CYBERNETICS AND CONTROL THEORY

BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS

V. I. Levenshtein

(Presented by Academician P. S. Novikov, January 4, 1965)
Translated from Doklady Akademii Nauk SSSR, Vol. 163, No. 4,
pp. 845-848, August, 1965
Original article submitted January 2, 1965

Investigations of transmission of binary information usually consider a channel model in which failures of the type 0 → 1 and 1 → 0 (which we will call reversals) are admitted. In the present paper (as in [1]) we investigate a channel model in which it is also possible to have failures of the form $0 \to \Lambda$, $1 \to \Lambda$, which are called deletions, and failures of the form $\Lambda \to 0$, $\Lambda \to 1$, which are called insertions (here Λ is the empty word). For such channels, by analogy to the combinatorial problem of constructing optimal codes capable of correcting s reversals, we will consider the problem of constructing optimal codes capable of correcting deletions, insertions, and reversals.

1. Codes Capable of Correcting Deletions and Insertions

By a binary word we will mean a word in the alphabet $\{0, 1\}$. By a code we will mean an arbitrary set of binary words that has fixed length.[1] We will say that a code K can correct s deletions (s insertions) if any binary word can be obtained from no more than one word in K by s or fewer deletions (insertions). This last property guarantees the possibility of unique determination of the initial code word from a word obtained as the result of some number $i\,(i \geq 0)$ of deletions and some number $j\,(j \geq 0)$ of insertions if $i + j \leq s$. The following assertion shows that all of the codes defined above are equivalent.

Lemma 1. Any code that can correct s deletions (like any code that can correct s insertions) can correct s deletions and insertions.

Proof (by contradiction). Assume that the same word z is obtained from a word x of length n by i_1 deletions and j_1 insertions, where $i_1 + j_1 \leq s$, and from a word y of length n by i_2 deletions and j_2 insertions, where $i_2 + j_2 \leq s$. If the symbols that were inserted (deleted) from at least one of the words x or y to obtain z are deleted from (inserted into) the word z, then, as we can easily see, we obtain a word that can be obtained from both x and y by no more than $\max(i_2 + j_1, j_2 + i_1)$ deletions (insertions). Because x and y have the same length, $j_1 - i_1 = j_2 - i_2$ and, consequently, $i_2 + j_1 = j_2 + i_1 = 1/2\,(i_1 + i_2 + j_1 + j_2) \leq s$, which proves Lemma 1.

Codes that can correct s deletions and insertions admit another, metric, description. Consider a function $\rho(x, y)$ defined on pairs of binary words and equal to the smallest number of deletions and insertions that transform the word x into y. It is not difficult to show that the function $\rho(x, y)$ is a metric, and that a code K can correct s deletions and insertions if and only if $\rho(x, y) > 2s$ for any two different words x and y in K.

Let B_n be the set of all binary words of length n. For an arbitrary word x in B_n, let $|x|$ denote the number of ones in x, and let $\|x\|$ be the number of runs[2] in the word x. We will now estimate the number $P_s(x)$ $[Q_s(x)]$ of different words that can be obtained from x by s deletions (s insertions). We have the bounds

$$C^{s}_{\|x\|-s+1} \leqslant P_s(x) \leqslant C^{s}_{\|x\|+s-1}, \tag{1}$$

$$\sum_{i=0}^{s} C_n^{\,i} 2^{s-i} \leqslant Q_s(x) \leqslant \sum_{i=0}^{s} C_n^{\,i} C_s^{\,i} 2^{s-i}. \tag{2}$$

In order to prove the upper bound in (1), note that each word obtained by deletion from x is uniquely determined by the number of symbols deleted

[1] The definitions given below are also meaningful if the code is taken to mean an arbitrary set of words (possibly of different lengths) in some alphabet containing r letters ($r \geq 2$). We should note, however, that in the case of words of different length Lemma 1 is generally not true.

