Publ. Math. Debrecen
56 / 3-4 (2000), 313–320

Superwords and context-free languages

By PÁL DÖMÖSI (Debrecen)

To the honour of Professor Kálmán Győry on his 60^{th} birthday

Abstract. An improvement of a polynomial algorithm is given to solve the following problem: Given a context-free language L and a finite list of nonempty words $w_1, \ldots, w_n$, let us decide whether or not there are words $z_0, \ldots w_n$ having $z_0 w_1 z_1 \ldots \ldots w_n z_n \in L$.

1. Preliminaries

We will consider the following problem. Given a finite ordered list of nonempty words $w_1, \ldots, w_n \in X^*$, a context-free language $L \subseteq X^*$, let us decide whether or not there is a word $z \in L$ such that $z = z_0 w_1 z_1 \ldots w_n z_n$ for some words $z_0, \ldots, z_n \in X^*$. In other words, given a regular language of the form $R = X^* w_1 X^* w_2 \ldots X^* w_n X^*$, $\lambda \notin \{w_1, \ldots, w_n\}$ and a context-free language L, let us decide whether or not $R \cap L$ is empty or not. We note that R is defined as the shuffle ideal generated by $w_1, \ldots, w_n \in X^*$. It is also said that $z = z_0 w_1 z_1 \ldots w_n z_n$ is a *superword* of the list of words $w_1, \ldots, w_n$. In this explanation, we want to decide whether or not $w_1, \ldots, w_n$ has a superword in the context-free language L.

To the solution of the above problem a polynomial algorithm is described in [1]. In this paper we show results leading to an improvement of this algorithm.

Mathematics Subject Classification: 68Q45.
Key words and phrases: formal languages, combinatorics on words and languages.
This work was supported by the Hungarian National Science Foundation (Grant No.'s T019392 and T030140).

For all notions and notations not defined here, see [4] and [5], [6]. Consider an *alphabet* X and the *free monoid* X^* generated by X. λ denotes the *identity* of X^*, $X^+ = X^* \setminus \{\lambda\}$, and $|p|$ is the *length* of $p \in X^*$. In addition, we will denote by $|H|$ the cardinality of a given set H. Finally, we shall consider a grammar in the form $G = (V, X, S, P)$, where, in order, V and X are the sets of *variables* and *terminals*, respectively, S denotes the *start symbol*, and P is the set of *productions*. Moreover, $L(G)$ denotes the *language generated by* G. We will assume $|V| > 1$ taking out of consideration trivial cases.

Theorem 1.1 [3]. *For any context-free grammar $G = (V, X, S, P)$ in Chomsky normal form and $z \in L(G)$, if $|z| \geq |V|2^{|V|}\ e$ $(e > 0)$, and e positions of z are excluded, then z has the form $uvwxy$ where $|vx| > 0$, neither v nor x contains any excluded position, and uv^iwx^iy is in $L(G)$ for all $i \geq 0$.*

By this statement, the following result is shown in [1].

Theorem 1.2 [1]. *Consider a context-free grammar $G = (V, X, S, P)$ in Chomsky normal form and a word $z_0w_1z_1 \dots w_nz_n \in L(G)$ with $\lambda \notin \{w_1, \dots, w_n\}$. There are words $z'_0, \dots, z'_n$ such that $z'_0w_1z'_1 \dots w_nz'_n \in L(G)$ and $|z'_0w_1z'_1 \dots w_nz'_n| < |V|2^{|V|}|w_1 \dots w_n|$.*

2. Main results

For any word $z \in X^*$, and positive integer $k \leq |z|$, we will speak about *k^{th} position of z*. That is, if $z = a_1 \dots a_n$, $a_1, \dots a_n \in X$, then we say that *a_k is in the k^{th} position of z*. In addition, sometimes we will distinguish *excluded* and *non-excluded positions* of z. Finally, if $a_k, \dots, a_{k+\ell}$ are in excluded positions of z then we also say that *$a_k \dots a_{k+\ell}$ consists of excluded positions*.

Given a context-free grammar G in Chomsky normal form, let T_z be a derivation tree for some $z \in L(G)$. We say that a subpath of T_z is *external* if its initial node is the root of the tree and its terminal node is either the first or the last position of z. In the same sense, we will speak about the external subpaths of a given subtree of T_z. An intermediate node of T_z is said to be a *branch point* if each of its children has an excluded descendant. On the other hand, define a node to be *free* if each of its children has no excluded descendant. (Recall that G is in Chomsky normal form. Thus

every node in T_z has not more than two children.) Of course, the leaves of T_z are neither branch points nor free nodes.

