}_\varepsilon (L_{x,y}) = 4$. Let us assume that $L(x) \cap L(y) \neq \emptyset$ and that there exists a grammar G' for $L_{x,y}$ with four rules. Since G' has to have three rules of the form (†) and all two-symbol words in $\{a, b, c\}^*$ are also in $L_{x,y}$, the fourth rule of G' would have to be $\delta \to \delta\delta$ but then G' does not generate $L_{x,y}$. Hence $\text{Prod}_\varepsilon (L_{x,y}) > 4$ if $L(x) \cap L(y) \neq \emptyset$. Now the theorem follows from the undecidability of Post's correspondence problem.

Let us now return once more to the foregoing proof. If $L(x) \cap L(y) = \emptyset$, then (††) implies $\text{Symb}_\varepsilon (L_{x,y}) \leqq 13$. If $L(x) \cap L(y) \neq \emptyset$, then $\text{Prod}_\varepsilon (L_{x,y}) > 4$ and therefore $\text{Symb}_\varepsilon (L_{x,y}) \geqq 15$. Hence we can again apply Post's correspondence theorem and derive our last result.

Theorem 6. *There is no effective way to construct to a given CFG G, ε-fre CFG G' such that $L(G') = L(G) - \{\varepsilon\}$ and* $\text{Symb}_\varepsilon (L(G) - \{\varepsilon\}) = \text{Symb}(G')$.

Remark. For every n let G_n be a CFG with the rules $\sigma \to (E_1E_2)^n$, $E_1 \to E_1a_1|\varepsilon$, $E_2 \to E_2a_2|\,\varepsilon$. Then $\text{Symb}(G_n) = 2n + 14$. Let G''_n be a grammar constructed from G_n using the technique of Theorem 4. Then $\text{Symb}(G''_n) = 20n + 4$, $L(G'_n) = L(G_n) - \{\varepsilon\}$ and for every $\varrho > 0$, $\text{Symb}(G''_n) > (10 - \varrho)\,\text{Symb}(G_n)$ for sufficiently large n. On the other hand the open problem is whether there exists a $K < 10$ such that $\text{Symb}_\varepsilon ((L) - \{\varepsilon\}) \leqq K\, \text{Symb}(L)$ for any CFL L.

(Received June 28, 1971.)

REFERENCES

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

[2] J. Gruska: Some classifications of context-free languages. Information and Control *14* (1969), 2, 152–179.

[3] J. Gruska: A few remarks on the index of context-free grammars and languages. Information and Control *19* (1971), 3, 216–223.

[4] J. Gruska: On the size of context-free grammars. Kybernetika *8* (1972) 3, 213–218.

RNDr Jozef Gruska, CSc.; Výskumné výpočtové stredisko, program OSN (Computing Research Centre, United Nations D.P.), Dúbravska 7, 865 31 Bratislava. Czechoslovakia.