THE AUTOMATIC CREATION OF LITERATURE ABSTRACTS (Auto-Abstracts)

H. P. Luhn

IBM

RESEARCH CENTER

INTERNATIONAL BUSINESS MACHINES CORPORATION
YORKTOWN HEIGHTS, NEW YORK

THE AUTOMATIC CREATION OF LITERATURE ABSTRACTS

by

H.P. Luhn

International Business Machines Corporation
Research Center
Yorktown Heights, New York

ABSTRACT: Excerpts of technical papers and magazine articles that serve the purposes of conventional abstracts have been created entirely by automatic means. In the process described, the complete text of an ariticle in machine-readable form is scanned by an IBM 704 data processor and analyzed in accordance with a standard program. Statistical information derived on word frequencey 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."

Research Report
RC-42
January 15, 1958

This paper was presented to the IRE National Convention at New York City on March 24, 1958, and has also been accepted for publication in the April issue of the IBM Journal of Research and Development.

Printed in U. S. A.

INTRODUCTION

The purpose of abstracts in scientific and engineering literature is to facilitate quick recognition of the topic of published papers. The objective is to save a prospective reader time and effort in deciding whether he can expect to find 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 problem has been widely experienced and is being aggravated by the ever-increasing output of literature. But there is another problem that is, perhaps, equally acute--that of achieving consistence and objectivity.

An abstracter's product is almost always influenced by his background, attitude, and disposition. The abstracter's own opinions or immediate interests may 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 would likely come up with a different product.

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

The automatic abstracting system outlined here begins with the document in machine-readable form and proceeds by a programmed sampling process comparable in effect to the scanning a reader would do if an abstract were not available. However, instead of sampling at random as a reader normally does when scanning, the objective of the new mechanical method is to select 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 have to 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 was deemed 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 processor such as the IBM 704 or 705 and, at the same time, to keep the scheme as simple as possible consistent with achieving adequate results.

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 the words which comprise it. 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 these two measurements in combination.

It should be emphasized that the system described here is based on the capabilities of machines - not of human beings. Therefore - and as regrettable as it may appear to some - 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. It can remember such findings and can perform arithmetic on those which can be counted. It can do all of this by virtue of instructions impressed on it by means of a suitable program. The human intellect must be-relied upon to prepare the program, but is needed for this one task only.

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 sign 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 functional role of tieing these words together, the type of significance sought here does not reside in such words. If, as shall be proposed, such common words can be substantially segregated 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. The method does not even propose to differentiate between word forms. Thus the variants "differ," "differentiate," "different," "differently," "difference," and "differential," are considered to be 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 merely taken and a word list compiled in descending order of frequency.

Procedures as simple as these are, of course, 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 the confines of 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 Figure 1. The presence in the