A subpath of T_z is called *distinguished* if

a) its initial node is either a branch point or the root of the tree, and its terminal node is a branch point;

b) none of its intermediate nodes is a branch point;

c) if it has no intermediate node then its initial node is the root of the tree, and simultaneously, it is not a branch point.

(Of course, it is also possible that the root of the tree is a branch point.) A subpath of T_z is said to be *reducible* if

a) each of its nodes is not a leaf of the tree;

b) apart from the terminal one, its nodes are not branch points;

c) its terminal node is a branch point;

d) there are distinct nodes having the same (nonterminal) label.

Given a context-free grammar G in Chomsky normal form, a word $z \in L(G)$, let T_z be a derivation tree with $z = z_0 w_1 z_1 \dots w_n z_n$, where $w_1, \dots, w_n$ denote (possibly empty) words consists of excluded positions, and $z_1, \dots, z_n$ denote (possibly empty) words having no excluded positions. T_z is *reducible* (with respect to $z_0, w_1, z_1, \dots, w_n, z_n$) if it has a reducible subpath. Otherwise we say that T_z is *minimal*.

We start with the following

Lemma 2.1. *If T_z is a reducible derivation tree with respect to $z_0, w_1, z_1, \dots, w_n, z_n$ then there are words $z'_0, \dots, z'_n$ with $|z'_0 \dots z'_n| < |z_0 \dots z_n|$ and $z'_0 w_1 z'_1 \dots w_n z'_n \in L(G)$.*

Proof. Suppose that T_z is reducible and let denote p one of its reducible subpaths. Thus, there exist distinct nodes in p having the same (nonterminal) label, say A, moreover, two strings of terminals, v and x, and two nonterminals B and C such that the derivation $A \Rightarrow BC \Rightarrow *vAx$ is represented in T_z. B and C cannot both dominate the lower A, therefore $|vx| > 0$. On the other hand, since there exists no intermediate branch point of the distinguished paths, we have that neither v nor x contains an excluded position. (Of course, the free children of the nodes of this path do not have excluded descendants.) Therefore, there are positive integers i, j with $0 \leq i < j \leq n$, $z_i = v'vv''$, $z_j = x'xx''$ having $z_0 w_1 z_1 \dots w_i v' v'' w_{i+1} z_{i+1} \dots w_j x' x'' w_{j+1} z_{j+1} \dots w_n z_n \in L(G)$. This completes the proof. □

Lemma 2.2. *Let T_z be a minimal derivation tree and consider its arbitrary distinguished subpath p. The free children of the intermediate nodes in p have not more than $2^{|V|-1}-1$ non-excluded descendants.*

PROOF. Consider a subpath p' containing all nodes of p apart from its initial node if the initial node of p is a branch point. Otherwise, (when the initial point of p is not a brainch point and then it is the root of the tree) let us assume $p'=p$. Since T_z is minimal, p' is not reducible. Consider the maximal derivation subtree $T_{z'}$ of T_z having the root as the initial node of p'. Omitting all of the descendants of the terminal node of p' (and p) from $T_{z'}$, we get a subtree $T_{z''}$ containing no path with distinct nodes having the same nonterminal label. Therefore, the subtree $T_{z''}$ has not more than $2^{|V|-1}$ leaves, where one of the leaves is the terminal node of p' (and p). The proof is complete. □

Lemma 2.3. *Let k be the number of the words in $w_1,\dots,w_n$ consisting of two or more letters. Suppose that T_z is a minimal derivation tree. Then T_z has not more than $|w_1\dots w_n|+n+k-1$ distinguished paths.*

PROOF. Denote by $w_{i_1},\dots,w_{i_k}$, $1\le i_1<\dots<i_k\le n$ the words with $|w_{i_j}|>1, j=1,\dots,k$.

First we assume that $|w_1|,\dots,|w_n|\le 2$. Clearly, then $|w_1\dots w_n|=n+k$. On the other side, the number of distinguished paths in T_z is either at most $2(n+k)-2$ if the root of T_z is a branch point or at most $2(n+k)-1$ if the root of T_z is not a branch point. Therefore, there exist not more than $|w_1\dots w_n|+n+k-1$ distinguished paths.

