INFORMATIQUE THÉORIQUE ET APPLICATIONS

A. EHRENFEUCHT
G. ROZENBERG

Each regular code is included in a maximal regular code

Informatique théorique et applications, tome 20, nº 1 (1986), p. 89-96
<http://www.numdam.org/item?id=ITA_1986__20_1_89_0>

Informatique théorique et Applications/Théoretical Informatics and Applications
(vol. 20, n° 1, 1985, p. 89 à 96)
rairma$407

EACH REGULAR CODE IS INCLUDED IN A MAXIMAL REGULAR CODE (*)

by A. EHRENFEUCHT ([1]) and G. ROZENBERG ([2])

Abstract. — *It is proved that each regular code is included in a maximal regular code. A corollary of this result settles an open question from* [R].

Résumé. — *On prouve que tout code rationnel est contenu dans un code rationnel maximal. Un corollaire de ce résultat répond à une question ouverte posée dans* [R].

INTRODUCTION

A language $C\ \Sigma^+$ is called a *code* if C^* is a free submonoid of Σ^* with base C. The theory of codes initiated by M. Schutzenberger [Sch] forms an interesting fragment of formal language theory. A code $C \subseteq \Sigma^+$ is called *maximal* if, for any $x \in \Sigma^* - C$, $C \cup \{x\}$ is not a code. All codes are subsets of maximal codes and the investigation of maximal codes forms an active research area within the theory of codes (*see*, e. g., [BPS], [P1], [R] and [SM]). In particular one is often interested in the problem of the following kind: given a code C of type X (e. g. finite or regular) is it possible to find a maximal code D of type X such that $C \subseteq D$?

It was shown in [R] that for finite codes this question gets a negative answer. Since then the following question remained open: is every finite code included in a maximal regular code? Obviously any finite (resp. regular) prefix code is included in a finite (resp. regular) maximal prefix code. Recently it was shown in [P2] that every *finite biprefix* code is included in a maximal biprefix regular code.

In this paper we provide a positive answer to the above question. As a matter of fact we prove a more general result (theorem 5): each *regular* code is included in a regular maximal code. We would like to emphasize the

(*) Received May 1984, revised April 1985.
([1]) Department of Computer Science, University of Colorado, Boulder, Colorado 80309.
([2]) Institute of Applied Mathematics and Computer Science, University of Leiden, Leiden, The Netherlands.

following: *the new result presented in this paper is theorem* 5; most of the other results is in one form or the other (and perhaps in a different terminology) retrievable from the literature. However we have decided to make this paper rather self-contained and to provide all the needed results with their (sometimes different from the literature) proofs carried out in a "uniform manner".

We assume the reader to be familiar with basic formal language theory — in particular with rudimentary theory of regular languages (*see*, e. g., [S]).

PRELIMINARIES

We use mostly standard language theoretic notation and terminology.

For a set A, $\# A$ denotes the cardinality of A.

For sets A, B, $A-B$ denotes the set theoretic difference of A and B.

For a word x, $|x|$ denotes its length and *first* (x) denotes the first letter of x; if $x=x_1 y x_2$ then y is called a *subword* of x (also referred to as a *segment* or a *factor* of x). The set of all subwords of x is denoted by $\mathrm{sub}(x)$ and for a language K, $\mathrm{sub}(K)=\bigcup_{x \in K} \mathrm{sub}(x)$.

A nonempty word x is called *bordered* if $x=yzy$ for a nonempty word y; otherwise x is called *unbordered.*

A language $C \subseteq \Sigma^+$ is called a *code* if every word $y \in C^+$ satisfies the following condition:

if $y=u_1 \ldots u_n$ and $y=x_1 \ldots x_m$ for $n, m \geqslant 1$ and $u_1, \ldots, u_n, x_1, \ldots, x_m \in C$ then $n=m$ and $u_i=x_i$ for $1 \leqslant i \leqslant n$. (In other words, y has a unique representation in C; subwords $u_1, \ldots, u_n$ of this representation are referred to as *C-blocks* of y).

A code $C \subseteq \Sigma^+$ is called *maximal* if, for each $x \in \Sigma^* - C$, $C \cup \{x\}$ is not a code.

In the sequel of this paper we consider an arbitrary but fixed alphabet Σ *where* $\sigma = \# \Sigma > 1$; *all languages we will consider are over* Σ.

For a language K and a positive integer n, $L_n(K)=\{w \in K: |w|=n\}$ and $\alpha_n(K)=\# L_n(K)$.

We will define now and recall a number of notions concerning languages — they will be central to our paper.

Let $K \subseteq \Sigma^+$.

(1) K is *dense* if $x \in \mathrm{sub}(K^*)$ for each $x \in \Sigma^*$.