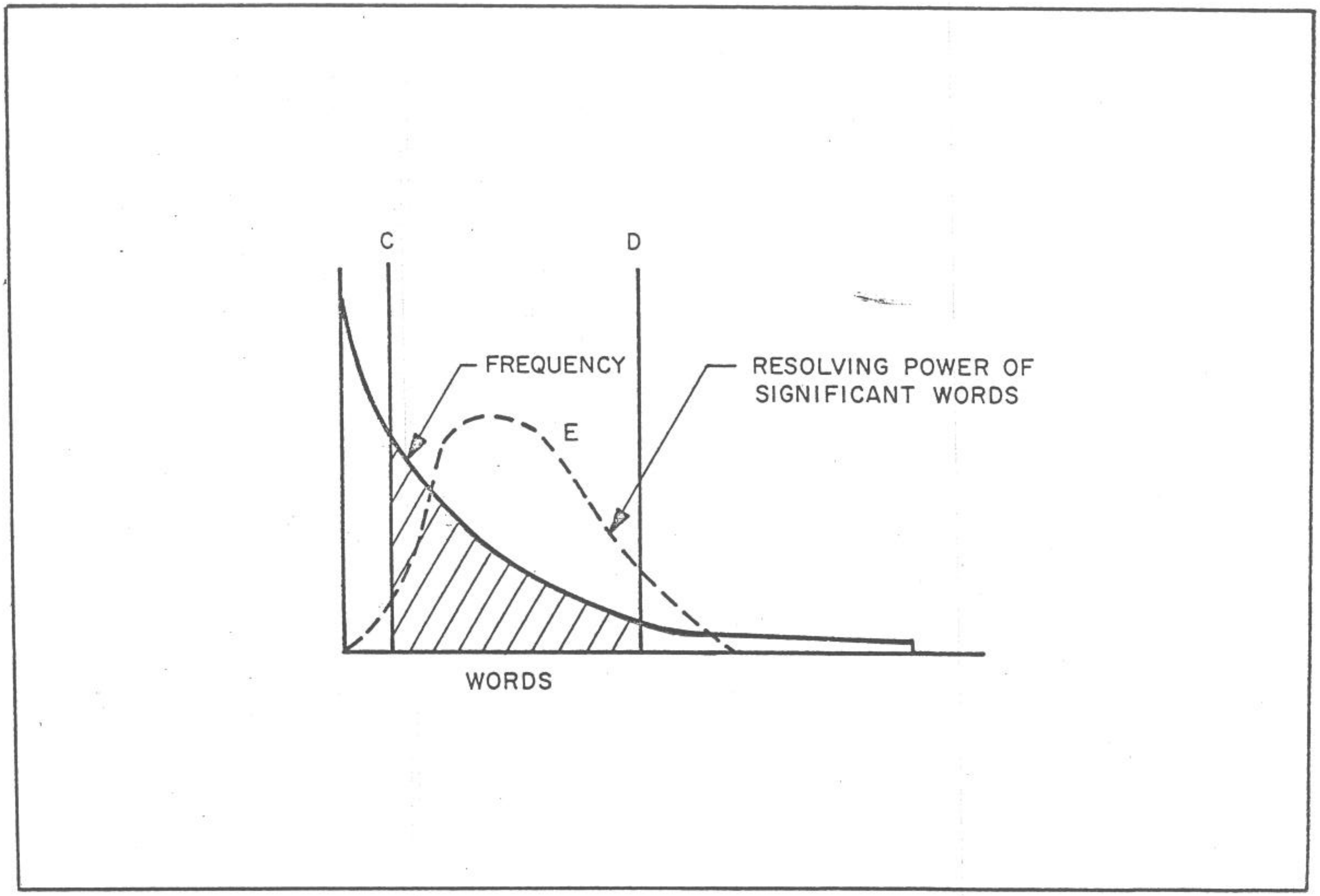


FIGURE I — CURVES OF WORD FREQUENCY AND RESOLVING POWER.

region of highest frequency of many of the words previously described as too common to have the type 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. If the line C in the figure were thought of as representing this cutoff, only words to its right would be considered suitable for indicating significance. Since degree of frequency has been proposed as 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, at times, words that cannot be put in the category of common words may fall to the left of line C. If the program has been properly formulated, this may be taken as an indication that these words have lost their discriminating power. (The word "cell" in an article on biology may be an example of this.) Thus it may be anticipated that, once established, the cutoff line may be stable over many different degrees of specialization within a field, or even over many different fields. Moreover, the resolving power would increase automatically with the need for finer resolution. In the case of a common word falling in the region to the right of line C, it 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 occurrring 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 of which it is a part.

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, etc., 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 nonsignificant 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.

It must be kept in mind that, in applying a statistical procedure 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 nonsignificant 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.

In one scheme of computing sentence significance the number of significant words in a bracketed cluster was divided by the square of the total number of words within the cluster. Tests based on this formula and performed on about 50 articles ranging from 300 to 4500 words each have yielded significantly successful results in producing sentences having the desired characteristics.

The resolving power of significant words derived under the method described is dependent 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 process may be performed on subdivisions of the article,and the highest ranking sentences of each of the divisions selected for the Auto-Abstract. In many cases the author

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

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 appended to this paper. Exhibit I shows 4 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 761-word article from the Science Section of the New York Times. A reproduction of this article is also 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 machinable form of the document.

The automatic creation of the abstracts 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, separate it into its individual words, and note the position of each word in the document, the sentence and paragraph in which it appeared, and the punctuation preceding and following it Concurrently, common words such as pronouns, prepositions, and articles weie deleted from the list by a table look-up routine. This operation was followed by a sorting program which arranged the remaining words in alphabetic order.

Subsequently, a count was performed of all words which were identical. In addition, different word forms of the same stem were consolidated so that thereafter they could be treated as a single word. This was done by a simple statistical analysis routine. The total of each set of such consolidated words was combined with the previously derived word count. 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. Results were then printed out from this tape.

EXTENDED APPLICATIONS

Although a "standard" abstract has thus far been assumed to simplify the explanation, extracts or condensations of literature are used for diverse purposes and may vary in length and orientation. Under certain circumstances a condensation of a document to a given fraction of the original might be desired. This could be readily accomplished with the system outlined by adjusting the cutoff value of sentence significance up or down. 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 a weighting of sentences could be achieved by assigning a premium value to a predetermined class of words.

The above features of the Auto-Abstract, variable length and emphasis, might at times be usefully combined. Where a long, comprehensive paper is involved, 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, etc. Such specific words could be selecte by the machine either because they are capitalized or by means of look up 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 the purpose of 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 being treated, 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 in that they are the product of a statistical analysis of the author's own words.

