DOCUMENT RESUME

ED 036 310 LI 001 849

| | |
|---|---|
| AUTHOR | ARTANDI, SUSAN |
| TITLE | AUTOMATIC INDEXING OF DRUG INFORMATION. PROJECT MEDICO FINAL REPORT. |
| INSTITUTION | RUTGERS, THE STATE UNIV., NEW BRUNSWICK, N.J. GRADUATE SCHOOL OF LIBRARY SERVICE. |
| SPONS AGENCY | PUBLIC HEALTH SERVICE (DHEW), WASHINGTON, D.C. NATIONAL LIBRARY OF MEDICINE. |
| PUB DATE | 70 |
| NOTE | 27P.; ED 022 504, ED 028 810, AND ED 028 811 ARE PROJECT PROGRESS REPORTS |
| EDRS PRICE | EDRS PRICE MF-$0.25 HC-$1.45 |
| DESCRIPTORS | *AUTOMATION, *COMPUTER PROGRAMS, COORDINATE INDEXES, *INDEXING, *INFORMATION RETRIEVAL, INFORMATION STORAGE, *INPUT OUTPUT ANALYSIS, STATISTICAL ANALYSIS, SUBJECT INDEX TERMS |
| IDENTIFIERS | MEDICO, *MODEL EXPERIMENT IN DRUG INDEXING BY COMPUTER |

ABSTRACT

THE BROAD OBJECTIVE OF THIS INVESTIGATION WAS TO EXPLORE THE POTENTIAL AND APPLICABILITY OF AUTOMATIC METHODS FOR THE INDEXING OF DRUG-RELATED INFORMATION APPEARING IN ENGLISH NATURAL LANGUAGE TEXT AND TO FIND OUT WHAT CAN BE LEARNED ABOUT AUTOMATIC INDEXING IN GENERAL FROM THE EXPERIENCE. MORE SPECIFIC OBJECTIVES WERE THE DEVELOPMENT, IMPLEMENTATION, AND EVALUATION OF AN INDEXING ALGORITHM WHICH WILL ENABLE THE COMPUTER TO ASSIGN AUTOMATICALLY INDEX TERMS TO DOCUMENTS. THIS FINAL PROJECT REPORT DESCRIBES THE AUTOMATIC INDEXING METHOD THAT WAS DEVELOPED IN WHICH INDEX TAGS FOR DOCUMENTS ARE GENERATED BY THE COMPUTER. THE COMPUTER SCANS THE TEXT OF PERIODICAL ARTICLES AND AUTOMATICALLY ASSIGNS TO THEM INDEX TERMS WITH THEIR RESPECTIVE WEIGHTS ON THE BASIS OF EXPLICITLY DEFINED TEXT CHARACTERISTICS. A MACHINE FILE OF DOCUMENT REFERENCES WITH THEIR ASSOCIATED INDEX TERMS IS AUTOMATICALLY PRODUCED WHICH CAN BE SEARCHED ON A COORDINATE BASIS FOR THE RETRIEVAL OF SPECIFIED DRUG-RELATED INFORMATION. A STATISTICAL EVALUATION OF THE OUTPUT OF THE INDEXING ALGORITHM AND INFORMATION CONCERNING THE SYSTEM'S ABILITY TO RESPOND TO SPECIFIC QUERIES IS GIVEN. (AUTHOR/JB)

# AUTOMATIC INDEXING OF DRUG INFORMATION

## Project MEDICO Final Report

## (LM-94 Grant)

by
Susan Artandi
Graduate School of Library Service
Rutgers, the State University

Graduate School of Library Service
Rutgers, the State University
New Brunswick, New Jersey

1970

# FOREWORD

The investigation described here was conducted under grant LM-94 from the Public Health Service National Library of Medicine.

The project was under the direction of Dr. Susan Artandi, Associate Professor, Rutgers, The State University, who was also Principal Investigator.

