MATHEMATICS

Al. A. MARKOV

1960

SovietRxiv

Abstract

Full Text

MATHEMATICS

Al. A. MARKOV

ON ALPHABETIC CODING

(Presented by Academician M. V. Keldysh, 16 I 1960)

1. The question is considered of the conditions for the mutual uniqueness of coding of the words of a certain alphabet by replacing each letter with some word of the same or of some other alphabet. It is proved that in the general case the problem reduces to the analogous problem for the case of coding by the same system of words a finite set of words of this alphabet. A necessary and sufficient condition is established for mutual uniqueness, and consequently also for the possibility of reversing the operator that carries out the coding.

2. Let $\mathfrak{A} : \{a_1, \dots, a_m\}$ be some alphabet with base m, and let $S(\mathfrak{A})$ be the set of all words from the letters of this alphabet. To each letter $a_i \in \mathfrak{A}$ let us assign some nonempty word u_i from the letters of the alphabet $\mathfrak{B} : \{b_1, \dots, b_r\}$, which may be the same alphabet, so that $u_i \in S(\mathfrak{B})$. Then to each word $a_{i_1} \dots a_{i_s} \in S(\mathfrak{A})$ there corresponds one and only one word $u_{i_1} \dots u_{i_s} \in S(\mathfrak{B})$. The system of words $\mathfrak{U} : \{u_1, \dots, u_m\}$ generates a mapping of the set $S(\mathfrak{A})$ into the set $S(\mathfrak{B})$, selecting in $S(\mathfrak{B})$ a subset $S_{\mathfrak{A}}(\mathfrak{B})$, to each word of which there corresponds at least one word from $S(\mathfrak{A})$ (in particular, it may be that $S_{\mathfrak{A}}(\mathfrak{B}) = S(\mathfrak{B})$). This mapping, in accordance with the terminology adopted in ([1]), we shall call **alphabetic coding**. Every finite subset $\widetilde{S}(\mathfrak{A})$ of the set $S(\mathfrak{A})$, called a dictionary, by means of the system $\mathfrak{U}$, which we shall call a **coding system of words**, is mapped into a finite subset $\widetilde{S}_{\mathfrak{A}}(\mathfrak{B})$ of the set $S_{\mathfrak{A}}(\mathfrak{B})$, forming the dictionary of codes of the words from $\widetilde{S}(\mathfrak{A})$. The length of a word α will be denoted by $l(\alpha)$; the dictionary consisting of all words of $S(\mathfrak{A})$ whose length does not exceed N, by $S^N(\mathfrak{A})$. Let $l(u_1 u_2 \dots u_m) = n$. Consider all possible relations of the form $u_i \alpha = \beta u_{i_1} \dots u_{i_k}$ such that $0 \leq l(\alpha) < l(u_{i_k})$ and β is a nonempty word, and denote the maximum k over all possible representations of this kind by K.

3. **Theorem 1.** *The alphabetic coding by the system $\mathfrak{U}$ is mutually unique or not simultaneously with the coding of its dictionary $S^N(\mathfrak{A})$, where*

$$N = \frac{(n-m)(K+1)}{2} - \delta(n-m)\frac{K-1}{2},$$

and the function $\delta(p)$ *is defined by the recursive relations* $\delta(0) = 0$, $\delta(p+1) = 1 - \delta(p)$.

Proof. Denote by $\mathfrak{F}$ the set of all words from $S_{\mathfrak{A}}(\mathfrak{B})$ that have more than one preimage in $S(\mathfrak{A})$. A word $\alpha \in \mathfrak{F}$ we shall call **basic** if it has at least two preimages in $S(\mathfrak{A})$, $a_{i_1} \dots a_{i_k}$ and $a_{j_1} \dots a_{j_l}$, such that for no μ and ν, $1 \leq \mu < k$, $1 \leq \nu < l$, does

$$u_{i_1} \dots u_{i_\mu} = u_{j_1} \dots u_{j_\nu}$$

hold. The set of basic words will be denoted by $[\mathfrak{F}]$. It is obvious that every word from $\mathfrak{F}$ can be obtained from basic words and from words having no more than one preimage in $S(\mathfrak{A})$, using only the operation of concatenation (i.e., adjoining one to another in some $\cdots$).

order). We shall construct a special construction, called a **source**, which generates exactly $[\mathfrak{F}]$, under the assumption that $\mathfrak{F}$ is nonempty. An analogous construction is used for describing language in structural linguistics as a grammar with finite states ([2]).

