H. P. Luhn

The Automatic Creation of Literature Abstracts*

Abstract: Excerpts of technical papers and magazine articles that serve the purposes of conventional abstracts have been created entirely by automatic means. In the exploratory research described, the complete text of an article in machine-readable form is scanned by an IBM 704 data-processing machine and analyzed in accordance with a standard program. Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the "auto-abstract."

Introduction

The purpose of abstracts in technical literature is to facilitate quick and accurate identification of the topic of published papers. The objective is to save a prospective reader time and effort in finding useful information in a given article or report.

The preparation of abstracts is an intellectual effort, requiring general familiarity with the subject. To bring out the salient points of an author's argument calls for skill and experience. Consequently a considerable amount of qualified manpower that could be used to advantage in other ways must be diverted to the task of facilitating access to information. This widespread problem is being aggravated by the ever-increasing output of technical literature. But another problem — perhaps equally acute — is that of achieving consistence and objectivity in abstracts.

The abstracter's product is almost always influenced by his background, attitude, and disposition. The abstracter's own opinions or immediate interests may sometimes bias his interpretation of the author's ideas. The quality of an abstract of a given article may therefore vary widely among abstracters, and if the same person were to abstract an article again at some other time, he might come up with a different product.

The application of machine methods to literature searching is currently receiving a great deal of attention and now indicates that both human effort and bias may be eliminated from the abstracting process. Although rapid progress is being made in the development of systems using modern electronic data-processing devices, their efficiency depends on availability of literary information in machine-readable form. It is evident that the transcription of existing printed text into this form would have to be done manually at this time. In the future, however, print-reading devices should be sufficiently developed for this task. For material not yet printed, tape-punching devices attached to typewriters and typesetting machines could readily produce machine-readable records as by-products.

This paper describes some exploratory research on automatic methods of obtaining abstracts. The system outlined here begins with the document in machine-readable form and proceeds by means of a programmed sampling process comparable to the scanning a human reader would do. However, instead of sampling at random, as a reader normally does when scanning, the new mechanical method selects those among all the sentences of an article that are the most representative of pertinent information. These key sentences are then enumerated to serve as clues for judging the character of the article. Thus, citations of the author's own statements constitute the "auto-abstract."

The programs for creating auto-abstracts must be based on properties of writing ascertained *by analysis of specific types of literature*. Because the use of abstracts is an established practice in science and technology, it seemed desirable to develop the method first for papers and articles in this area. A primary objective of the development was to arrive at a system that could take full advantage of the capabilities of a modern electronic data-processing system such as the IBM 704 or 705, while at the same time keeping the scheme as simple as possible.

*Presented at IRE National Convention, New York, March 24, 1958.

Measuring significance

To determine which sentences of an article may best serve as the auto-abstract, a measure is required by which the information content of all the sentences can be compared and graded. Since the suitability of each sentence is relative, a value can be assigned to each in accordance with the quality criterion of significance.

The "significance" factor of a sentence is derived from an analysis of its words. It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnishes a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.

It should be emphasized that this system is based on the capabilities of machines, not of human beings. Therefore, regrettable as it might appear, the intellectual aspects of writing and of meaning cannot serve as elements of such machine systems. To a machine, words can be only so many physical things. It can find out whether or not certain such things are similar and how many of them there are. The machine can remember such findings and can perform arithmetic on those which can be counted. It can do all of this by means of suitable program instructions. The human intellect need be relied upon only to prepare the program.

Establishing a set of significant words

The justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject. This means of emphasis is taken as an indicator of significance. The more often certain words are found in each other's company within a sentence, the more significance may be attributed to each of these words. Though certain other words must be present to serve the important function of tying these words together, the type of significance sought here does not reside in such words. If such common words can be segregated substantially by non-intellectual methods, they could then be excluded from consideration.

This rather unsophisticated argument on "significance" avoids such linguistic implications as grammar and syntax. In general, the method does not even propose to differentiate between word forms. Thus the variants *differ, differentiate, different, differently, difference* and *differential* could ordinarily be considered identical notions and regarded as the same word. No attention is paid to the logical and semantic relationships the author has established. In other words, an inventory is taken and a word list compiled in descending order of frequency.