(2) K is *fast* if there exists a positive integer n such that for each $w \in \mathrm{sub}(K^*)$ there exist $x, y \in \Sigma^*$ such that $|xy| \leqslant n$ and $xwy \in K^*$.

(3) K is *rich* if there exists a positive integer e such that $\alpha_m(K^*) \geqslant \sigma^m/e$ for infinitely many positive integers m.

RESULTS

In this section we investigate the problem how various properties of a code (such as: fast, dense, rich, regular and maximal) influence each other. Once this relationship is explored we can settle the problem of completing a regular code to a regular maximal code.

Our first result is known (*see* [SM]). However for the sake of completeness we provide its proof (which is different from the proof in [SM]).

THEOREM 1: *Each maximal code is dense.*

Proof: First we prove the following result.

CLAIM 1: *Let C be a code that is not dense. There exists an unbordered word w_c such that $w_c \notin \mathrm{sub}(C^*)$.*

Proof of Claim 1: Since C is not dense, there exists a word $z \notin \mathrm{sub}(C^*)$. Let $b \in \Sigma$ be such that $b \neq \mathit{first}(z)$ and let $w_c = zb^{|z|}$. Clearly w_c is unbordered. Moreover $w_c \notin \mathrm{sub}(C^*)$, because $z \notin \mathrm{sub}(C^*)$.

Thus claim 1 holds. ■

Now we prove theorem 1 as follows.

Let C be a maximal code.

Assume to the contrary that C is not dense. Then let w_c be an unbordered word satisfying the statement of claim 1.

Consider $D = C \cup \{w_c\}$. Let y be an arbitrary word in D^+. Since w_c is unbordered, y has a unique representation of the form $y = x_0 w_c x_1 w_c \ldots w_c x_n$, where $n \geqslant 0$ (that is if $y = u_0 w_c u_1 w_c \ldots w_c u_m$ where $m \geqslant 0$ *then* $m = n$ and $u_i = x_i$ for $1 \leqslant i \leqslant n$). Since C is a code and $w_c \notin \mathrm{sub}(C^*)$, y has a unique representation in D. Thus D is a code.

Since $C \subseteq D$ and $w_c \notin \mathrm{sub}(C^*)$ we get a contradiction (to the fact that C is maximal).

Consequently C must be dense and theorem 1 holds. ■

THEOREM 2: *Each rich code is maximal.*

Proof: Let C be a rich code and let e be a positive integer constant satisfying the definition of richness for C.

Assume to the contrary that C is not maximal. Let z be a word such that $B=C \cup \{z\}$ is a code; let $|z|=t$.

Let k be a positive integer. Let $n_1, \ldots, n_k$ be a sequence of positive integers such that:

$$n_1 < n_2 < \ldots < n_k \quad \text{and} \quad \alpha_{n_i}(C^*) \geqq \frac{\sigma^{n_i}}{e}. \tag{1}$$

(Since C is rich and e satisfies the definition of richness of C, such a sequence exists.)

Consider $r=n_1+n_2+\ldots+n_k+kt$. Clearly:

$$\alpha_r(B^*) \leqq \sigma^r. \tag{2}$$

On the other hand let us consider an arbitrary permutation $i_1, \ldots, i_k$ of the set $\{1, \ldots, k\}$. Let $y_{i_1} \in L_{n_{i_1}}(C^*), \ldots, y_{i_k} \in L_{n_{i_k}}(C^*)$ and let $\gamma(i_1, \ldots, i_k)=y_{i_1} z y_{i_2} z \ldots y_{i_k} z$. Since B is a code, if $(j_1, \ldots, j_k)$ is a permutation of $\{1, \ldots, k\}$ different from $(i_1, \ldots, i_k)$, then $\gamma(i_1, \ldots, i_k) \neq \gamma(j_1, \ldots, j_k)$. Consequently from (1) it follows that:

$$\frac{\sigma^{n_1}}{e}\frac{\sigma^{n_2}}{e}\ldots\frac{\sigma^{n_k}}{e}k! \leqq \alpha_r(B^*). \tag{3}$$

From (2) and (3) it follows that:

$$k! \leqq e^k \sigma^{tk}=(e\,\sigma^t)^k. \tag{4}$$

Since $e\,\sigma^t$ is a constant (independent of k), there exists a positive integer k_0 such that, for all $s > k_0$, $s! > (e\,\sigma^t)^s$. Consequently (4) yields a contradiction (k was chosen to be an arbitrary positive integer).

Thus C must be maximal and theorem 2 holds. ■

THEOREM 3: *Each regular code is fast.*

Proof: Obvious. ■

THEOREM 4: *Each dense and fast code is rich.*

Proof: Let C be a code that is dense and fast. Then there exists a finite set F of ordered pairs of words from Σ^* such that for each $w \in \Sigma^*$ there exists $(x, y) \in F$ such that $xwy \in C^*$. Let $q=\max\{|xy| : (x, y) \in F\}$, $f=\# F$ and $d=f\,\sigma^q$.

