Theoretical Computer Science 163 (1996) 277–281

Theoretical Computer Science

Note

A shrinking lemma for indexed languages

Robert H. Gilman [1]

Department of Mathematics, Stevens Institute of Technology, Hoboken, NJ 07030, USA

Received September 1995
Communicated by M. Nivat

Abstract

This article presents a lemma in the spirit of the pumping lemma for indexed languages but easier to employ.

1. Introduction

The pumping lemma for context-free languages has been extended to stack languages [5] and indexed languages [3], but these generalizations are rather complicated. In this article we take a slightly different approach by concentrating only on that part of the context-free pumping lemma which says that if $uvwxy \in L$, then $uwy \in L$, and by employing a theorem on divisibility of words which is not used in [5] or [3]. Our result, Theorem A, is relatively easy to state and strong enough to verify the examples given in [3] of languages which are not indexed. On the other hand, it does not afford a proof that the finiteness problem for indexed languages is solvable as does [3, Theorem 5.1].

Indexed languages were introduced by Aho [1,2]. A brief introduction appears in [4, Ch. 14]. Our original motivation for Theorem 1 was the investigation of finitely generated groups for which the language of words defining the identity is indexed.

2. A result on indexed languages

Before stating our result we fix some notation. Σ is a finite alphabet, $|w|$ is the length of $w \in \Sigma^*$, and for each $a \in \Sigma$, $|w|_a$ is the number of a's in w.

[1] The author was partially supported by NSF grant DMS-9401090.

S0304-3975/96/$15.00 © 1996—Elsevier Science B.V. All rights reserved
PII S0304-3975(96)00244-8

Theorem A. *Let L be an indexed language over Σ and m a positive integer. There is a constant $k > 0$ such that each word $w \in L$ with $|w| \geqslant k$ can be written as a product $w = w_1 \cdots w_r$ for which the following conditions hold.*

(1) $m < r \leqslant k$.

(2) *The factors w_i are nonempty words.*

(3) *Each choice of m factors is included in a proper subproduct which lies in L.*

By (3) we mean that the chosen factors occur in a product $w_{i_1} \cdots w_{i_t} \in L$ with $1 \leqslant i_1 < \cdots < i_t \leqslant r$ and $m \leqslant t < r$. The proof of Theorem A is given in the next section.

Corollary 1. *Let L be an indexed language. There is a constant $k > 0$ such that if $w \in L$ and $|w| > k$, then there exists $v \in L$ with $(1/k)|w| \leqslant |v| < |w|$.*

Proof. Take $m = 1$ in Theorem A and choose a factor of maximum length. $\square$

By taking m to be the number of letters in Σ and arguing similarly we obtain a result on the Parikh mapping.

Corollary 2. *Let L be an indexed language over Σ. There is a constant $k > 0$ such that if $w \in L$ and $|w| > k$, then there exists $v \in L$ with $(1/k)|w|_a \leqslant |v|_a \leqslant |w|_a$ for each $a \in \Sigma$ and $|v|_a < |w|_a$ for some $a \in \Sigma$.*

Corollary 1 has the following immediate consequence.

Corollary 3 [3, Theorem 5.2]. *If f is a strictly increasing function on the positive integers, and $L = \{a^{f(n)}\}$ is an indexed language, then $f = O(k^n)$ for some positive integer k.*

Corollary 4 [3, Theorem 5.3]. *The language $L = \{(ab^n)^n \mid n \geqslant 1\}$ is not indexed.*

Proof. Suppose L is indexed, and apply Theorem A to L with $m = 1$. Pick $w = (a^n b)^n$ with $n > k$ and consider the decomposition $w = w_1 \cdots w_r$. As $r \leqslant k$, at least one factor w_i must contain two or more a's. Choose that w_i to be in the proper subproduct v. But then v contains a subword $ab^n a$, which is impossible as $v \neq w$. $\square$

3. Proof

The proof of Theorem A depends on a result about divisibility of words. We say that v divides w and write $v \prec w$ if v is a subsequence of w. For example $ac \prec abc$. By a theorem of Higman [6, Theorem 6.1.2] every set of words defined over a finite alphabet and pairwise incomparable with respect to divisibility is finite. We will use this result in the following form.

Lemma 1. *Let m be a positive integer and Y a language over a finite alphabet Δ. Y contains a finite subset X with the property that for any $y \in Y - X$ with m letters distinguished there is an $x \in X$ such that $x \precneqq y$ and x includes all the distinguished letters of y.*

Proof. Let Δ' be the union of Δ with m pairwise disjoint copies of itself, and define Y' be the language of all words over Δ' which project to Y and contain exactly one letter from each of the m copies of Δ. By Higman's theorem X', the set of all words in Y' each of which is not divisible by any word in Y' except itself, is finite. For any $y' \in Y'$ if we take x' to be a word of minimum length among all words in Y' dividing y', then $x' \in X'$. Further x' contains all the letters of y' from $\Delta' - \Delta$.

