数理解析研究所講究録
910 巻 1995 年 1-4

On Subwords of Languages[1]

Pál DÖMÖSI
Institute of Mathematics and Informatics
L. Kossuth University
Egyetem tér 1
4032 Debrecen
Hungary
domosi@math.klte.hu

and

Masami ITO
Faculty of Science
Kyoto Sangyo University
Kyoto 603
Japan
ito@ksuvx0.kyoto-su.ac.jp

Abstract: For any (formal) language L, we consider the language $Sub(L)$ of all subwords of elements in L and define the function $f_L : N \to N$ having the possibly minimal complexity such that $p \in Sub(L)$ implies $qpr \in L$ for some pair q, r of words with $| qr | \leq f_L(| p |)$ (where $| p |$ denotes the length of p). We show that, for any regular language L, there exists a constant f_L of this type. Moreover, if L is context-free, then it can be found a linear f_L. Using well-known results, we give an example for a context-sensitive language L having only non-recursive f_L.

[1]This work was supported by grants of the Soros Foundation, the Hungarian National Foundation for Scientific Research (OTKA T4295,T7608), Kyoto Sangyo University, and Grant-in-Aid for Scientific Research (No. 06640092), Ministry of Education, Science and Culture of Japan.

1. Introduction

For all notions and notations not defined here, see [1 - 3]. An *alphabet* is a finite nonempty set. The elements of an alphabet are called *letters.* A *word* over an alphabet X is a finite string consisting of letters of X. For any alphabet X, let X^* denote the *free monoid* generated by X, i.e. the set of all words over X including the empty word λ and $X^+ = X^* \setminus \{\lambda\}$. The *length* of a word w, in symbols $| w |$, means the number of letters in w when each letter is counted as many times as it occurs. By definition, $| \lambda | = 0$. If u and v are words over an alphabet X, then their *catenation* uv is also a word over X. Especially, for any word uvw, we say that v is a *subword* of uvw. A *language* over X is a set $L \subseteq X^*$. We extend the concept of catenation for the class of languages as usual. Therefore, if L_1 and L_2 are languages, then $L_1L_2 = \{p_1p_2 \mid p_1 \in L_1, p_2 \in L_2\}$. Let p be a word. We put $p^0 = \lambda$ and $p^n = p^{n-1}p (n > 0)$. Thus $p^k (k \geq 0)$ is the *k-th power* of p. If there is no danger of confusion, then sometimes we identify p with the singleton set $\{p\}$. Thus we will write p^* and p^+ instead of $\{p\}^*$ and $\{p\}^+$, respectively. The set of all subwords of any word p is denoted by $Sub(p)$. For any language L, we put $Sub(L) = \cup\{Sub(p) \mid p \in L\}$. L is *dense* if $Sub(L) = X^*$. A *generative grammar* is an ordered quadruple $G = (V_N, V_T, S, P)$ where V_N and V_T are disjoint alphabets, $S \in V_N$, and P is a finite set of ordered pairs (W, Z) such that Z is a word over the alphabet $V = V_N \cup V_T$ and W is a word over V containing at least one letter of V_N. The elements of V_N are called *nonterminals* and those of V_T *terminals.* S is called the *start symbol.* Elements (W, Z) of P are called *productions* and are written $W \to Z$. A word Q over V *derives directly* a word R, in symbols, $Q \Rightarrow R$, if and only if there are words Q_1, Q_2, Q_3, R_1 such that $Q = Q_2Q_1Q_3, R = Q_2R_1Q_3$ and $Q_1 \to R_1$ belongs to P. Q *derives* R, or in symbols, $Q \Rightarrow *R$ if and only if there is a finite sequence of words $W_0, \ldots, W_k (k \geq 0)$ over V where $W_0 = Q, W_k = R$ and $W_i \Rightarrow W_{i+1}$ for $0 \leq i \leq k - 1$. Thus for every $W \in (V_N \cup V_T)^*$ we have $W \Rightarrow *W$. The *language* $L(G)$ *generated by* G is defined by $L(G) = \{w \mid w \in V_T^*, S \Rightarrow *w\}$.

2. Results

Suppose that G is *regular.* Then each production is one of the forms $W \to wZ$ or $W \to w$ where $W, Z \in V_N$ and $w \in V_T^*$. It is obvious that for any $p \in Sub(L(G))$, there exists a derivation $W_1 \Rightarrow q_1W_2 \Rightarrow \ldots \Rightarrow q_1 \ldots q_iW_{i+1} \Rightarrow q_1 \ldots q_ip_1W_{i+2} \Rightarrow \ldots \ldots \Rightarrow q_1 \ldots q_ip_1 \ldots p_mW_{i+m+1}$ with $W_1, \ldots, W_{i+m} \in V_N, W_1 = S, W_{i+m+1} \in V_N \cup \{\lambda\}$, and $p = p_1 \ldots p_m$, such that the word $W_1 \ldots W_{i+1}$ has no letters with double occurrences. Clearly, then $i <| V_N |$. On the other hand, we may suppose without loss of generality that there exists a positive integer t such that every nonterminal W has a derivation $W \Rightarrow *p_W$ with $p_W \in V_T^*$ and $| p_W | \leq t$. We get the following result.