Now we suppose that our statement holds for every $w_1,\dots,w_n$ with $k\le\ell$, where ℓ is a fixed nonnegative integer. Denote by s_i, $i=1,\dots,n$ the number of the distinguished paths which are subpaths of the external paths of the subtree T_{w_i}. Consider a decomposition $z=z'_0w'_1z'_1\dots z'_{n-2}w'_{n-1}z'_{n-1}$ with $z'_0=z_0$, $w'_1=w_1$, $z'_1=z_1$, …, $w'_{i-1}=w_{i-1}$, $z'_{i-1}=z_{i-1}$, $w'_i=w_iz_iw_{i+1}$, $z'_i=z_{i+1}$, $w'_{i+1}=w_{i+2}$, $z'_{i+1}=z_{i+2}$, …, $w'_{n-1}=w_n$, $z'_{n-1}=z_n$ for some $1\le i<n$. Of course, if T_z is minimal with respect to $z_0,w_1,z_1,\dots,w_n,z_n$ then it is also minimal with respect to $z'_0,w'_1,z'_1,\dots,w'_{n-1},z'_{n-1}$. Thus it is enough to prove that in this case the number of distinguished paths is not more than $|w'_1\dots w'_{n-1}|+n+k-1$ $(=|w'_1\dots w'_{n-1}|+(n-1)+(k+1)-1)$. Denote by t_i the number of all distinguished paths which are subpaths of the external paths of the derivation subtree $T_{w'_i}$. It is clear that the new distinguished paths in T_z (with

respect to $z'_0, w'_1, z'_1, \ldots, w'_{n-1}, z'_{n-1}$) are subpaths of the external paths of the derivation subtree $T_{w'_i}$ and thus t_i is bounded by $s_i + s_{i+1} + |z_i|$. Therefore, using our inductive assumption, the number of distinguished paths is not more than $|w'_1 \ldots w'_{n-1}| + n + k - 1$. This ends the proof. □

Now we show an improvement of Theorem 1.2 in [1] (assuming $|V|>1$).

Theorem 2.4. *Given a context-free grammar $G = (V, X, S, P)$ in Chomsky normal form and a word $z_0 w_1 z_1 \ldots w_n z_n \in L(G)$ with $\lambda \notin \{w_1, \ldots, w_n\}$, let k be the number of the words in $w_1, \ldots, w_n$ consisting of two or more letters. There are words $z'_0, \ldots, z'_n$ such that $z'_0 w_1 z'_1 \ldots w_n z'_n \in L(G)$ and $|z'_0 w_1 z'_1 \ldots w_n z'_n| \le 2^{|V|-1}(|w_1 \ldots w_n| + n + k - 1) - n - k + 1$.*

Proof. Consider a derivation tree T_z of the word $z{=}z'_0 w_1 z'_1 \ldots w_n z'_n$. Exclude positions in z such that $w_1, \ldots, w_n$ are (possibly empty) words consisting of excluded positions, and $z'_0, \ldots, z'_n$ are (possibly empty) words having non-excluded positions. Using Lemma 2.1, we may assume that T_z is a minimal derivation tree. Then, by Lemma 2.3, T_z has not more than $|w_1 \ldots w_n| + n + k - 1$ distinguished paths. On the other hand, using Lemma 2.2, for every distinguished subpath p of T_z, the free children of the intermediate nodes in p have not more than $2^{|V|-1} - 1$ excluded descendants. In addition, it is obvious that each of the excluded positions is a descendant of an unambigously determined free child of a node of a given distinguished path in T_z. (Note that the root of T_z may have free children unless it is a branch point. In the other cases, only the intermediate nodes of the distinguished paths may have free children.) Therefore, $|z'_0 \ldots z'_n| \le (2^{|V|-1} - 1)(|w_1 \ldots w_n| + n + k - 1)$ which implies $|z| \le |w_1 \ldots w_n| + (2^{|V|-1} - 1)(|w_1 \ldots w_n| + n + k - 1) = 2^{|V|-1}(|w_1 \ldots w_n| + n + k - 1) - n - k + 1$. □

Given a context-free grammar $G = (V, X, S, P)$ in Chomsky normal form, let us construct the grammar $G' = (V', X', S', P')$ in Chomsky normal form such that $V' = V \cup \{\Omega\}$, $X' = X \cup \{\omega\}$, $P' = P \cup \{A \to A\Omega, A \to \Omega A : A \in V\} \cup \{\Omega \to \Omega\Omega, \Omega \to \omega\}$ and $S' = S$, where Ω and ω are new nonterminal and terminal symbols, respectively. The following statement is obvious.