Define X to be the union of the projection of X' to Δ^* with the set of all words in Y of length less than m. Suppose that $y \in Y - X$ has m distinguished letters. Since $|y| \geqslant m$, we can pick $y' \in Y'$ projecting to y so that the distinguished letters of y correspond to the letters of y' in $\Delta' - \Delta$. By the preceding paragraph y' is divisible by an $x' \in X'$ which contains those letters. It follows that the projection of x' to Σ^* is the desired word x. $\square$

Notice that x might be a subsequence of y in more than one way. Lemma 1 asserts only that there is some subsequence of y which includes the distinguished letters and whose product is x.

Fix an indexed language L over Σ, and let G be an indexed grammar for L. Let G have sentence symbol S, nonterminals N, and indices F. $(NF^* + \Sigma)^*$ is the set of sentential forms. By [1, Theorem 4.5] we may assume G is in normal form, i.e.,

(1) S does not appear on the righthand side on any production;

(2) There are no ε-productions except perhaps $S \to \varepsilon$;

(3) Each production has one of the forms $A \to BC, Af \to B, A \to Bf$, or $A \to a$, where $A, B, C \in N$, $f \in F$, and $a \in \Sigma$.

We are using the definition of indexed grammar from [6]; this definition is slightly different from the original.

We write $\alpha \xrightarrow{*} \beta$ to indicate that the sentential form β can be derived from the sentential form α via productions of G, and we use $\beta \cdot \omega$ to denote the sentential form obtained by appending the index string ω to the index string of every nonterminal in the sentential form β. It follows from the way derivations are defined in indexed grammars that if $\alpha \xrightarrow{*} \beta$, then $\alpha \cdot \omega \xrightarrow{*} \beta \cdot \omega$. Conversely if $\alpha \cdot \omega \xrightarrow{*} \beta \cdot \omega$ and if every nonterminal occurring in that derivation has an index string with suffix ω, then $\alpha \xrightarrow{*} \beta$.

Lemma 2. *Let m be a positive integer and $A\omega$ a sentential form in NF^*. There is a finite set of sentential forms $X \subset (N + \Sigma)^*$ with the property that if $A\omega \xrightarrow{*} \beta \in (N + \Sigma)^* - X$, and m symbols of β are distinguished, then there is $\alpha \in X$ such that $A\omega \xrightarrow{*} \alpha \precneqq \beta$, and α includes all the distinguished symbols of β.*

Proof. Apply Lemma 1 to the language of all sentential forms in $(N+\Sigma)^*$ derivable from $A\omega$. $\square$

Consider a derivation $S \xrightarrow{*} w \in L$, and let Γ be the corresponding derivation tree. Let each vertex p of Γ have label $\lambda(p)$, and define a subtree $\Gamma(p)$ with root p as follows. If $\lambda(p)$ is a terminal or nonterminal, then $\Gamma(p)$ consists of p and all its descendants. Otherwise $\lambda(p) = Af\omega$ for some nonterminal A, index f, and string of indices ω. In this case along each path emanating from p there will be a first vertex, perhaps a leaf of Γ, at which f is consumed. Define $\Gamma(p)$ to be the union of all the paths from p up to and including these first vertices. The subtrees $\Gamma(p)$ play an important role in [3]; we will use them here in a slightly different way than they are used there.

Let $\gamma(p)$ be the sentential form obtained by concatenating the labels of the leaves of $\Gamma(p)$ in order; if p is a leaf, $\gamma(p) = \lambda(p)$. Since $\Gamma(p)$ is a subtree of a derivation tree, $\lambda(p) \xrightarrow{*} \gamma(p)$. If $\lambda(p) = Af\omega$, then by construction all vertices of $\Gamma(p)$ except its leaves have labels of the form $B\omega' f\omega$. The leaves are labelled by terminals or labels form $B\omega$. Deleting all the suffixes ω yields a derivation tree for $Af \xrightarrow{*} \beta(p)$ where $\gamma(p) = \beta(p)\cdot\omega$. Extend the definition of $\beta(p)$ to all other vertices p of Γ by defining $\beta(p) = \gamma(p)$ when $\lambda(p)$ is a terminal or nonterminal.