Let $\mathfrak{M}$ be a collection, i.e. an unordered aggregate in which repetitions of objects are allowed, of all ends of words from U (we assume that a word is not an end of itself), containing the empty word Λ as one of its elements. $\mathfrak{M}$ contains $n-m+1$ elements. First of all, note that, if $\mathfrak{F}$ is nonempty, then there is a pair of words u_i and u_j in U such that one of them is the beginning of the other. The set of pairs of words from U possessing this property determines the set $\mathfrak{R} \subset \mathfrak{M}$ of ends of words from U, whose deletion from the corresponding words also gives words from U.

The construction is a finite directed graph ([3]) having $n-m+1$ vertices, to each of which we assign one and only one element of $\mathfrak{M}$, and no two different vertices are assigned the same element. The vertex corresponding to Λ is numbered 1; from 2 to $p+1$ we number the vertices corresponding to the elements of $\mathfrak{R}$ (p is the number of elements in $\mathfrak{R}$); the remaining vertices are numbered arbitrarily from $p+2$ to $n-m+1$. Vertex 1 is joined by edges to each of $2, 3, \dots, p+1$, directed from 1 to i, $2 \leq i \leq p+1$. To each edge $(1, i)$ we assign the word from U which is the beginning of the word determining the element of $\mathfrak{R}$ assigned to i.

Let $\alpha \in \mathfrak{M}$. Among all words $u_{i_1} \cdots u_{i_k}$ for which α is an initial segment, consider those such that $l(u_{i_1} \cdots u_{i_k}) \geq l(\alpha) > l(u_{i_1} \cdots u_{i_{k-1}})$, and denote by $\mathfrak{S}(\alpha)$ the set of those elements of $\mathfrak{M}$ which are the ends of u_{i_k}, obtained by deleting from $u_{i_1} \cdots u_{i_k}$ the first $l(\alpha)$ letters. Let the vertex corresponding to α be j. We join vertex j by edges, directed from j, to all vertices corresponding to elements of $\mathfrak{S}(\alpha)$, and to each we assign the word α. Proceeding in this way with every element of $\mathfrak{M}$ and the vertex corresponding to it, we obtain all the connections between the vertices of the graph—the source with finite states.

We shall say that every sequence of vertices of the form $i_0, i_1, \dots, \dots, i_{k-1}, i_k, 1$, where $i_0 = 1$, $i_l \neq 1$, $l = 1, 2, \dots, k$, and i_l and i_{l+1} are joined by an edge directed to i_{l+1}, $l = 0, 1, \dots, k$, determines a unique word, which is the concatenation of the words assigned to the edges $(1, i_1)$, $(i_1, i_2), \dots, (i_k, 1)$, in the same order in which these edges occur in the sequence. Such a sequence will be called **proper**, and the number k its **length**. If $a_{i_1} \cdots a_{i_s}$ and $a_{j_1} \cdots a_{j_t}$ are two prototypes of a basic word, then it follows from the construction of the source that this word is generated by some proper sequence of vertices.