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 constituting the selected sentences, these sentences could not have attained their status based on the frequency of word usage.

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 conjuction 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 making of an Auto-Abstract is probably the first example of a machine doing the equivalent of what normally is a completely intellectual task in the field of literature evaluation.

EXHIBIT I

ABSTRACT CREATED ENTIRELY BY AUTOMATIC MEANS[†]

Article: Amodeo S. Marrazzi, "Messengers of the Nervous System," Scientific American, Vol. 196, No. 2, February 1957.

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, releas 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. (7.4)

[†] Sentences with significance factors of six and over as computed by the formula discussed on page 5 were selected. The significance factor is given at the end of each sentenc.

SIGNIFICANT WORDS IN DESCENDING ORDER OF FREQUENCY

(Common words omitted)

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

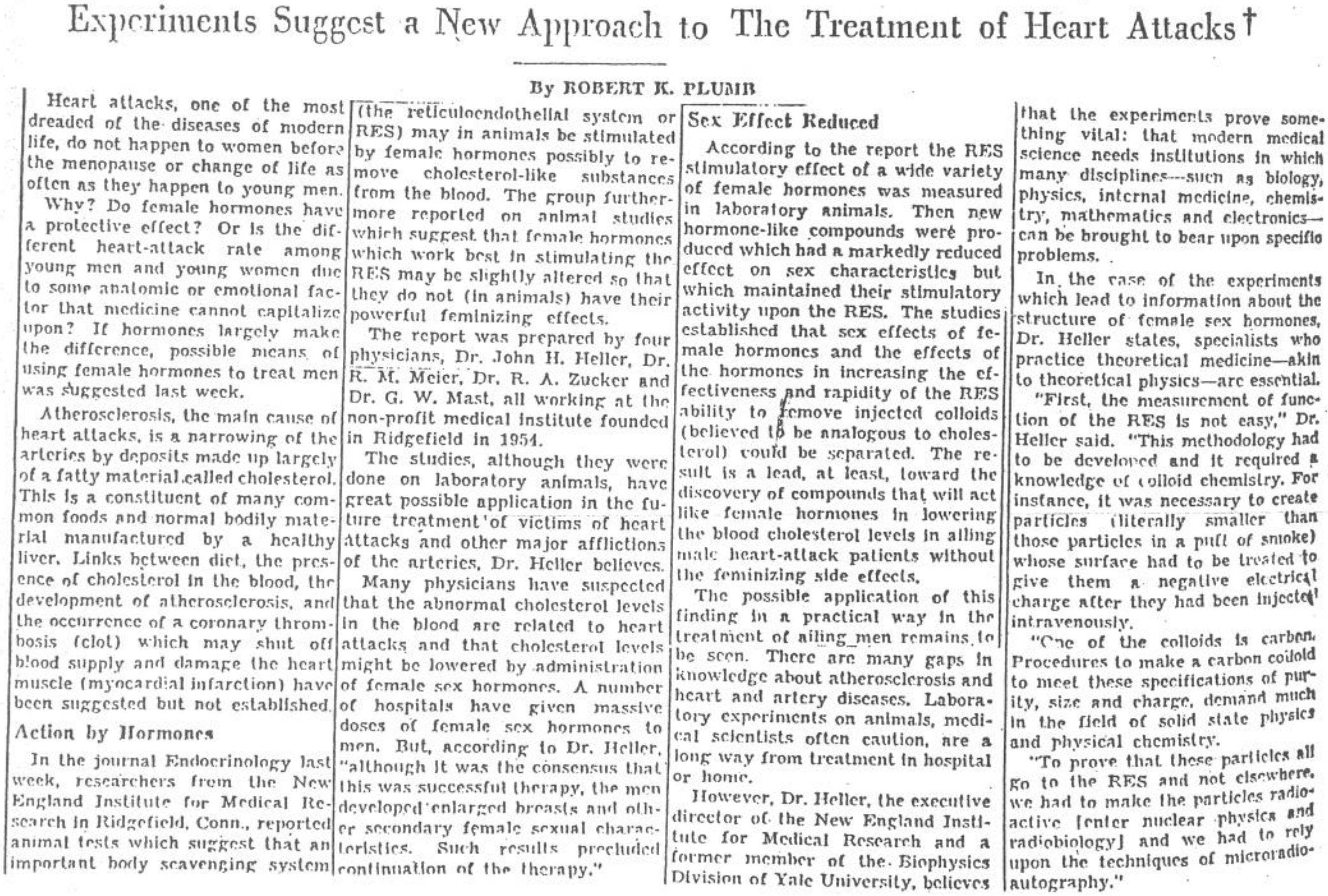
Words in document:		2326)	
Different words:	741	)	ratio = 4.1
less common words	170	)	
	571	571)	
Words of frequency 5 and over:		478	
Different words of frequency 5 and over:		39	
Average frequency:		12	

EXHIBIT II

AUTOMATIC SELECTION OF A "BEST" SENTENCE

Article: Robert K. Plumb, "Experiments Suggest a New Approach to the Treatment of Heart Attacks," The New York Times, September 22, 1957.