[2] By a run in a word x we mean a maximal subword consisting of identical symbols. For example, the word x = 01101 has 4 runs.

from each run, so $P_s(x)$ is no greater than the number of combinations of s items taken $\|x\|$ at a time. On the other hand, it is easy to see that if one symbol is eliminated from any s pairwise nonadjacent runs in x, all of the words thus obtained will be different. This leads to the lower bound in (1), if we note that the number of such words is equal to the number of ordered partitions of the number $\|x\|-s$ into $s+1$ non-negative terms, where only two may, perhaps, be equal to zero. The upper bound in (2) follows from the fact that each word obtained from $x=\sigma_1 \ldots \sigma_n$ by s insertions can be obtained in the following manner. For some $i\,(i=0, 1, \ldots, s)$, choose i indices $n_1, \ldots, n_i$ $(1 \le n_1 < \ldots < n_i \le n)$ and $i+1$ words $\beta_1, \ldots, \beta_i, \beta_{i+1}$ such that the sum of their lengths is s and such that each of the first i words β_j is nonempty and does not end in the symbol σ_{n_j}; then, insert each word β_j $(j=1, \ldots, i)$ into the word x before the symbol σ_{n_j}, and insert β_{i+1} before the symbol σ_n. The lower bound in (2) follows from the fact that if each of the words $\beta_1, \ldots, \beta_i$ has length 1, all of the words obtained from x in this way are different.

We should note that (1) and (2) imply that $P_1(x) = \|x\|$ and $Q_1(x)=n+2$.

Let $L_s(n)$ denote the power (number of words) of a maximal code in B_n that can correct s deletions and insertions.

Lemma 2.[3] For fixed s and $n \to \infty$

$$2^s(s!)^2 2^n / n^{2s} \lesssim L_s(n) \lesssim s!\, 2^n / n^s. \tag{3}$$

Proof. Let K be a maximal code in B_n that can correct s deletions and insertions, and for arbitrary $k\,(1 \le k < n/2)$, let $L_s(n) = L_k' + L_k''$, where L_k' is the number of words $x \in K$ such that $k < \|x\| < n-k$. By the definition of K, $\sum\limits_{x\in K} P_s(x) \le 2^{n-s}$, and because of maximality, $\sum\limits_{x\in K} R_{2s}(x) \ge 2^n$, where $R_{2s}(x)$ is the number of words at a distance of 2s or less [in the metric $\rho(x, y)$] from x. It follows from (1) and (2) that $2^{n-s} \ge L_k' C_{k-s}^{\,s}$ and

$$2^n \leqslant (L_k' C^s_{n-k+s} + L_k'' C^s_{n+s-1}) \sum_{i=0}^{s} C^i_{n-1} C_s^{\,i} 2^{s-i}.$$

Estimate (3) follows from these last inequalities when we note that $L_k'' \leqslant 2\left(\sum\limits_{i=1}^{k} C^{i-1}_{n-1} + \sum\limits_{i=n-k}^{n} C^{i-1}_{n-1}\right) = 2\sum\limits_{i=0}^{k} C^i_n$ (since the number of words in B_n with i runs is $2C^{i-1}_{n-1}$), and we use the fact that $\sum\limits_{i=0}^{k} C_n{}^i = o\left(\frac{2^n}{n^{2s}}\right)$ when $k=[n/2-(s n \ln n)^{1/2}]$ and $n \to \infty$ (see, for example, [2]).