Other Rutgers University personnel participating in various phases of the work reported here were:

Mr. Stanley Baxendale
Associate Professor, Dept. of Computer Sciences

Dr. Edward H. Wolf
Assistant Professor, Statistics Center

Mr. Donald R. King
Associate Professor, Dept. of Computer Sciences

Mrs. Gillian McElroy
Mr. Charles W. Davis
Mrs. Ellen Altman
Research Assistants

A great deal of assistance and advice was received from Dr. Thomas H. Mott, Jr., Dean, Graduate School of Library Service, formerly Director, Center for Computer and Information Services and Chairman, Department of Computer Sciences, Rutgers University.

Extensive and valuable consultation was provided by Mr. Charles T. Meadow, National Bureau of Standards, formerly IBM Corporation, and Mr. Donald L. Dimitry, IBM Corporation.

Preceding this Final Report three progress reports were published.

Project MEDICO, First Progress Report, by Susan Artandi and Stanley Baxendale, 1968.

This report describes in detail the indexing algorithm and the indexing program. Some modifications in the program were made later and the modified version is included in the Third Progress Report.

The effectiveness of weights and links in automatic indexing. Project MEDICO, Second Progress Report, by Susan Artandi and Edward H. Wolf, 1968.

This second report describes work related to the statistical evaluation of the output of the indexing algorithm.

> Project MEDICO, Third Progress Report, by Susan Artandi and Stanley Baxendale, 1969.

The MEDICO index file, the searching method, and the search program for the automatically created index file are described in this Third report which also includes the revised indexing program.

Other publications relating to the Project:


Artandi, S. Automatic indexing of drug information. In Proceedings of the American Documentation Institute, 30th Annual Meeting, October 1967. pp. 148-151.

Artandi, S. and Edward H. Wolf. The effectiveness of automatically generated weights and links in mechanical indexing. American Documentation, 20:198-202, July 1969.

Artandi, S. Automatic indexing of medical literature--The MEDICO system. Paper presented at the III International Congress of Medical Librarianship, Amsterdam, The Netherlands, May 1969.

Artandi, S. Computer indexing of medical articles. Journal of Documentation, October, 1969.

## ABSTRACT

An automatic indexing method is described in which index tags for documents are generated by the computer. The computer scans the text of periodical articles and automatically assigns to them index terms with their respective weights on the basis of explicitly defined text characteristics. A machine file of document references with their associated index terms is automatically produced which can be searched on a coordinate basis for the retrieval of specified drug-related information. A statistical evaluation of the output of the indexing algorithm and information concerning the system's ability to respond to specific queries is given.

## TABLE OF CONTENTS

# I. SUMMARY

The broad objective of the investigation was to explore the potential and applicability of automatic methods for the indexing of drug-related information appearing in English natural language text and to find out what can be learned about automatic indexing in general from the experience. More specific objectives were the development, implementation, and evaluation of an indexing algorithm which will enable the computer to assign automatically index terms to documents.

In fully automatic indexing the computer takes the place of the human indexer. It is programmed to scan the natural language text of the document and to assign index tags to it on the basis of explicitly defined text characteristics. The human indexer no longer makes a separate judgment for each document since in automatic indexing the computer repeatedly applies the same algorithm in the indexing of each document.

Project MEDICO builds on earlier automatic book indexing research done at Rutgers.[1, 2, 3] While chemistry texts were the bases of that work, in Project MEDICO, the primary emphasis was on drug-related information appearing in English language periodical articles published in the medical literature.

In addition to assigning index terms to documents, the MEDICO algorithm was designed to compute weights automatically for the various index terms and to establish links between index terms and modifiers. The output of the indexing program is a machine searchable file of document references and their associated index tags. Each document record in the file includes the following: author, title, bibliographic citation, and index terms with their respective weights and Chemical Abstracts Registry Numbers. Although the variety of words used in natural language text is taken into consideration in indexing, the output of the indexing program utilizes a controlled vocabulary to facilitate more convenient and less ambiguous searching.

The MEDICO index file is a direct file on magnetic tape on which index records are sequenced according to document accession number. The primary access points of the file

include four hierarchical levels, and generic searches are easily implemented. Boolean searches provide for the retrieval of highly specific information. Prior to searching, the Boolean expressions corresponding to the natural language query are formulated by the human searcher. Normalization of the query to make it compatible with the index language is accomplished automatically by the computer. Several queries can be processed simultaneously and the list of references relating to each query is printed out as a separate unit.