Auto Abstract: The result is a lead, at least, toward the discovery of compounds that will act like female hormones in lowering the blood cholesterol levels in ailing male heart-attack patients without the feminizing side effects.



Experiments Suggest a New Approach to The Treatment of Heart Attacks†

By ROBERT K. PLUMB

Heart attacks, one of the most dreaded of the diseases of modern life, do not happen to women before the menopause or change of life as often as they happen to young men.

Why? Do female hormones have a protective effect? Or is the different heart-attack rate among young men and young women due to some anatomic or emotional factor that medicine cannot capitalize upon? If hormones largely make the difference, possible means of using female hormones to treat men was suggested last week.

Atherosclerosis, the main cause of heart attacks, is a narrowing of the arteries by deposits made up largely of a fatty material called cholesterol. This is a constituent of many common foods and normal bodily material manufactured by a healthy liver. Links between diet, the presence of cholesterol in the blood, the development of atherosclerosis, and the occurrence of a coronary thrombosis (clot) which may shut off blood supply and damage the heart muscle (myocardial infarction) have been suggested but not established.

Action by Hormones

In the journal Endocrinology last week, researchers from the New England Institute for Medical Research in Ridgefield, Conn., reported animal tests which suggest that an important body scavenging system (the reticuloendothelial system or RES) may in animals be stimulated by female hormones possibly to remove cholesterol-like substances from the blood. The group furthermore reported on animal studies which suggest that female hormones which work best in stimulating the RES may be slightly altered so that they do not (in animals) have their powerful feminizing effects.

The report was prepared by four physicians, Dr. John H. Heller, Dr. R. M. Meier, Dr. R. A. Zucker and Dr. G. W. Mast, all working at the non-profit medical institute founded in Ridgefield in 1954.

The studies, although they were done on laboratory animals, have great possible application in the future treatment of victims of heart attacks and other major afflictions of the arteries, Dr. Heller believes.

Many physicians have suspected that the abnormal cholesterol levels in the blood are related to heart attacks and that cholesterol levels might be lowered by administration of female sex hormones. A number of hospitals have given massive doses of female sex hormones to men. But, according to Dr. Heller, "although it was the consensus that this was successful therapy, the men developed enlarged breasts and other secondary female sexual characteristics. Such results precluded continuation of the therapy."

Sex Effect Reduced

According to the report the RES stimulatory effect of a wide variety of female hormones was measured in laboratory animals. Then new hormone-like compounds were produced which had a markedly reduced effect on sex characteristics but which maintained their stimulatory activity upon the RES. The studies established that sex effects of female hormones and the effects of the hormones in increasing the effectiveness and rapidity of the RES ability to remove injected colloids (believed to be analogous to cholesterol) could be separated. The result is a lead, at least, toward the discovery of compounds that will act like female hormones in lowering the blood cholesterol levels in ailing male heart-attack patients without the feminizing side effects.

The possible application of this finding in a practical way in the treatment of ailing men remains to be seen. There are many gaps in knowledge about atherosclerosis and heart and artery diseases. Laboratory experiments on animals, medical scientists often caution, are a long way from treatment in hospital or home.

However, Dr. Heller, the executive director of the New England Institute for Medical Research and a former member of the Biophysics Division of Yale University, believes that the experiments prove something vital: that modern medical science needs institutions in which many disciplines---such as biology, physics, internal medicine, chemistry, mathematics and electronics—can be brought to bear upon specific problems.

In the case of the experiments which lead to information about the structure of female sex hormones, Dr. Heller states, specialists who practice theoretical medicine—akin to theoretical physics—are essential.

"First, the measurement of function of the RES is not easy," Dr. Heller said. "This methodology had to be developed and it required a knowledge of colloid chemistry. For instance, it was necessary to create particles (literally smaller than those particles in a puff of smoke) whose surface had to be treated to give them a negative electrical charge after they had been injected intravenously.

"One of the colloids is carbon. Procedures to make a carbon colloid to meet these specifications of purity, size and charge, demand much in the field of solid state physics and physical chemistry.

"To prove that these particles all go to the RES and not elsewhere, we had to make the particles radioactive [enter nuclear physics and radiobiology] and we had to rely upon the techniques of microradioautography."

†Reproduced by permission of the New York Times.