Inverse Star, Borders, and Palstars

Narad Rampersad
Department of Mathematics
University of Liège
Grande Traverse, 12 (Bat. B37)
4000 Liège
Belgium
narad.rampersad@gmail.com

Jeffrey Shallit
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
Canada
shallit@cs.uwaterloo.ca

Ming-wei Wang
Microsoft Corporation
Redmond, WA
USA
m2wang@gmail.com

September 1, 2010

Abstract

A language L is closed if $L = L^*$. We consider an operation on closed languages, L^{-*}, that is an inverse to Kleene closure. It is known that if L is closed and regular, then L^{-*} is also regular. We show that the analogous result fails to hold for the context-free languages. Along the way we find a new relationship between the unbordered words and the prime palstars of Knuth, Morris, and Pratt. We use this relationship to enumerate the prime palstars, and we prove that neither the language of all unbordered words nor the language of all prime palstars is context-free.

1 Inverse star

Let L be a language such that $L = L^*$. Then, following [3], we say that L is *closed*. Brzozowski [2] studied the the "smallest" language M such that $L = M^*$.

Definition 1. For closed languages L, define

$$L^{-*} = \bigcap_{S^*=L} S.$$

Brzozowski proved

Theorem 2. *If L is closed then $(L^{-*})^* = L$. Furthermore $L^{-*} = L - L^2$. If L is regular and closed, then so is L^{-*}.*

In this note we show that the class of context-free languages is not closed under the operation $-*$. First, though, we take a digression to discuss products of palindromes.

2 Palstars, prime palstars, and unbordered words

In this section we find a new connection between the prime palstars (as introduced in Knuth, Morris, and Pratt [4]) and the unbordered words.

We start with some definitions. By w^R we mean the reverse of the word w. A *palindrome* is a word w such that $w = w^R$. In this paper we will only be concerned with the nonempty palindromes of even length:

$$\texttt{PAL} = \{xx^R \ : \ x \in \Sigma^+\}.$$

A *palstar* is an element of the language $\texttt{PALSTAR} := \texttt{PAL}^*$.

A word x is a *prime palstar* if it is a palstar and cannot be written as the product of two palstars. Evidently a prime palstar must itself be a palindrome. The first few prime palstars over $\{0, 1\}$ are $00, 0110, 010010, 011110, 01000010, 01011010, 01111110$, and their complements, obtained by mapping 0 to 1 and vice versa. The language of all prime palstars is denoted $\texttt{PRIMEPALSTAR}$.

Theorem 3 (Knuth-Morris-Pratt [4])**.** *Every palstar has a unique factorization into prime palstars.*

The proof of this theorem depends on the following lemma:

Lemma 4 (Knuth-Morris-Pratt [4])**.** *No prime palstar is a proper prefix of another prime palstar.*

Corollary 5. *If w is a palindrome of even length, then its factorization into prime palstars must be of the form $w = x_1 x_2 \cdots x_n$, where $x_i = x_{n+1-i}$ for $1 \leq i \leq n$.*

Proof. Suppose $w = x_1 \cdots x_n$ is the factorization into prime palstars x_i. If $n = 1$ we are done. Otherwise, since w ends with x_n, it must begin with $x_n^R = x_n$. Hence either x_1 is a prefix of x_n, or vice versa. By Lemma 4 we must have $x_1 = x_n$. Using the same argument on the shorter palindrome ${x_1}^{-1} w x_1^{-1}$, we derive the remaining equalities. □

We now turn to borders. A word is said to be *bordered* if it has some nonempty prefix that is also a suffix. Otherwise, it is *unbordered*. Unbordered words are also called *bifix-free* in the literature [5].

Equivalently, a word w is bordered if it can be written in the form xyx for some nonempty word x. For example, `entanglement` begins and ends with the string `ent`.

Given two words of the same length $x = a_1 a_2 \cdots a_n$ and $y = b_1 b_2 \cdots b_n$, their *perfect shuffle* $x Ш y$ is defined by $x Ш y = a_1 b_1 \cdots a_n b_n$.

Theorem 6. *A word w is a prime palstar if and only if there exists an unbordered word z such that $w = z Ш z^R$.*

Proof. Suppose w is not a prime palstar. If w is not an even length palindrome then it is certainly not of the form $z Ш z^r$. Suppose then that w is an even length palindrome and hence is of the form $z Ш z^R$. We will show that z is bordered. Since w is not a prime palstar we can factor w into a product of prime palstars. Then by Corollary 5 such a factorization must look like $x \cdots x$ for some palindrome x. Then when we "unshuffle" w into z and z^R, we get that z starts with the odd-indexed letters of x and ends with the odd-indexed letters of x^R. But $x = x^R$, so z starts and ends with the same word.

On the other hand, suppose $w = x Ш y$. By comparing the symbols x to y we see that if $y \neq x^R$, then w is not a palindrome. So assume $y = x^R$. Now if x is bordered, then we can write it as $x = zuz$ for some nonempty string z. Then $w = (zuz) Ш (zuz)^R = (z Ш z^R)(u Ш u^R)(z Ш z^R)$. This gives a factorization of w as a product of two or three nonempty palstars (according to whether u is empty or nonempty). □

An example of this theorem in English is `noon`, which is a prime palstar, and is the shuffle of the unbordered word `no` with its reversal.

3 Enumeration of palstars

As far as we know, up to now no one has enumerated the palstars. However, our argument above allows us to do so, based on enumeration of the unbordered words.