Lemma 2.5 [1]**.**

$$L(G') = \{\omega^* x_1 \omega^* x_2 \ldots \omega^* x_n \omega^* : x_1 \ldots x_n \in L(G), x_1, \ldots, x_n \in X\}.$$

Next we prove a little bit modified version of a result in [1].

Theorem 2.6. *Given a context-free grammar $G = (V, X, S, P)$ in Chomsky normal form, let $w_1, \dots w_n$ be an arbitrary list of nonempty words. Moreover, let k be the number of the words in $w_1, \dots, w_n$ consisting of two or more letters. There are words $z_0, \dots, z_n$ with $z_0 w_1 z_1 \dots w_n z_n \in L(G)$ if and only if $z'_0 w_1 z'_1 \dots w_n z'_n \in L(G')$ holds for some words $z'_0, \dots, z'_n \in X'^*$ $(= (X \cup \{\omega\})^*)$ with $|z'_i| = (2^{|V|-1} - 1)(|w_1 \dots w_n| + n + k - 1)$, $i = 0, \dots, n$.*

Proof. First we suppose $z'_0 w_1 z'_1 \dots w_n z'_n \in L(G')$ for some words $z'_0, \dots, z'_n \in X'^*$ $(= (X \cup \{\omega\})^*)$ with $|z'_i| = (2^{|V|-1} - 1)(|w_1 \dots w_n| + n + k - 1)$. Consider the words z_i, $i = 0, \dots, n$ such that for every $i = 0, \dots, n$, we omit all occurrences of the letter ω in z'_i. By Lemma 2.5, $z_0 w_1 z_1 \dots w_n z_n \in L(G)$.

Conversely, we now suppose $z_0 w_1 z_1 \dots w_n z_n \in L(G)$ for some $z_0, \dots$ $\dots, z_n$. By Theorem 2.4, we may assume $|z_0 w_1 z_1 \dots w_n z_n| \leq 2^{|V|-1}(|w_1 \dots$ $\dots w_n| + n + k - 1) - n - k + 1$. In other words, $|z_0 z_1 \dots z_n| \leq (2^{|V|-1} - 1)(|w_1 \dots w_n| + n + k - 1)$, which implies $|z_i| \leq (2^{|V|-1} - 1)(|w_1 \dots w_n| + n + k - 1)$, $i = 0, \dots, n$. Hence, for every $i = 0, \dots, n$, we can define $z'_i = z_i \omega^{(2^{|V|-1}-1)(|w_1 \dots w_n| + n + k - 1) - |z_i|})$ such that, by Lemma 2.5, our conditions hold again. The proof is complete. □

3. Generalized CYK algorithm

Using the well-known CYK-algorithm (Cocke–Younger–Kasami algorithm), on the basis of Theorem 2.6, a cubic time algorithm can be given to the discussed problem. Apart from the value of the constant p, this algorithm is the same as in [1] for which, to the completeness of the paper, we give a short description. (For a more detailed description, see [1].)

Let $G = (V, X, S, P)$ be a context-free grammar having Chomsky normal form and let $w_1, \dots, w_n \in X^+$. Consider the grammar $G' = (V', X', S', P')$ such that $V' = V \cup \{\Omega\}$, $X' = X \cup \{\omega\}$, $P' = P \cup \{A \to A\Omega, A \to \Omega A : A \in V\} \cup \{\Omega \to \Omega\Omega, \Omega \to \omega\}$, where Ω and ω are new nonterminal and terminal symbols, respectively. Let k be the number of the words in $w_1, \dots, w_n$ consisting of two or more letters.

Now we give the formal description of our algorithm.

begin

$p := \omega^{(|2^{|V'|-1}-1)(|w_1 \ldots w_n|+n+k-1)} w_1 \omega^{(|2^{|V'|-1}-1)(|w_1 \ldots w_n|+n+k-1)} w_2 \ldots$
$\omega^{(|2^{|V'|-1}-1)(|w_1 \ldots w_n|+n+k-1)} w_n \ \omega^{(|2^{|V'|-1}-1)(|w_1 \ldots w_n|+n+k-1)}$;

for $i := 1$ **to** $|p|$ **do**
if the i^{th} symbol of p is ω then **do**
$V_{i,1} := V \cup \{\Omega\}$;
else do
$V_{i,1} := \{A \mid \exists a \in X$ such that $A \to a$ is a production in G
and the i^{th} symbol in p equals to $a\}$;
for $j := 2$ **to** $|p|$ **do**
for $i := 1$ to $|p| - j + 1$ **do**
begin
$V_{i,j} := \emptyset$;
for $k := 1$ **to** $j - 1$ **do**
$V_{i,j} := V_{i,j} \cup \{A \mid A \to BC$ is a production in G', B is in $V_{i,k}$
and C is in $V_{i+k,j-k}\}$;
end

$z_0 w_1 z_1 \ldots z_n w_n \in L(G)$ for some $z_0, \ldots, z_n \in X^*$ iff $S \in V_{1,|p|}$.

end

Finally, we consider the following example.

