PSYCHOLOGY

Machine Learning from Noisy Information

A METHOD of approach is described here in connexion with international telegraph (Morse) code, but not much change is necessary with other systems of substitution ciphers[1]. At this stage no reference is made to a vocabulary or to the syntactic and semantic context[2].

Let us start with zero noise-level and come later to the simulation of noise.

As is known, there are five different basic elements in the Morse code: dot, dash, intra-letter space, inter-letter space, and inter-word space. (For the sake of simplicity, we will now ignore punctuation marks. Their inclusion does not, however, represent any difficulty.) By using single-letter frequency charts for English and establishing the word-length distribution function with regard to the number of letters in a word (an analysis of 4,000 words, for example, has given an average word-length of 4·49 letters), we can show that the respective frequencies of the foregoing basic elements are $f_1 = 0{\cdot}308$, $f_2 = 0{\cdot}198$, $f_3 = 0{\cdot}297$, $f_4 = 0{\cdot}155$ and $f_5 = 0{\cdot}045$.

A computer can easily pick out the three basic elements that represent spaces since two identical of these can never be found in sequence. (We may note here that considerable research has been done in the related field of computer recognition and transcription of manual Morse, such as by Blair[3], Freimer *et al.*[4] and Gold[5].) When the information becomes noisy, as set out later, the machine looks for the elements with minimum number of repetitions in sequence. A comparison between the frequencies of occurrence of the elements so selected allows completion of the further classification (intra-letter, inter-letter and inter-word) due to the significant differences between f_3, f_4 and f_5. In a similar fashion dots and dashes can also be distinguished.

The next step is to apply the well-known statistical procedure used to fit different basic frequency distributions to an empirical one. It is necessary to minimize:

$$\chi^2 = \sum_{(\alpha)} \frac{(f_s^{(\alpha)} - f_p^{(\alpha)})^2}{f_p^{(\alpha)}} \qquad (1)$$

where the summation is extended over the whole alphabet; $f_p^{(\alpha)}$ is the frequency of occurrence of a particular letter (α) in the established chart (population value); $f_s^{(\alpha)}$ is the sample value of the same quantity for the letter assumed to be the same. The minimization is carried out by systematically changing the equivalence table (representations) of the Morse code. The starting set is usually very near to the optimum but, of course, it need not necessarily be the true set. A more reliable, but much more complicated, procedure would be the minimization of:

$$\chi^2 = w_1 \sum_{(\alpha)} \frac{(f_s^{(\alpha)} - f_p^{(\alpha)})^2}{f_p^{(\alpha)}} + w_2 \sum_{(\alpha,\beta)} \frac{(f_s^{(\alpha,\beta)} - f_p^{(\alpha,\beta)})^2}{f_p^{(\alpha,\beta)}} + {} + w_3 \sum_{(\alpha,\beta,\gamma)} \frac{(f_s^{(\alpha,\beta,\gamma)} - f_p^{(\alpha,\beta,\gamma)})^2}{f_p^{(\alpha,\beta,\gamma)}} \qquad (2)$$

where also the frequencies of occurrence of digraphs and trigraphs, $f^{(\alpha,\beta)}$ and $f^{(\alpha,\beta,\gamma)}$, are involved[6]; w_1, w_2 and w_3 are certain weighting factors, undetermined at present, representing the varying importance of each term.

The noisiness of the information is simulated by over-writing a given percentage of basic elements by other (or sometimes the same) ones, at random. The aim of this procedure is to find the minimum redundancy necessary at different noise-levels so that it is still possible to decipher the Morse code. Fig. 1 shows the nature of an expected diagram. The curve represents the average relationship between noise-level (*NL*) and minimum information content (*MIC*) of messages. We note that it may well be idealized in the sense that even noiseless messages cannot always be recognized without errors, and there could exist a certain amount of 'irreducible' error.

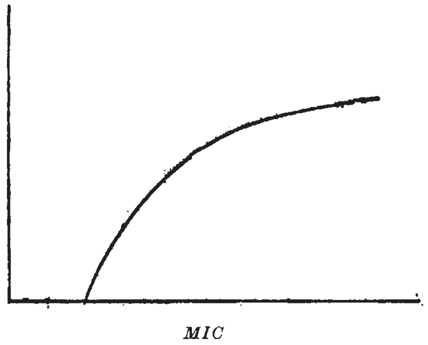

Fig. 1. Expected average relationship between noise-level (*NL*) and minimum information content (*MIC*) of messages

N. V. FINDLER*

Carnegie Institute of Technology,
Pittsburgh, Penn.

* Present address: University of Kentucky, Lexington.

[1] Findler, N. V., Computer for Cryptanalysis. Part (b), Paper *BI* 3.3, Summarized *Proc. First Conf. Automatic Computing and Data Processing*, Australia (1960).

[2] Evens, M. W., *et al.*, *Mass. Inst. Tech. Lincoln Lab. Rep.*, No. 54, *G*–0022 (1960).

[3] Blair, C. B., *J. Assoc. Comp. Mach.*, **6**, 429 (1959).

[4] Freimer, M., *et al.*, *Inst. Rad. Eng. Trans. Information Theory IT–5*, 25 (1959).

[5] Gold, B., *Inst. Rad. Eng. Trans. on Information Theory IT*–5, 17 (1959).

[6] Baddeley, A. D., *et al.*, *Nature*, **186**, 414 (1960).

Influence of Motivation, Perception and Attention on Age-related Changes in Short-term Memory

THE responses made by young normal adults to various methods of dichotic stimulation have been investigated and reported by Broadbent[1]. His observations have led him to postulate that both perceptual and storage mechanisms are necessary for the correct sequential recall of such simultaneous stimulation.

I have put forward the hypothesis that since increasing age in normal individuals also seems to affect learning capacity, and as such impairment may also depend on changes in some short-term storage process, it could be anticipated that responses to dichotic stimulation might also show associations with age.

Broadbent has suggested that the first half-set of digits recalled from each dichotic span does not involve the storage process. If age primarily affects storage then the reproduction of the first half-set recalled should not be affected by advancing years. If the second half-set recalled must pass through the storage process then the recall of these digits should be affected by age. Investigations by Inglis and Caird[2] and Mackay and Inglis[3] show that this appears to be the case.

The results of these two experiments, carried out on different subjects (total No. = 280) by different experimenters, proved to be in very close agreement. As age increases there is little or no significant impairment in the ability to recall the half-spans reproduced first. Progressively and significantly greater difficulty is, on the other hand, shown in the reproduction of the second half-spans as age advances. Furthermore, the longer the span