Procedures as simple as these, of course, are rewarding from the standpoint of economy. The more complex the method, the more operations must the machine perform and therefore the more costly will be the process. But in this case an even more fundamental justification for simplicity can be found in the nature of technical writing. Within a technical discussion, there is a very small probability that a given word is used to reflect more than one notion. The probability is also small that an author will use different words to reflect the same notion. Even if the author makes a reasonable effort to select synonyms for stylistic reasons, he soon runs out of legitimate alternatives and falls into repetition if the notion being expressed was potentially significant in the first place.

A word list compiled in accordance with the method outlined will generally take the form of the diagram in Fig. 1. The presence in the region of highest frequency of many of the words previously described as too common to have the type of significance being sought would constitute "noise" in the system. This noise can be materially reduced by an elimination technique in which text words are compared with a stored common-word list. A simpler way might be to determine a high-frequency cutoff through statistical methods to establish "confidence limits." If the line C in the figure represents this cutoff, only words to its right would be considered suitable for indicating significance. Since degree of frequency has been proposed as a criterion, a lower boundary, line D, would also be established to bracket the portion of the spectrum that would contain the most useful range of words. Establishing optimum locations for both lines would be a matter of experience with appropriately large samples of published articles. It should even be possible to adjust these locations to alter the characteristics of the output.

The curve for the degree of discrimination, or "resolving power," of the bracketed words in the figure might look something like the dotted line, E. It is apparent that words that cannot be put in the category of common words may sometimes fall to the left of line C. If the program has been properly formulated, the location of these words on the diagram would indicate their loss of discriminatory power. The word "cell" in an article on biology may be an example of this. It may be anticipated that the cutoff line, once established, may be stable over many different degrees of specialization within a field, or even over many different fields. Moreover, the resolving power would increase with the need for finer resolution. The case of a common word falling in the region to the right of line C can be tolerated because of its lesser degree of interference.

Establishing relative significance of sentences

As pointed out earlier, the method to be developed here is a probabilistic one based on the physical properties of written texts. No consideration is to be given to the meaning of words or the arguments expressed by word combinations. Instead it is here argued that, whatever the topic, the closer certain words are associated, the more specifically an aspect of the subject is being treated. Therefore, wherever the greatest number of frequently occurring different words are found in greatest physical proximity to each other, the probability is very high that

the information being conveyed is most representative of the article.

The significance of degree of proximity is based on the characteristics of spoken and written language in that ideas most closely associated intellectually are found to be implemented by words most closely associated physically. The divisions of written text into sentences, paragraphs, chapters, et cetera, is another physical manifestation of the graduating degree of association of ideas. These aspects have been discussed in detail in an earlier paper by the writer.*

