ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

486

ON COMMUTATIVE LANGUAGES

BY

TIMO LEPISTÖ

HELSINKI 1971
SUOMALAINEN TIEDEAKATEMIA

doi:10.5186/aasfm.1971.486

Communicated 13 November 1970 by ARTO SALOMAA

KESKUSKIRJAPAINO
HELSINKI 1971

On commutative languages

1. Consider the alphabets

$$I_r = \{x_1, x_2, \ldots, x_r\} \quad (r \geqq 1)$$

and

$$A_\omega = \{\alpha, \beta, \gamma, \alpha_1, \beta_1, \gamma_1, \ldots\},$$

where A_ω is an infinite alphabet such that A_ω and each I_r is disjoint. The elements of A_ω are regular expressions, which denote, in the usual way (cf. [2], pp. 1—4), the languages over I_r. As usual we call some equation $\alpha = \beta$ valid if α and β denote the same language, i.e. $|\alpha| = |\beta|$. Let δ_r, $r = 1, 2, \ldots$ denote the set of all schemas of valid equations between regular expressions over A_ω such that a valid equation always results whenever each letter of A_ω appearing in X or Y is substituted by some regular expression over I_r. The interseciion of all sets δ_r is denoted by δ_ω. It is proved (cf. [2], p. 128) that

$$\delta_2 = \delta_3 = \ldots = \delta_\omega \tag{1}$$

and δ_2 is properly included in δ_1.

In the following we consider commutative languages, i.e. we assume that the catenation is commutative. Thus the order of letters in a word does not matter, but only the number of occurrences of each letter. More specifically, let c be the operator defined for languages such that $c(L)$ is the language consisting of all such words which are obtained by permuting the letters in some word belonging to L.

For regular expressions X and Y, the equation $X = Y$ is said to be c-valid if and only if the languages $c(|X|)$ and $c(|Y|)$ are equal. Clearly, all valid equations are c-valid but not vice versa. Denote by C_r, $= 1, 2, \ldots$, the set of equations $X = Y$, where X and Y are regular expressions over A_ω such that whenever the letters of A_ω appearing in X or Y are substituted by some regular expressions over I_r, then the resulting equation is c-valid. We denote by C_ω the intersection of all sets C_r, $r = 1, 2, \ldots$. It is obvious that

$$C_1 \supset C_2 \supset C_3 \supset \ldots \supset C_\omega \tag{2}$$

and

$$\delta_1 = C_1 \quad \delta_r \subset C_r \quad (r = 2, 3, \ldots) .$$

The problem: are the inclusions in (2) proper or not (as in (1)) is presented by SALOMAA (cf. [2], p. 142). In this paper we prove

Theorem. *The inclusions in* (2) *are not proper, i.e.*

$$C_1 = C_2 = \ldots = C_\omega .$$

This problem is independently solved also by LINNA in a recent paper [1], but his proof is essentially different from our proof.

2. Consider the proof of the above theorem. We first show that

$$C_2 = C_3 = \ldots = C_\omega . \tag{3}$$

In order to prove (3) assume the contrary: there is an equation

$$X = Y \tag{4}$$

which belongs to C_2 but not to C_ω. This implies that there is a natural number $r \geqq 3$ such that (4) does not belong to C_r. Hence, there are some regular expressions over I_r such that the equation $X_r = Y_r$ resulting from (4) by substituting these regular expressions for letters of A_ω appearing in X or Y, is not c-valid. Without loss of generality, we may assume that there is a word P over I_r such that $P \in |X_r|$ and there is no word Q in $|Y_r|$ such that the number of the occurrences of x_i in Q is the same as the number of the occurrences of x_i in P for all x_i $(i = 1, 2, \ldots, r)$.