Nielsen [5] has shown that if a_n denotes the number of unbordered words of length n over an alphabet of size k, then

$$a_n = \begin{cases} k, & \text{if } n = 1; \\ ka_{n-1} - a_{n/2}, & \text{if } n \text{ even}; \\ ka_{n-1}, & \text{if } n \text{ odd and } > 1. \end{cases}$$

(Also see [1].) Furthermore, he showed that $a_n \sim c_k k^n$, where c_k is a constant that tends to 1 as $k \to \infty$, and $c_2 \doteq .2677868$.

It follows that if b_n is the number of prime palstars of length $2n$, then $b_n = a_n$. In particular, about 27% of all binary palindromes are prime palstars.

4 Context-free languages and inverse star

We now apply the results in Section 2 to prove that the class of context-free languages is not closed under inverse star.

Clearly `PALSTAR` = `PAL`* is context-free. We have `PRIMEPALSTAR` = `PALSTAR`$^{-*}$. So it suffices to show that `PRIMEPALSTAR` is not context-free. Suppose it were. First, we need the following result.

Theorem 7. *The language U of unbordered words over an alphabet of size at least 2 is not context-free.*

Proof. Assume it is. Without loss of generality the alphabet is $\Sigma = \{0, 1, \ldots\}$. Consider

$$U' := U \cap 1\ 0^+\ 1\ 0^+\ 1\ 0^+\ 1\ 0^+,$$

the intersection of U with a regular language. Then

$$U' := \{1\ 0^a\ 1\ 0^b\ 1\ 0^c\ 1\ 0^d\ :\ (a < d) \text{ and } ((a \neq c) \text{ or } (b < d))\}.$$

Since the context-free languages are closed under intersection with a regular language, it suffices to prove U' is not context-free.

To do this, we use Ogden's lemma [6]. Choose

$$z = \overbrace{10^{n+n!}}^{A}\ \overbrace{10^{n+1+n!}}^{B}\ \overbrace{10^{n}}^{C}\ \overbrace{10^{n+1+n!}}^{D} \in U',$$

and distinguish the third block of 0's, the one corresponding to C. Write $z = uvwxy$. Then by Ogden's lemma vwx must contain at most n distinguished positions and vx at least one.

If vx contains a 1, then by pumping we get a string with too many 1's. Thus vx contains 0's only, and each of v, x is contained in a single block of zeros.

Case 1: v contains 0's from block A, and x contains 0's from block C. Then consider $uv^2wx^2y = 1\ 0^{a'}\ 1\ 0^{b'}\ 1\ 0^{c'}\ 1\ 0^{d'}$. It has $a' \geq d'$, a contradiction.

Case 2: v contains 0's from block B, and x contains 0's from block C. Then consider $uv^iwx^iy = 1\ 0^{a'}\ 1\ 0^{b'}\ 1\ 0^{c'}\ 1\ 0^{d'}$, where $i = (n!/|x|) + 1$. Then this string has $a' = c'$, $b' \geq d'$, a contradiction.

Case 3: vx contains 0's from block C. Then as in the previous case, choose $i = (n!/|vx|) + 1$. The resulting string has $a' = c'$ and $b' \geq d'$, a contradiction.

Case 4: v contains 0's from block C, and x contains 0's from block D. Consider $uv^iwx^iy = 1\ 0^{a'}\ 1\ 0^{b'}\ 1\ 0^{c'}\ 1\ 0^{d'}$ with $i = 0$ to get $a' \geq d'$, a contradiction. □

Now, using this result, we can prove our last result:

Theorem 8. *Over an alphabet of two or more letters,* `PRIMEPALSTAR` *is not context-free.*

Proof. Consider the morphisms g and h defined as follows: $g(a) = 00$, $g(b) = 01$, $g(c) = 10$, $g(d) = 11$, and $h(a) = h(b) = 0$, $h(c) = h(d) = 1$. Then the effect of $h \circ g^{-1}$ is to extract the odd-indexed letters from an even-length word.

Assume that `PRIMEPALSTAR` is context-free. Then $h(g^{-1}(\texttt{PRIMEPALSTAR}))$ would be context-free. But by Theorem 6 $h(g^{-1}(\texttt{PRIMEPALSTAR})) = U$, the language of unbordered words, which we have shown in Theorem 7 to be non-context-free. □

References

[1] G. Blom. Problem 94-20. *SIAM Review* **36** (1994), 657. Solution by O. P. Lossers, **37** (1995), 619–620.

[2] J. Brzozowski. Roots of star events. *J. ACM* **14** (1967), 466–477.

[3] J. Brzozowski, E. Grant, and J. Shallit. Closures in formal languages and Kuratowski's theorem. In V. Diekert and D. Nowotka, editors, *Developments in Language Theory, 13th International Conference, DLT 2009*, Vol. 5583 of *Lecture Notes in Computer Science*, pp. 125–144. Springer-Verlag, 2009.

[4] D. E. Knuth, J. Morris, and V. Pratt. Fast pattern matching in strings. *SIAM J. Comput.* **6** (1977), 323–350.

[5] P. T. Nielsen. A note on bifix-free sequences. *IEEE Trans. Inform. Theory* **IT-19** (1973), 704–706.

[6] W. Ogden. A helpful result for proving inherent ambiguity. *Math. Systems Theory* **2** (1968), 191–194.