Theorem 2.1. *For any regular language L there exists a positive integer k having the property that $p \in Sub(L)$ implies $qpr \in L$ for some pair q, r of words with $| qr | \leq k$.* □

Now we assume that G is *context-free.* Then every production has the form $W \to Z$, where $W \in V_N$ and $Z \in (V_N \cup V_T)^*$. We may assume without loss of generality that for a suitable positive integer t every nonterminal W has a derivation $W \Rightarrow *p_W$ with $p_W \in V_T{}^*$ and $| p_W | \leq t$.

Denote s the maximal length of the right side of the productions. First we show that for any derivation $A \Rightarrow *q'ar', A \in V_N, a \in V_T$ there exists a pair $q, r \in V_T{}^*$ such that $A \Rightarrow *qar, | qar | \leq (| V_N | s-1)t+1$, moreover, $q = \lambda$ provided $q' = \lambda$ and $r = \lambda$ provided $r' = \lambda$. If $A \Rightarrow *q'ar'$ holds for some pair $q', r' \in (V_N \cup V_T)^*$, then there exist productions $W_i \rightarrow Q_i W_{i+1} R_i, i = 1, \ldots, j, j \geq 1, W_1 = A, W_{j+1} = a$ with $W_1 (= A), \ldots, W_j \in V_N$ such that the word $W_1 \ldots W_j$ has only distinct letters. Then $j \leq | V_N |$. Thus the length of $Q_1 \ldots Q_j a R_j \ldots R_1$ is not greater than $| V_N | s$ and it has not more than $| V_N | s - 1$ nonterminals. Therefore, we can obtain a derivation $A \Rightarrow *qar$ where $qar \in V_T{}^*$ and $| qar | \leq (| V_N | s-1)t+1$. Especially, if $q' = \lambda$ then for any derivation $A \Rightarrow *Q_1 \ldots Q_j a R_j \ldots R_1 \Rightarrow *ar'$ we obtain $Q_1 \ldots Q_j \Rightarrow *\lambda$. Hence we may assume $q = \lambda$ whenever $q' = \lambda$. Similarly, if $r' = \lambda$, then for any derivation $A \Rightarrow *Q_1 \ldots Q_j a R_j \ldots R_1 \Rightarrow *q'a$ we obtain $R_j \ldots R_1 \Rightarrow *\lambda$. Consequently, we may assume $r = \lambda$ whenever $r' = \lambda$.

Let us consider a positive integer $n > 1$. Now we suppose that for any derivation $A \Rightarrow *q'pr', A \in V_N, p \in V_T{}^+, | p | < n$ there exists a pair $q, r \in V_T{}^*$ such that $A \Rightarrow *qpr, | qpr | \leq ((| V_N | s-1)t+1)(2 | p | -1)$, moreover, $q = \lambda$ provided $q' = \lambda$ and $r = \lambda$ provided $r' = \lambda$. Prove that the n-length words preserve these properties. Take an n-length word $p' \in V_T{}^*$ such that $A \Rightarrow *q'p'r'$ holds for some pair $q', r' \in$ $\in (V_N \cup V_T)^*$. Then there exist productions $W_i \rightarrow Q_i W_{i+1} R_i, i = 1, \ldots, j,$ $j \geq 1$ with $W_1 (= A), \ldots, W_j \in V_N$ such that the word $W_1 \ldots W_j$ has only distinct letters. Furthermore, $W_{j+1} = Z_1 \ldots Z_m$ where $Z_1, ..., Z_m \in V_N \cup V_T, m \geq 2$, $| Q_1 \ldots Q_j R_j \ldots R_1 | \leq | V_N | s - 2$. Moreover, $Z_1 \Rightarrow *w_1 p_1$, $Z_m \Rightarrow *p_m w_2$, $| p_1 |, | p_m | > 0, Z_\ell \Rightarrow *p_\ell, \ell = 2, \ldots, m-1, p' = p_1 \ldots p_m$, and $w_1, w_2 \in V_T{}^*$. Of course, using our inductive assumptions, $| w_1 p_1 | \leq ((| V_N | s-1)t+1)(2 | p_1 | -1)$ and $| p_m w_2 | \leq ((| V_N | s-1)t+1)(2 | p_m | -1)$. Then for an appropriate derivation $A \Rightarrow qp'r$ $(q, r \in V_T{}^*)$ we have that $qp'r$ has not more letters than $(| V_N | s-2)t+$ $+ | p_2 | + \ldots + | p_{m-1} | + ((| V_N | s-1)t+1)(2 | p_1 | + 2 | p_m | -2)$ $(m \geq 2)$. Therefore, $| qp'r | < ((| V_N | s-1)t+1)(2n-1)$. On the other hand, for any derivation $A \Rightarrow *Q_1 \ldots Q_j w_1 p' w_2 R_j \ldots R_1 \Rightarrow *p'r'$ we obtain $Q_1 \ldots Q_j w_1 \Rightarrow *\lambda$. Hence we may assume $q = \lambda$ whenever $q' = \lambda$. Similarly, if $r' = \lambda$, then for any derivation $A \Rightarrow *Q_1 \ldots Q_j w_1 p' w_2 R_j \ldots R_1 \Rightarrow q'p'$ we obtain $w_2 R_j \ldots R_1 \Rightarrow *\lambda$. Consequently, we may assume $r = \lambda$ whenever $r' = \lambda$. Therefore, the word p' preserves the properties of our inductive assumptions. Especially, if $A = S$ and $A \Rightarrow *q'p'r'$ with $q'p'r' \in V_T{}^*$, then by definition $p' \in sub(L(G))$. Thus, if k is a positive integer with $k \geq 2(| V_N | s-1)t+1$, then we receive the following result.