Denote by a_i the number of the letters x_i in the word P $(i = 1, 2, \ldots, r)$ and consider the following function f mapping the set $W(I_r)$ into the set $W(I_2)$ ($W(I_r)$ denotes the set of all words over I_r):

$$f(x_1) = x_1^{p^{r-1}} x_2 ,$$

$$f(x_2) = x_1^{p^{r-2}} x_2 ,$$

$$— — — — — —$$

$$f(x_r) = x_1 x_2 ,$$

$$f(\lambda) = \lambda ,$$

$$f(P'Q') = f(P')f(Q'), \text{ for all } P' , Q' \in W(I_r) ,$$

where the prime p is so chosen that

$$p > a_1 + a_2 + \ldots + a_r . \tag{5}$$

If α is a regular expression over I_r, then α_f is defined to be the regular expression over I_2, obtained from α by replacing each letter x_i by $f(x_i)$, $i = 1, 2, \ldots , r$. Thus, if Q is an arbitrary word belonging to $W(I_r)$, then

(6) $$f(Q) \in |\,\alpha_f\,| \text{ if } Q \in |\,\alpha\,| .$$

Let us denote by b_i the number of the letters x_i in the word $Q(\in W(I_r))$. Consider the system

(7) $$a_1 p^{r-1} + a_2 p^{r-2} + \ldots + a_r = b_1 p^{r-1} + b_2 p^{r-2} + \ldots + b_r ,$$

(8) $$a_1 + a_2 + \ldots + a_r = b_1 + b_2 + \ldots + b_r ,$$

(9) $$a_i \geqq 0 , \; b_i \geqq 0 \;\; (i = 1, 2, \ldots , r) .$$

We can show that this system has only one solution $b_i = a_i$ $(i = 1, 2, \ldots , r)$. Indeed, if the system has another solution, it then follows from (8) that there exist t and k $(t > k)$ such that

$$\begin{cases} a_t \neq b_t , \\ a_k \neq b_k , \\ a_i = b_i \text{ if } i < t \text{ and } i > k . \end{cases}$$

Hence, by (7),

$$p \mid (a_k - b_k) \text{ or } a_k = b_k + \nu p \;\; (\nu \neq 0) .$$

If $\nu > 0$, this yields, by (9),

$$\sum_{i=1}^{r} a_i \geqq a_k \geqq p$$

contradicting (5). On the other hand, if $\nu < 0$, we have again a contradiction with (5), because

$$\sum_{i=1}^{r} a_i = \sum_{i=1}^{r} b_i \geqq b_k \geqq p .$$

We have thus shown that only the words in which the number of the letters x_i $(i = 1, 2, \ldots , r)$ is exactly the same as in P can be mapped by f to the words in which the number of x_i $(i = 1, 2)$ is the same as in $f(P)$. It then follows, by (6),

$$f(P) \in c({}_f|(X_r)|) \text{ and } f(P) \notin c(|(Y_r)_f|) .$$

Hence (4) does not belong to C_2. This is a contradiction. Therefore the equation (3) holds true.

Finally we prove that

(10) $$C_1 = C_2 .$$

The proof is about the same as in the preceding case. However, we must choose the homomorphism f in the different way:

$$f(x_1) = x_1^p\ ,\quad f(x_2) = x_1^q\ ,$$

where the distinct primes p and q are so chosen that $p\,, q > a_1 + a_2$. The above system (7), (8), (9) must now be replaced by the system

$$\left\{\begin{array}{l} a_1 p + a_2 q = b_1 p + b_2 q\ , \\ a_i \geqq 0\ ,\quad b_i \geqq 0\ (i = 1, 2)\ . \end{array}\right. \tag{11}$$

If the system has another solution than $a_1 = b_1$, $a_2 = b_2$, then $a_1 \neq b_1$, $a_2 \neq b_2$ and

$$p \mid (a_2 - b_2)\ ,\quad q \mid (a_1 - b_1)\ .$$

This implies that $b_2 > a_2$ and $b_1 > a_1$, contradicting the system (11). Consequently the system has only one solution $a_i = b_i$ $(i = 1, 2)$ and we can conclude in the same way as in the precedig case that (10) holds true. Our theorem is thus proved.

University of Turku
Technical University of Tampere.

References

[1] Linna, M.: *The set of schemata of c-valid equations between regular expressions is independent of basic alphabet* Ann. Univ. Turku. Series A I

[2] Salomaa, A.: *Theory of Automata* International Series of Monographs in Pure and Applied Mathematics, Vol. 100. Pergamon Press 1969.

Printed January 1971