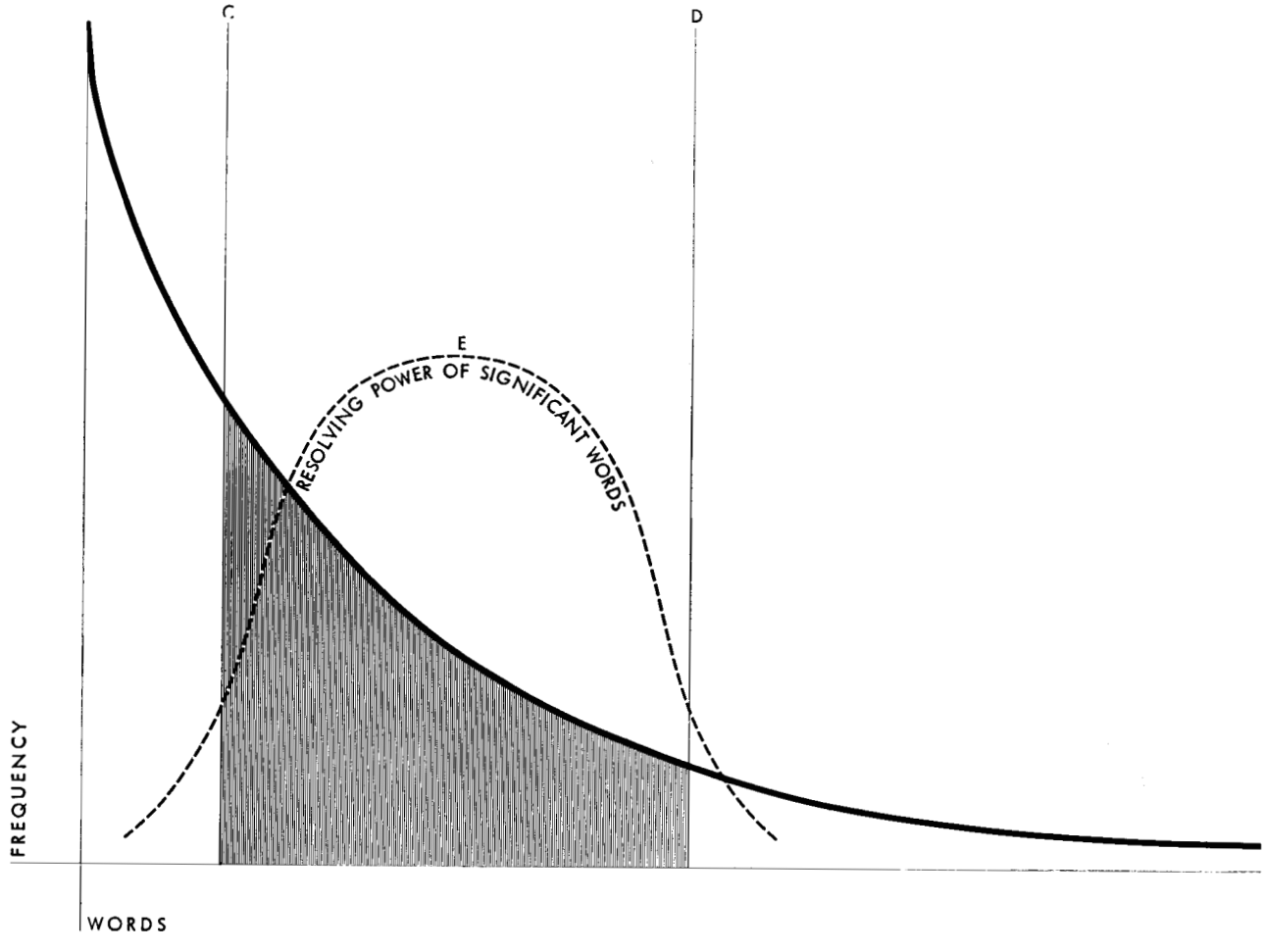
From these considerations a "significance factor" can be derived which reflects the number of occurrences of significant words within a sentence and the linear distance between them due to the intervention of non-significant words. All sentences may be ranked in order of their significance according to this factor, and one or several of the highest ranking sentences may then be selected to serve as the auto-abstract.

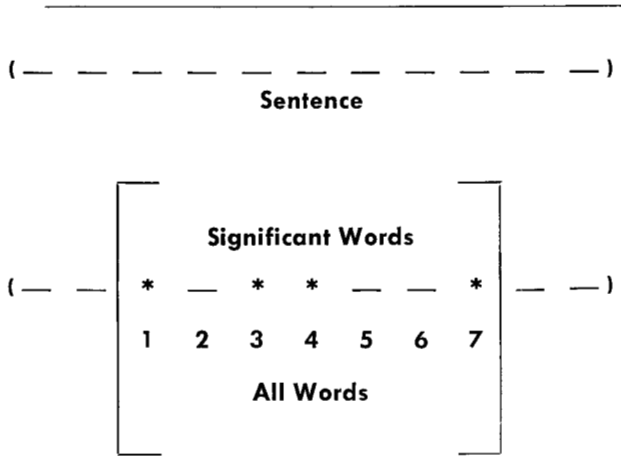
*H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development*, 1, No. 4, 309-317 (October 1957).

It must be kept in mind that, when a statistical procedure is applied to produce such rankings, the criterion is the relationship of the significant words to each other rather than their distribution over a whole sentence. It therefore appears proper to consider only those portions of sentences which are bracketed by significant words and to set a limit for the distance at which any two significant words shall be considered as being significantly related. A significant word beyond that limit would then be disregarded from consideration in a given bracket, although it might form a bracket, or cluster, in conjunction with other words in the sentence. An analysis of many documents has indicated that a useful limit is four or five non-significant words between significant words. If with this separation two or more clusters result, the highest one of the several significance factors is taken as the measure for that sentence.

A scheme for computing the significance factor is given by way of example in Fig. 2. It consists of ascertaining the extent of a cluster of words by bracketing, counting the number of significant words contained in the cluster, and dividing the square of this number by the

Figure 1 **Word-frequency diagram.**
Abscissa represents individual words arranged in order of frequency.





Portion of sentence bracketed by and including significant words not more than four non-significant words apart. If eligible, the whole sentence is cited.