Conversely, let $i_0, i_1, \dots, i_{k+1}$ be a proper sequence of vertices and let the word generated by it be $\alpha_0\alpha_1 \cdots \alpha_k$, where the word α_ν is generated by the edge $(i_\nu, i_{\nu+1})$. Since $\alpha_1 \in \mathfrak{M}$, there exist two words $u_i, u_j \in U$ such that $\alpha_0 = u_i$, $\alpha_0\alpha_1 = u_j$, and we can construct a prototype in $S(U)$ from the given sequence in two ways: starting with either of the words u_i or u_j; moreover these two prototypes will be different, since $u_i \neq u_j$. In order to prove that the generated word is basic, it remains only to note that the process of constructing a proper sequence, as follows from the definition of the source, can continue at the l-th step only under the condition that the word corresponding to i_l can determine some nonempty end of one of the words U; otherwise $i_{l+1} = 1$, or i_l is not a marked end for any edge, i.e. the process either terminates or cannot be continued. And this means precisely the impossibility of the equality $u_i u_{i_1} \cdots u_{i_s} = u_j u_{j_1} \cdots u_{j_t}$ for any s and t strictly smaller than the lengths of the corresponding prototypes of the generated word, i.e. this word is basic. Thus, the set of words generated by the source we have constructed is $[\mathfrak{F}]$.

If $\mathfrak{F}$ is nonempty, then there is at least one closed contour passing through vertex 1, and consequently there exists a contour without self-intersections passing through this vertex. This means that there exists a basic word generated by a proper sequence of length not greater than $n - m$.

Noting that the edges $(1, i_1), (i_2, i_3), \dots, (i_{2l}, i_{2l+1}), \dots$ carry words, each of which, under transition to a proimage, gives not more than one letter, while the edges $(i_1, i_2), (i_3, i_4), \dots, (i_{2l-1}, i_{2l}), \dots$ carry not more than K letters from $\mathfrak{A}$ for one of the proimages, and the length of the other cannot exceed the length of the first, we find that, if $\mathfrak{F}$ is nonempty, then there exist at least two words in $S(\mathfrak{A})$ of length not exceeding

$$K\left[\frac{n-m}{2}\right] + \left[\frac{n-m+1}{2}\right] = \frac{(n-m)(K+1)}{2} - \delta(n-m)\frac{K-1}{2},$$

mapped into one and the same word from $S_{\mathfrak{A}}(\mathfrak{B})$, which proves the assertion of the theorem; moreover, in the course of the proof we have obtained a source generating the whole set of basic words $[\mathfrak{F}]$, which is of independent interest when the coding is not one-to-one.

4. For the practical solution of the question of one-to-one coding, the geometric construction is inconvenient because of its cumbersomeness. It is

much simpler to construct the connection matrix [3] of the oriented graph described, $\|a_{ij}\|$, putting, however, $a_{ii} = 0$ for all i, $1 \leq i \leq n - m + 1$. In the case where $\mathfrak{M}$ contains identical elements, we identify all those vertices of the graph to which one and the same word corresponds. Let the number of vertices of the resulting graph be $T \leq n - m + 1$. The estimate in Theorem 1 in this case can be lowered, namely

$$N = K\left[\frac{T+1}{2}\right] - \delta(T).$$

On the basis of the theorem proved and Theorem 4 from [3] we can formulate an analytic criterion for the one-to-one property of alphabetic coding, and consequently of any dictionary coding, namely:

Theorem 2. *In order that the system $\mathfrak{A}$ carry out one-to-one coding, it is necessary and sufficient that*

$$\sum_{l=1}^{T} a_{1l}\,|a_{ij}\mathfrak{A}_{11}|_{ll} = 0.$$

Here

$$|a_{ij}\mathfrak{A}_{lk}|_{kl} = \sum a_{ki_1}a_{i_1i_2}\ldots a_{i_rl},$$

where the summation is over all arrangements $i_1 i_2 \ldots i_r$, $r = 0, 1, \ldots, T-2$—the quasi-minor* of the matrix $\|a_{ij}\|$.

Research Physico-Technical Institute
of Gorky State University
named after N. I. Lobachevsky

Received
12 I 1960

REFERENCES

[1] L. N. Korolev, DAN, **113**, No. 4 (1957).
[2] N. Chomsky, Inf. and Control, **2**, No. 2, 137 (1959).
[3] G. N. Povarov, UMN, **11**, issue 5 (71), 195 (1956).

* On the calculation of quasi-minors see [3].

Note: Figure translations are in progress. See original paper for figures.