Theorem 1.

$$L_1(n) \sim 2^n / n. \tag{4}$$

Proof. In virtue of Lemma 2, it is sufficient to prove that

$$L_1(n) \geqslant 2^n / (n+1). \tag{5}$$

In order to prove this, we will use one of the Varshamov-Tenengol'ts [3] constructions. Consider the class of codes $K^a_{n,m}$, where each $K^a_{n,m}$ $(a=0, 1, \ldots, m-1)$ is defined as the set words $\sigma_1 \ldots \sigma_n$ in B_n such that $\sum\limits_{i=1}^{n} \sigma_i i \equiv a \pmod{m}$. We will show that for $m \ge n+1$, each code $K^a_{n,m}$ can correct one deletion. As the result of one deletion, assume that a word $x=\sigma_1 \ldots \sigma_n$ in $K^a_{n,m}$ has been transformed into the word $x'=\sigma_1' \ldots \sigma'_{n-1}$. We can then assume that we know $|x'|$ and the smallest non-negative residue of $a - \sum\limits_{i=1}^{n-1} \sigma'_i i$ mod m, which we will denote by a'. In order to restore the word x from x', it is clearly sufficient to know: 1) which of the binary symbols 0 or 1 has been eliminated and 2) either the number (which we denote by n_0) of zeros to the left of the deleted symbol if this symbols is 1, or the number (which we denote by n_1) of ones to the right of the deleted symbol, if this symbol is 0. But it follows from the definition of $K^a_{n,m}$ and the numbers n_0 and n_1 that when $m \ge n+1$ we have either $a' = |x'| + 1 + n_0$ (if the symbol 1 has been deleted) or $a' = n_1$ (if the symbol 0 has been deleted), and $n_1 \le |x'|$. As a result, depending on whether a' is larger than $|x'|$ or not, we can determine which of the binary symbols has been deleted, and then find n_0 or n_1. As a result, by Lemma 1, each code $K^a_{n,m}$ can, for $m \ge n+1$, correct one deletion or insertion. Since each of the words in B_n belongs to the same one of the m codes $K^a_{n,m}$ $(a=0, 1, \ldots, m-1)$, at least one of these codes contains no fewer than $2^n/m$ words, which, for $m=n+1$, yields estimate (5).

2. Codes that Can Correct Deletions, Insertions, and Reversals

We will say that a code K can correct s deletions, insertions, and reversals if any binary word can be obtained from no more than one word in K by s or fewer deletions, insertions,

[3] In what follows the notation $f(n) \lesssim g(n)$ will mean that $\overline{\lim\limits_{n\to\infty}} f(n) / g(n) \leqslant 1$,, while the notation $f(n) \sim g(n)$ will mean that $\lim\limits_{n\to\infty} f(n) / g(n) = 1$.

or reversals. It can be shown that the function r(x, y) defined on pairs of binary words as equal to the smallest number of deletions, insertions, and reversals that will transform x into y is a metric, and that a code K can correct s deletions, insertions, and reversals if and only if $r(x, y) > 2s$ for any two different words x and y in K. Let $M_s(n)$ denote the power of the maximal code in B_n that can correct s deletions, insertions, and reversals.

Theorem 2.

$$2^{n-1}/n \leqslant M_1(n) \leqslant 2^n/(n+1). \tag{6}$$

Proof. The upper bound is Hamming's estimate [4] for codes that can correct one reversal. In order to prove the lower bound, it is sufficient to show that all of the codes $K^a_{n,m}$ defined in the proof of Theorem 1 are, when $m \geq 2n$, capable of correcting one deletion, insertion, or reversal. The fact that these codes can correct deletions or insertions has already been proved. We should note, furthermore, that if no more than one reversal is required to change a word $\sigma_1 \ldots \sigma_n$ in $K^a_{n,m}$ into a word $\sigma_1' \ldots \sigma_n'$, the smallest of the non-negative residues of $a - \sum_{i=1}^{n} \sigma_i' i$ and $\sum_{i=1}^{n} \sigma_i' i - a$ mod 2n is larger than or equal to j, where j is the index of the reversed symbol (or $j = 0$ if there is no reversal).