Figure 2 **Computation of significance factor.** *The square of the number of bracketed significant words (4) divided by the total number of bracketed words (7) = 2.3.*

total number of words within this cluster. The results based on this formula, as performed on about 50 articles ranging from 300 to 4,500 words each, have been encouraging enough for further evaluation by a psychological experiment involving 100 people. This experiment will determine on an objective basis the effectiveness of the abstracts generated.

The resolving power of significant words derived under the method described depends on the total number of words comprising an article and will decrease as the total number of words increases. In order to overcome this effect, the abstracting process may be performed on subdivisions of the article, and the highest ranking sentences of each of these divisions may then be selected and combined to constitute the auto-abstract. In many cases the author provides such divisions as part of the organization of his paper, and they may therefore serve for the extended process. Where such deliberate divisions are absent they can be made arbitrarily in accordance with some criteria established by experience. These divisions would be arranged in such a way that they overlap each other, for lack of any simple means of mechanically detecting the exact point of the author's transition to a new subject subdivision.

A more detailed account of these and other computing methods, as well as details on programming electronic data-processing machines for this procedure, will be given in subsequent papers.

By way of example, two auto-abstracts are included in this paper. Exhibit 1 shows four selected sentences of a 2,326-word article from *The Scientific American*. A table of word frequency is also given. Exhibit 2 shows the highest ranking sentence of a 783-word article from the Science Section of *The New York Times*. A complete reproduction of this article is given.

Machine procedures

The abstracts described in this paper were prepared by first punching the documents on cards. Punctuation marks in the printed text not available on the standard key punch were replaced by other key-punch characters. The cards thus produced constitute the machine-readable form of the document.

The abstracting process was initiated by transcribing the card record onto magnetic tape by means of an auxiliary card-to-tape unit. The resulting tape was introduced into an IBM 704 data-processing machine, which was programmed to read the taped text to separate it into its individual words, to note the position of each word in the document, the sentence and paragraph in which it appeared, and to note the punctuation preceding and following it. Concurrently, common words such as pronouns, prepositions, and articles were deleted from the list by a table-lookup routine. This operation was followed by a sorting program which arranged the remaining words in alphabetic order.

The next step of the machine operation was a consolidation of words which are spelled in the same way at their beginning, such as *similar* and *similarity*. This procedure was a simple statistical analysis routine consisting of a letter-by-letter comparison of pairs of succeeding words in the alphabetized list. From the point where letters failed to coincide, a combined count was taken of the non-similar subsequent letters of both words. When this count was six or below, the words were assumed to be similar notions; above six, different notions. Although this method of word consolidation is not infallible, errors up to 5% did not seem to affect the final results of the abstracting process. The machine then counted the occurrence of similar words derived in this way. Words of a stipulated low frequency were then deleted from the list and locations of the remaining words were sorted into order. These words thereby attained the status of "significant" words.

The significance factor for each sentence was determined by a computing routine in accordance with the formula previously mentioned. All sentences which scored above a predetermined cutoff value were written on an output tape along with their respective values. The basis for this cutoff value depends on the amount of detailed information needed for a given type of abstract. Results were then printed out from this tape.

Extended applications

Although a standard abstract has thus far been assumed in order to simplify the explanation of the machine process, extracts or condensations of literature are used for diverse purposes and may vary in length and orientation.

Exhibit 1

Source: The Scientific American, Vol. 196, No. 2, 86-94, February, 1957

Title: Messengers of the Nervous System

Author: Amodeo S. Marrazzi

Editor's Sub-heading: The internal communication of the body is mediated by chemicals as well as by nerve impulses. Study of their interaction has developed important leads to the understanding and therapy of mental illness.

Auto-Abstract*

It seems reasonable to credit the single-celled organisms also with a system of chemical communication by diffusion of stimulating substances through the cell, and these correspond to the chemical messengers (e.g., hormones) that carry stimuli from cell to cell in the more complex organisms. (7.0)†

Finally, in the vertebrate animals there are special glands (e.g., the adrenals) for producing chemical messengers, and the nervous and chemical communication systems are intertwined: for instance, release of adrenalin by the adrenal gland is subject to control both by nerve impulses and by chemicals brought to the gland by the blood. (6.4)

The experiments clearly demonstrated that acetylcholine (and related substances) and adrenalin (and its relatives) exert opposing actions which maintain a balanced regulation of the transmission of nerve impulses. (6.3)

It is reasonable to suppose that the tranquilizing drugs counteract the inhibitory effect of excessive adrenalin or serotonin or some related inhibitor in the human nervous system. (7.3)

*Sentences selected by means of statistical analysis as having a degree of significance of 6 and over.
†Significance factor is given at the end of each sentence.