Example. Consider the grammar $G = (V, X, S, H')$ with $V = \{\mathcal{A}, \mathcal{B}, \mathcal{D}, \mathcal{H}, \mathcal{I}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{S}, \mathcal{T}, \mathcal{U}, \mathcal{Y}, \Diamond\}$, $X = \{A, B, D, H, I, O, P, R, T, U, Y, \Box\}$, $H' = \{\mathcal{A} \to \mathcal{AP}$, $\mathcal{A} \to \mathcal{AY}$, $\mathcal{B} \to \mathcal{BI}$, $\mathcal{D} \to \mathcal{DA}$, $\mathcal{H} \to \mathcal{HD}$, $\mathcal{I} \to \mathcal{IR}$, $\mathcal{O} \to \mathcal{OU}$, $\mathcal{O} \to \mathcal{O}\Diamond$, $\mathcal{P} \to \mathcal{PP}$, $\mathcal{P} \to \mathcal{PY}$, $\mathcal{R} \to \mathcal{RT}$, $\mathcal{S} \to \mathcal{HA}$, $\mathcal{T} \to \mathcal{TH}$, $\mathcal{T} \to \mathcal{TO}$, $\mathcal{Y} \to \mathcal{YO}$, $\mathcal{Y} \to \mathcal{Y}\Diamond$, $\Diamond \to \Diamond\mathcal{B}$, $\Diamond \to \Diamond\mathcal{T}$, $\Diamond \to \Diamond\mathcal{Y}$, $\mathcal{A} \to A$, $\mathcal{B} \to B$, $\mathcal{D} \to D$, $\mathcal{H} \to H$, $\mathcal{I} \to I$, $\mathcal{O} \to O$, $\mathcal{P} \to P$, $\mathcal{R} \to R$, $\mathcal{T} \to T$, $\mathcal{U} \to U$, $\mathcal{Y} \to Y$, $\Diamond \to \Box\}$, words HA, R, D, TO, Y. Using our algorithm, (by a long computation) we get the existence of words α, β, γ, δ, ε, ζ with $\alpha HA\beta R\gamma D\delta TO\varepsilon Y\zeta \in L(G)$. For example, let $\alpha = \lambda$, $\beta = PPY\Box BI$, $\gamma = TH$, $\delta = AY\Box$, $\varepsilon = \Box$, $\zeta = OU$. Then the following superword of HA, R, D, TO, Y is obtained.

HAPPY □ BIRTHDAY □ TO □ YOU

References

[1] P. Dömösi and M. Ito, Multiple keyword pattern in context-free languages, in: Words, sequences, grammars, languages: where biology, computer science, linguistics and mathematics meet I. (C. Martin-Vide and V. Mitrana, eds.), *Kluwer*, *Dordrecht*, 2000, (*in print*).

[2] P. Dömösi and M. Ito, Characterization of languages by length of their subwords, Proc. Int. Conf. in Semigroups and it Applications, Semigroups (K. P. Shum, Y. Gao, M. Ito and Y. Fong, eds.), *Springer-Verlag*, 1998, 117–129.

[3] P. Dömösi, M. Ito, M. Katsura and C. Nehaniv, A new pumping property of context-free languages, Proc. DMTCS'96, Combinatorics, Complexity, & Logic, *Springer-Verlag*, 1997, 187–193.

[4] J. E. Hopcroft and J. D. Ullman, Introduction to Automata Theory, Languages, and Computation, *Addison-Wesley*, *Reading , Massachusetts, Menlo Park, California, London, Amsterdam, Don Mils, Ontario, Sidney*, 1979.

[5] Gy. E. Révész, Introduction to Formal Languages, *McGraw-Hill*, *New York, etc.*, 1983.

[6] A. Salomaa, Formal Languages, *Academic Press*, *New York, London*, 1973.

pál dömösi
institute of mathematics and informatics
lajos kossuth university
debrecen, egyetem tér 1, h–4032
hungary

E-mail: domosi@ math.klte.hu

(Received August 16, 1999; revised February 2, 2000)