CLAIM 2: *For each positive integer n there exists a positive integer $m \leqq n+q$ such that $\alpha_m(C^*) \geqq \sigma^m/d$.*

Proof of claim 2: Let for each $w \in \Sigma^*$, *pair* (w) be a fixed element (x, y) of F such that $xwy \in C^*$.

Let n be a positive integer. Let:

$$E(n, x, y) = \{w \in L_n(\Sigma^*): pair(w) = (x, y)\}.$$

Clearly for some $(x_0, y_0) \in F$, $\# E(n, x_0, y_0) \geqq \sigma^n/f$. Let $p = |x_0 y_0|$. Then $\alpha_{n+p}(C^*) \geqq \# E(n, x_0, y_0) \geqq \sigma^n/f$.

Hence:

$$\alpha_{n+p}(C^*) \geqq \frac{\sigma^n}{f} = \frac{\sigma^{n+p}}{f\sigma^p} \geqq \frac{\sigma^{n+p}}{f\sigma^q} \geqq \frac{\sigma^{n+p}}{d}.$$

Thus if we choose $m = n+p$ we get $m \leqq n+q$ and claim 2 holds. ■

Now theorem 4 follows directly from claim 2. ■

REMARK: Theorems 2 and 4 together are more general than theorem 7.4 (due to Schutzenberger) from [E]. However, it is pointed out by D. Perrin in [P3] that a proof of the general case can be retrieved from the proof of theorem 9.3 in [E]. ■

THEOREM 5: *Let C be a regular code. There exists a code D which is dense, fast, regular and such that $C \subseteq D$.*

Proof: Let C be a regular code.

We consider separately two cases.

(i) C is dense.

Then the theorem follows from theorem 3 (take $D = C$).

(ii) C is not dense.

Then, by claim 1, there exists an unbordered word w_c such that $w_c \notin \text{sub}(C^*)$.

Let:

$$A = \{w_c x_1 w_c x_2 \ldots w_c x_n w_c : n \geqq 1, x_i \notin C^* \text{ and } w_c \notin \text{sub}(x_i)\}$$

and let $D = C \cup \{w_c\} \cup A$.

CLAIM 3: *D is a code.*

Proof of Claim 3: Let $y \in D^+$. Since w_c is unbordered, y has a unique representation of the form $y = x_1 w_c x_2 w_c \ldots w_c x_n$ (that is we can uniquely distinguish all occurrences of w_c in y).

This representation provides the basis for the division of y into D-blocks which is obtained as follows:

(1) A subword $w_c x_j w_c x_{j+1} \ldots w_c x_{j+l} w_c$ constitutes a D-block (corresponding to A) if $2 \leqq j \leqq n-1$, $j+l \leqq n-1$, $x_j, \ldots, x_{j+l} \notin C^*$ and x_{j-1}, $x_{j+l+1} \in C^*$; such a D-block is referred to as an *A-block*.

(2) All occurrences of w_c not involved in A-blocks are also D-blocks.

(3) All x_i'*s which are not involved in* A-blocks must be in C^* and so they are uniquely divisible in D-blocks (really C-blocks).

The definition of A and the fact that $w_c \notin \mathrm{sub}(C^*)$ and w_c is unbordered guarantee that such a division is unique.

Hence D is a code and claim 3 holds. ■

CLAIM 4: *D is dense.*

Proof of claim 4: Let $u \in \Sigma^*$.

Consider $y = w_c u w_c$. Reasoning as in the proof of claim 3 we get a (unique) representation of y in D^+.

Thus D is dense and claim 4 holds. ■

CLAIM 5: *D is regular.*

Proof: Obvious. ■

CLAIM 6: *D is fast.*

Proof: This follows from claim 5 and theorem 3. ■

Now theorem 5 follows from claims 3 through 5. ■

Our results yield two interesting corollaries. The first one solves an open problem from the theory of codes (*see*, e. g., [R] *and* [P2]). As a matter of fact it provides a more general result: Restivo has asked ([R]) whether an arbitrary *finite* code can be completed to a maximal regular code—we show that even an arbitrary *regular* code can be completed to a maximal regular code.

COROLLARY 1: *Let C be a code. If C is regular, then there exists a code D such that* $C \subseteq D$, *D is maximal and D is regular.*

Proof: Let C be a regular code.

By theorem 5 there exists a regular code D such that $C \subseteq D$, D is fast and dense.

Thus, by theorem 4, D is rich and so, by theorem 2, D is maximal.

Hence corollary 1 holds. ■

Secondly, we notice that theorems 1 through 4 provide an alternative proof of the theorem by Schutzenberger (*see* [E], p. 94).