Significant words in descending order of frequency (common words omitted).

46	nerve	12	body	6	disturbance	4	accumulate
40	chemical	12	effects	6	related	4	balance
28	system	12	electrical	5	control	4	block
22	communication	12	mental	5	diagram	4	disorders
19	adrenalin	12	messengers	5	fibers	4	end
18	cell	10	signals	5	gland	4	excitation
18	synapse	10	stimulation	5	mechanisms	4	health
16	impulses	8	action	5	mediators	4	human
16	inhibition	8	ganglion	5	organism	4	outgoing
15	brain	7	animal	5	produce	4	reaching
15	transmission	7	blood	5	regulate	4	recording
13	acetylcholine	7	drugs	5	serotonin	4	release
13	experiment	7	normal			4	supply
13	substances					4	tranquilizing

Total word occurrences in the document	2326
Different words in document:	
Total of different words	741
Less different common words	170
Different non-common words	571
Ratio of all word occurrences to different non-common words	~4:1
Non-common words having a frequency of occurrence of 5 and over:	
Total occurrences	478
Different words	39

Exhibit 2

Source: The New York Times, September 8, 1957, page E11

Title: Chemistry Is Employed in a Search for New Methods to Conquer Mental Illness

Author: Robert K. Plumb

SCIENCE IN REVIEW

Chemistry Is Employed in a Search for New Methods to Conquer Mental Illness

By ROBERT K. PLUMB

By coincidence this week-end in New York City marks the end of the annual meeting of the American Psychological Association and the begining of the annual meeting of the American Chemical Society.

Psychologists and chemists have never had so much in common as they now have in new studies of the chemical basis for human behavior. Exciting new finds in this field were also discussed last week in Iowa City, Iowa, at the annual meeting of the American Physiological Society and at Zurich, Switzerland, at the Second International Congress for Psychiatry.

Two major recent developments have called the attention of chemists, physiologists, physicists and other scientists to mental diseases: It has been found that extremely minute quantities of chemicals can induce hallucinations and bizarre psychic disturbances in normal people, and mood-altering drugs (tranquilizers, for instance) have made long-institutionalized people amenable to therapy.

Money to finance resreach on the physical factors in mental illness is being made available. Progress has been achieved toward the understanding of the chemistry of the brain. New goals are in sight.

At the psychiatrists meeting in Zurich last week, four New York City physicians urged their colleagues to broaden their concept of "mental disease," and to probe more deeply into the chemistry and metabolism of the human body for answers to mental disorders and their prevention.

Blood May Tell

Dr. Felix Marti-Ibanez and three brothers, Dr. Mortimer D. Sackler, Dr. Raymond R. Sackler and Dr. Arthur M. Sackler cited evidence that the blood chemistry of victims of schizophrenia is different from that of normal people. Perhaps multiple biological factors are responsible for this chemical change, they suggested.

Mental disease is a "developmental process" and long duration of a disorder may result in "permanent alteration of anatomy and physiology," they said. They urged that trials of new drugs which affect the brain should be concentrated on complex studies of the mechanism of action of the drugs. The variety of substances capable of producing profound mental effects is a new armory of weapons for use in investigating biological mechanisms underlying mental disease, they said.

The sources of behavioral disturbance are many and they may come from external as well as internal forces, the four reported. This concept has already proven practical, for instances, when it enabled psychiatrists to predict that the administration of ACTH and cortisone could produce psychosis.

"It led some years ago to the development of a blood test which was 80 per cent accurate in the identification of schizophrenic patients," they said. "It permitted us on physiologic grounds to deny that the psychoneuroses and the psychoses were lesser and greater degrees of the same disease process, and, in fact, to affirm that they represented opposite and even mutually exclusive directions of physiologic disturbances," they said.

Chemicals now available should be used not only to bring relief to the mentally sick but also to uncover the biological mechanisms of the disease processes themselves. "Only then will the metabolic era mature and bring to fruition man's long hoped for salvation from the ravages of mental disease," they reported.