By using the same method as we used to prove Lemma 2, we can show that for fixed s and $n \to \infty$

$$\left((2s)! \Big/ \sum_{i=0}^{s} 2^{-i} C_{2s}^{2i} C_{2i}^{\ i} \right) \frac{2^n}{n^{2s}} \leqslant M_s(n) \leqslant s!\, \frac{\ddot{\bar{2}}^n}{n^s}. \tag{7}$$

3. Use of Codes for Transmission (Without Synchronizing Symbols) Over Channels that Delete, Insert, and Reverse

Let $l'_{s,n}$ ($l''_{s,n}$; $l_{s,n}$; $m_{s,n}$) denote a channel in which no more than s deletions (insertions; deletions and insertions; deletions, insertions, and reversals) occur in each segment of length n. We agree to write the sequence obtained at a channel output from an arbitrary infinite sequence $z_1 z_2 \ldots$ of words in a code J in the form $z_1' z_2' \ldots$, where z_i' denotes the word obtained from the code word z_i as the result of failures in the channel. We will call a code J admissible for a given channel if there exists a finite automaton[4] that maps any sequence $z_1' z_2' \ldots$ into the sequence $z_1 z_2 \ldots$. In order for a code J to be admissible for the channels defined above, it is necessary (but generally not sufficient) that it be a code capable of correcting s failures of the appropriate types. The following assertion is useful for construction of admissible codes: for any binary words α and β, the codes K and $K_{\alpha,\beta} = \{\alpha x \beta,\ x \in K\}$ can correct the same number of failures of the types under discussion. This statement follows from the obvious equations $\rho(\alpha x \beta, \alpha y \beta) = \rho(x, y)$, $r(\alpha x \beta, \alpha y \beta) = r(x, y)$. In what follows, the word $\beta\alpha$ will play the role of a separator between code words, although it is generally distorted by the channel.

We should also note the important fact that in contrast to the case of the channel l_{sn}, in the case of the channels $l''_{s,n}$, $l_{s,n}$, $m_{s,n}$ no code J permits, when $s \geq 2$ (i.e., in channels with two or more insertions) determination of the end of the word z_1' from any sequence $z_1' z_2' \ldots$. This means that, in the cases indicated, decoding must start with the assumption that not only can there be failures in the channel, but there can be failures due to improper location of the beginning of a word z_i' (decoding failures). The idea at the basis of the constructions proposed below for the indicated channels is that as a result of treating decoding failures as channel failures, no more than s failures occur in each code word. This is achieved by decreasing the length of the code and appropriately selecting a separator $\beta\alpha$. The following statements hold: 1) if a code K in B_{n-2s-1} can correct s deletions, then the code $J = K_{1^s, 0^s}$ is admissible for the channel $l'_{s,n}$; 2) if a code K in B_{n-4s} can correct s insertions, then $J = K_{\Lambda, 1^s 0^s}$ is admissible for the channel $l''_{s,n}$; 3) if K $B_{n-4(s+1)^2-2s}$ can correct s deletions, insertions, and reversals (insertions and deletions), $J = K_{\Lambda, (1^{s+1} 0^{s+1})^{s+1} 1^s}$ is admissible[5] for the channel $m_{s,n}$ ($l_{s,n}$).

LITERATURE CITED

1. F. F. Sellers Jr., IRE Trans., IT-8, No. 1 (1962).

2. W. Feller, An Introduction to Probability Theory and Its Applications [Russian translation], 1964.

[4] In some generalized sense (see, for example [5]).

[5] It can be shown that if a code K in B_{n-7} can correct one deletion, insertion, or reversal (e.g., $K = K_{n-7, 2(n-7)}$), the code $J = K_{11, 01}$ is admissible.

3. R. R. Varshamov and G. M. Tenengol'ts, Avtomatika i telemekhanika, 26, No. 2 (1965).

4. R. W. Hamming, Bell Syst. Techn. J., 29, No. 2 (1950).

5. V. I. Levenshtein, Problemy kibernetiki, No. 11, 1964.

All abbreviations of periodicals in the above bibliography are letter-by-letter transliterations of the abbreviations as given in the original Russian journal. *Some or all of this periodical literature may well be available in English translation.* A complete list of the cover-to-cover English translations appears at the back of this issue.