COROLLARY 2: *Let C be a regular code. Then C is maximal if and only if C is dense.*

Proof: It follows directly from theorems 1 through 4. ■

DISCUSSION

We have established a number of relationships between dense, fast, rich, maximal and regular codes. Using these relationships we were able to demonstrate that each regular code is included in a maximal regular code.

In particular we have demonstrated that each rich code is maximal and each maximal code is dense. Hence each rich code is dense. We provide now a "direct" proof of this result—we believe it sheds a different light on this relationship.

COROLLARY 3: *Each rich code is dense.*

Proof: Let C be a rich code.

Assume that C is not dense. Hence there exists a word $z \notin \mathrm{sub}(C^*)$; let $|z|=t$. Let n be an arbitrary positive integer; n can be represented in the form $n=k_1 t+k_2$ for some $k \geqq 0$ and $k_2 < t$. An arbitrary word from $L_n(C^+)$ can be (starting from the left end) divided into k_1 consecutive subwords of length t leaving a suffix of length k_2. Thus:

$$\alpha_n(C^+) < (\sigma^t - 1)^{k_1} \sigma^{k_2}.$$

Consequently:

$$\frac{\alpha_n(C^+)}{\sigma^n} < \frac{(\sigma^t - 1)^{k_1} \sigma^{k_2}}{\sigma^n} = \frac{(\sigma^t - 1)^{k_1} \sigma^{k_2}}{\sigma^{tk_1} \sigma^{k_2}} = \left(1 - \frac{1}{\sigma^t}\right)^{k_1}.$$

Hence:

$$\lim_{n \to \infty} \frac{\alpha_n(C^+)}{\sigma^n} = 0,$$

which contradicts the fact that C is rich.

Consequently C must be dense and the result holds. ∎

To put some of the dependencies we have demonstrated in a better perspective we provide now the following result.

THEOREM 6: *There exists a maximal code which is not rich.*

Proof: Consider the family of all full binary trees in which leafs are labelled by a and all inner nodes are labelled by b. Consider now all postfix notations for these trees—in this way we get the language $P \subseteq \{a, b\}^+$. It is well known that P is a code (every forest of full binary trees has a unique representation in the postfix notation).

Consider an arbitrary word $z \in \{a, b\}^+ - P$. Clearly $a^{|z|+1} z \in P^+$ (we parse $a^{|z|+1} z$ from right to left assigning $+1$ to a and -1 to b; then each subword yielding by summation weight $+1$ is a tree corresponding to an element of

P). Hence $P \cup \{z\}$ is not a code, because $a^{|z|+1}z$ would have two different representations in P^+. Thus P is a maximal code.

On the other hand it is known (*see*, e. g., [F], ch. III, sect. 3) that:

$$\lim_{n \to \infty} \frac{\alpha_n(P^+)}{2^n} = 0.$$

(Here one considers random walks on the line of positive integers where a represents a "step up" and b represents a "step down". It turns out that the probability of starting in 0 and not returning to 1 in up to n steps equals 1 in the limit.)

Hence P is not rich and the theorem holds. ■

Perhaps the most significant open question in the area of "extending codes to their maximal counterparts" is (*see* [P2]): can every biprefix regular code be extended to a maximal biprefix regular code? An answer to this question will certainly make the picture of the whole area clearer.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NSF grant MCS 79-03838. The authors are indebted to J. Karhumaki and D. Perrin for their comments on the first version of this paper.

REFERENCES

[BPS] J. Berstel, D. Perrin and M. P. Schutzenberger, *The Theory of Codes* (to appear).

[E] S. Eilenberg, *Automata, Languages and Machines*, Vol. A, 1974, Academic Press, New York and London.

[F] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, 1950, J. Wiley.

[P1] D. Perrin Ed., *Theorie des codes*, L.I.T.P. Publication, Paris, 1979.

[P2] D. Perrin, *Completing Biprefix Codes*, Lecture Notes in Computer Science, Vol. 140, 1982, pp. 397-406.

[P3] D. Perrin, *Séries formelles et combinatoire du monoide libre*, in J. Berstel Ed., Series Formelles, L.I.T.P., Paris, 1977.

[R] A. Restivo, *On Codes Having no Finite Completions*, in Automata, Languages and Programming, S. Michaelson Ed., Edinburgh University Press, 1976, pp. 38-44.

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

[Sch] M. P. Schutzenberger, *Une théorie algébrique du codage*, Seminaire Dubreuil-Pisot, année 55-56, exp. No. 15, Inst. Henri-Poincaré, Paris, 1956.

[SM] M. P. Schutzenberger and R. S. Marcus, *Full Decodable Code Word Sets*, I.R.E. Trans on Inf. Theory, Vol. 5, 1959, pp. 13-15.