Chemistry of the Brain

At the psychologist's meeting here, a technique for tracing electrical activity in specific portions of the animal brain was described by researchers from the University of California at Los Angeles. They reported that deep brain implants in cat brains were used to record electrical discharges created as the animals respond to stimulations to which they had been conditioned. In this way the California group reported, it is possible to track the sequence in which the brain brings its various parts into play in learning. Specific areas of memory in the brain may be located. Furthermore, the electrical pathways so traced out can be blocked temporarily by the use of chemicals. This poses new possibilities for studying brain chemistry changes in health and sickness and their alleviation, the California researchers emphasized.

The new studies of brain chemistry have provided practical therapeutic results and tremendous encouragement to those who must care for mental patients. One evidence that knowledge in the interdisciplinary field is accumulating fast came last week in an announcement from Washington.

This was the establishment by the National Institute of Mental Health of a clearing house of information on psychopharmacology. Literature in the field will be classified and coded so that staff members can answer a wide variety of technical and scientific questions. People working in the field are invited to send three copies of papers or other material — even informal letters describing work they may have in progress—to the Technical Information Unit of the center in Silver Spring, Md.

Exhibit 2 **Auto-Abstract**

Two major recent developments have called the attention of chemists, physiologists, physicists and other scientists to mental diseases: It has been found that extremely minute quantities of chemicals can induce hallucinations and bizarre psychic disturbances in normal people, and mood-altering drugs (tranquilizers, for instance) have made long-institutionalized people amenable to therapy. (4.0)

This poses new possibilities for studying brain chemistry changes in health and sickness and their alleviation, the California researchers emphasized. (5.4)

The new studies of brain chemistry have provided practical therapeutic results and tremendous encouragement to those who must care for mental patients. (5.4)

A condensation of a document to a given fraction of the original could be readily accomplished with the system outlined by adjusting the cutoff value of sentence significance. On the other hand, a fixed number of sentences might be required irrespective of document length. Here it would be a simple matter to print out exactly that number of the highest ranking sentences which fulfilled the requirement.

In many instances condensations of documents are made emphasizing the relationship of the information in the document to a special interest or field of investigation. In such cases sentences could be weighted by assigning a premium value to a predetermined class of words.

These two features of the auto-abstract, variable length and emphasis, might at times be usefully combined. In the case of a long, comprehensive paper, several condensed versions could be prepared, each of a length suitable to the requirements of its recipient and biased to his particular sphere of interest.

Along these same lines, a specificity ranking technique might prove feasible. If none of the sentences in an article attained a certain significance factor, it would be possible to reject the article as too generalized for the purpose at hand.

In certain cases an abstract might be amplified by following it with an enumeration of specifics, such as names of persons, places, organizations, products, materials, processes, et cetera. Such specific words could be selected by the machine either because they are capitalized or by means of lookup in a stored special dictionary.

Auto-abstracting could also be used to alleviate the translation burden. To avoid total translation initially, auto-abstracts of appropriate length could be produced in the original language and only the abstracts translated for subsequent analysis.

Finally, the process of deriving key words for encoding documents for mechanical information retrieval could be simplified by auto-abstracting techniques.

Conclusions

The results so far obtained for technical articles have indicated the feasibility of automatically selecting sentences that will indicate the general subject matter, very much as do conventional abstracts. What such auto-abstracts might lack in sophistication they will more than compensate for by their uniformity of derivation. Because of the absence of the variations of human capabilities and orientation, auto-abstracts have a high degree of reliability, consistency, and stability, as they are the product of a statistical analysis of the author's own words. In many cases the abstract obtained is the type generally referred to as the "indicative" abstract.

Once auto-abstracts are generally available, their users will learn how to interpret them and how to detect their implications. They will realize, for instance, that certain words contained in the sample sentences stand for notions which must have been elaborated upon somewhere in the article. If this were not so for a substantial portion of the words in the selected sentences, these sentences could not have attained their status based on word frequency.

There is, of course, the chance that an author's style of writing deviates from the average to an extent that might cause the method to select sentences of inferior significance. Since the title of the paper is always given in conjunction with the auto-abstract, there is a high probability that it will favorably supplement the abstract. However, there will always be a residue of inadequate results, and it appears to be entirely feasible to establish criteria by which a machine may recognize such exceptions and earmark them for human attention.

If machines can perform satisfactorily within the range outlined in this paper, a substantial and worthwhile saving in human effort will have been realized. The auto-abstract is perhaps the first example of a machine-generated equivalent of a completely intellectual task in the field of literature evaluation.

Received December 2, 1957