Theorem 2.2. *For any context-free language L there exists a positive integer k having the property that $p \in Sub(L)$ implies $qpr \in L$ for some pair q, r of words with $| qr | \leq k | p |$.* □

Finally, it is well-known [2] that, for each recursively enumerable language $L' \subseteq$ $\subseteq X^*$, there is a context-sensitive language $L \subseteq \{a^i b \mid i \geq 0\} X^*$ with $a, b \notin X$ such that for each $p \in L'$ there is a word $a^i bp \in L$, and for each $a^i bp \in L$ we have $p \in L'$.

We may assume, for example, that $L = \{c^n d \mid n \in M\}(c \neq d)$ where M is an arbitrary recursively enumerable but non-recursive subset of positive integers. Let $f_L : N \longrightarrow N$ be a mapping of the set of all positive integers into itself such that for any $p \in Sub(L)$ there exists a pair q, r with $qpr \in L$ and $\mid qr \mid \leq f_L(\mid p \mid)$. If f_L is recursive, then for any positive integer k, we can costruct the language $L_k = \{a^m bc^k d \mid m \leq f_L(k+2)\}$ such that $k \in M$ implies $bc^k d \in Sub(L)$, which leads to $bc^k d \in Sub(L_k)$ and $L \cap L_k \neq \emptyset$. (Observe that $bc^k d \in Sub(a^i bc^j d), i, j \geq 0$ if and only if $k = j$. Hence $m \leq f_L(k+2)$ for some $a^m bc^k d \in L$ provided $bc^k d \in Sub(L)$.) Conversely, if $L \cap L_k \neq \emptyset$ then $bc^k d \in Sub(L)$, which results $k \in M$. But L is context-sensitive, thus it is recursive [2]. Then it can be decidable whether $L \cap L_k$ is empty. Therefore, M is recursive, a contradiction. This means that f_L is non-recursive. Thus we have the following statement.

Theorem 2.3. *Let L be a language and $f_L : N \longrightarrow N$ be a function such that for any $p \in Sub(L)$ there exists a pair q, r with $qpr \in L$ and $\mid qr \mid \leq f_L(\mid p \mid)$. There exists a context-sensitive language which has no recursive function f_L having this property.* □

We close our paper with some examples which show that we can not extend our results in general.

Example 2.1. Consider the language $L = \{a^n b^n \mid n \geq 1\} \cup bX^*(X = \{a, b\})$. It satisfies the conditions of Theorem 2.1 with $k = 1$ but it is inherently context-free. Therefore, the converse of Theorem 2.1 does not hold.

Example 2.2. $L = \{a^n b^n c^n \mid n \geq 1\}$ satisfies the conditions of Theorem 2.2 with $k = 2$. And it is well-known that L is inherently context-sensitive. (More precisely, it is inherently indexed.) Thus the converse of Theorem 2.2 is invalid.

Example 2.3. For any positive integer k define the language $L = \{a^{k|p|} p \mid p \in X^*\}$ $(X = \{a, b\})$. It is clear that for any positive integer n, $a^{kn} b^n$ is the shortest word in $L(G)$ which contains b^n as subword. Thus, for any positive integer n, there exists an n-length word $p \in Sub(L)$ such that $qpr \in L$ implies $\mid qr \mid \geq k \mid p \mid$. It is easy to prove that L is context-free. (Actually L is a linear dense language.) Consequently, we can not extend our Theorem 2.1 for the class of context-free languages.

References

1. A.V.Aho, Indexed Grammars - An Extension of Context-Free Grammars, *Journ. of ACM*,**15** (1968), pp.647-671.
2. A. Salomaa, *Formal Languages*, Academic Press, New York, London,1973.
3. H.J. Shyr, *Free Monoids and Languages*, National Chung-Hsing University, Taichung, Taiwan, R.O.C., Ho Min Book Company,1991.