It follows from Lemma 2 that there is a finite set of sentential forms $Z \subset (N+\Sigma)^*$ such that for any of the finitely many sentential forms $A\omega \in N \cup NF$ if $A\omega \xrightarrow{*} \beta \in (N+\Sigma)^* - Z$ and m symbols of β are distinguished, then there is $\alpha \in Z$ such that $A\omega \xrightarrow{*} \alpha \precneqq \beta$, and α includes all the distinguished symbols of β. Since it does no harm to enlarge Z, we may assume Z contains all elements of $(N+\Sigma)^*$ of length at most m.

Lemma 3. *Let $C \geqslant 2$ be an upperbound for the lengths of elements of Z. Suppose $\beta(p) \notin Z$ but $\beta(q) \in Z$ for all vertices q which are proper descendants of p, then $|\beta(p)| \leqslant C^2$.*

Proof. If p is a leaf, then $|\beta(p)| = 1$. Suppose p has two descendants, q_1, q_2. It follows from the normal form for G that $\beta(p) = \beta(q_1)\beta(q_2)$, and consequently $|\beta(p)| \leqslant 2C$. Finally, if p has a single descendant, q, then the derivation $\lambda(p) \xrightarrow{*} \gamma(p)$ begins with application of a production of the form $A \to a$, $Af \to B$ or $A \to Bf$. In the first case $|\beta(p)| = |a| = 1$. In the second case $\lambda(p)$ must be $Af\omega$ whence $\beta(p) = B$ and again $|\beta(p)| = 1$.

Consider the last case. We have $\lambda(p) = A\omega$ and $\lambda(q) = Bf\omega$. Further $\beta(p)$ is the product of the terms $\beta(q')$ as q' ranges over the leaves of $\Gamma(q)$. Since $\beta(q) \in Z$, there are at most C terms; and as each $\beta(q') \in Z$, we have $|\beta(p)| \leqslant C^2$. $\square$

To complete the proof of Theorem A choose $k = C^2 + 2$ and suppose $S \xrightarrow{*} w \in L$ with $|w| \geqslant k$. Let Γ be the corresponding derivation tree and p_0 its root. Clearly, $\beta(p_0) = w \notin Z$, and so we may choose p to satisfy the hypothesis of Lemma 3. Note that $\beta(p) \notin Z$ implies $|\beta(p)| > m$; in particular p is not a leaf.

If $\lambda(p) = A$, then $\beta(p) = a_1 \cdots a_t$ is a subword of w and $m < t \leqslant C^2$. Consequently, $w = w'a_1 \cdots a_t w''$ exhibits w as a product of more than m and at most k nonempty factors. Suppose m of the factors in this product are distinguished. If not all these factors are letters a_i, distinguish more letters to bring the total of distinguished letters a_i to m. By definition of Z there is a word $u \in Z$ such that $A \xrightarrow{*} u \precneqq a_1 \cdots a_t$ and u contains all the distinguished letters of $a_1 \cdots a_t$. It follows that $v = w'uw''$ contains the distinguished factors of w and satisfies all the conditions of Theorem A.

Finally, $\lambda(p) = Af\omega$ implies $\beta(p) = z_1 \cdots z_t$ with $m < t \leqslant C^2$ and each $z_i \in N \cup \Sigma$. Further $\gamma(p) = \beta(p) \cdot \omega$. Consequently, $w = w'u_1 \cdots u_t w''$ where each u_i is the subword derived from $z_i \cdot \omega$ in the derivation $S \xrightarrow{*} w$. Because G is in normal form, none of the u_i's is the empty word. As before there exists $\alpha \in Z$ such that $Af \xrightarrow{*} \alpha \precneqq \beta(p)$ and α contains all the z_i's for which u_i is distinguished. We have $\alpha \cdot \omega \xrightarrow{*} u$ where u is the subproduct of $u_1 \cdots u_t$ corresponding to the z_i's in α. It follows that $v = w'uw''$ satisfies the conditions of Theorem A.

References

[1] A. Aho, Indexed grammars – an extension of context-free grammars, *J. ACM* **15** (1968) 647–671.

[2] A. Aho, Nested stack automata, *J. ACM* **16** (1969) 383–406.

[3] T. Hayashi, On derivation trees of indexed grammars: an extension of the uvwxy-theorem, *Publ. RIMS Kyoto Univ.* **9** (1973) 61–92.

[4] J.E. Hopcroft and J. D. Ullman, Introduction to Automata Theory, Languages, and Computation (Addison-Wesley, Reading, MA, 1979).

[5] W. Ogden, Intercalation theorems for stack languages, *Proc. 1st Annual ACM Symp. on the Theory of Computing* (1969) 31–42.

[6] J. Sakarovitch and I. Simon, Subwords, in: M. Lothaire, ed., *Combinatorics on Words*, Encyclopedia of Mathematics and Its Applications, Vol. 17 (Addison Wesley, Reading, MA, 1983) 105–142.