An important aspect of the project was the statistical evaluation of the output of the automatic indexing algorithm. The statistical tests were designed to examine the validity of the assumptions which formed the bases of the indexing algorithms, with primary emphasis on those developed for the computation of weights and for the generation of links. The tests also included a comparison of the output generated from the full text of the document and from the processing of the abstracts or summaries of the same articles.

A comparison of the weights assigned to terms using the MEDICO and manual procedures gave the same value for 71 percent of the terms generated by either procedure. A moderate increase in agreement (78 percent) was observed for terms assigned a weight of 3 by at least one of the methods. Ninety-eight percent of the weights assigned by manual and machine methods were found either in agreement or to differ by a weight of no more than 1. Further investigation should demonstrate the effectiveness of weighting using two weights instead of three.

Seventy-two percent of the links generated by the full text scan of the articles in the MEDICO procedure were relevant. While writing style did not appear to have an effect on the proportion of agreements on weights for the manual and machine methods, the percentage of relevant links observed was found to be dependent on the author's writing style. The proportion of relevant links decreased as the average length of the sentences increased.

A comparison of the index terms generated from full text with those which were generated from reduced text showed that the proportion of terms indexed from reduced text is greater for those terms which had high weights in the full text analysis. Eighty-six percent of terms having a weight of 3, 46 percent of the terms having a weight of 2, and 11 percent of the terms having a weight of 1 in the full text indexing were also generated from reduced text.

Another aspect of the evaluation was concerned with the system's ability to respond to specific queries. The relatively small size of the retrieval file, and the limited scope of the subject matter and the necessity to use artificial questions placed considerable limitations on this second phase of the evaluation. While performance scores were calculated, they are not considered as satisfactory bases for broad generalizations. The qualitative results indicate that search strategy is an important factor governing the performance of the system. One of the strengths of the system is its allowance for a flexible search strategy.

The choice of computer for the project was determined by the availability of the IBM 7040 computer which was part of the installation of the Center for Computer and Information Services at Rutgers University. The main programs are written in FORTRAN with some sub-routines written in assembly language.

## References

1. Artandi, S. Book indexing by computer. Ph.D. Thesis, Rutgers, 1963.

2. Artandi, S. Mechanical indexing of proper nouns. Journal of Documentation 19:187-196, December 1963.

3. Artandi, S. Automatic book indexing by computer. American Documentation 15:250-257, October 1964.

## II. THE AUTOMATIC INDEXING METHOD

### The Indexing Algorithm

The automatic indexing algorithm developed in Project MEDICO is based on the characteristics and the position of strings of characters constituting words in English natural language text. This general approach has been characteristic of the various experimental approaches in the field because of limitations imposed by the inadequate understanding of the relationship between the meaning of text and the words which appear in it. Thus, automatic indexing algorithms have not dealt successfully with information that is implicit rather than explicit in text. Taking these limitations into consideration, the problem of defining for the computer the presence of information that should be indexed must be re-stated, at least for the time being, as follows: how to determine text characteristics in terms of the characteristics of the strings of characters appearing in it which will indicate to the computer the presence of information to be indexed and will cause the computer to take a particular action.

In the automatic indexing methods which have been developed such things as the frequency of occurrence of words, the co-occurrence of words, the relative position of words, and the pattern of the strings of characters constituting words have formed the bases of some experimental methods. The MEDICO algorithm uses location and co-occurrence as a basis for the assignment of modifiers to index terms, relative frequency of occurrence for the computation of weights, and the characteristics of string patterns and a stored dictionary for the selection of index terms.

Input to the computer consisted of the text of English language periodical articles in machine readable form and of the abstracts or summaries of the same articles.

A practical limitation was placed on the scope of the experiment by selecting test documents dealing with a particular drug group, namely anticonvulsants. For a working definition the drugs were considered anticonvulsants which were classified as such in the major drug dictionaries and in the open literature used in the compilation of the dictionary.

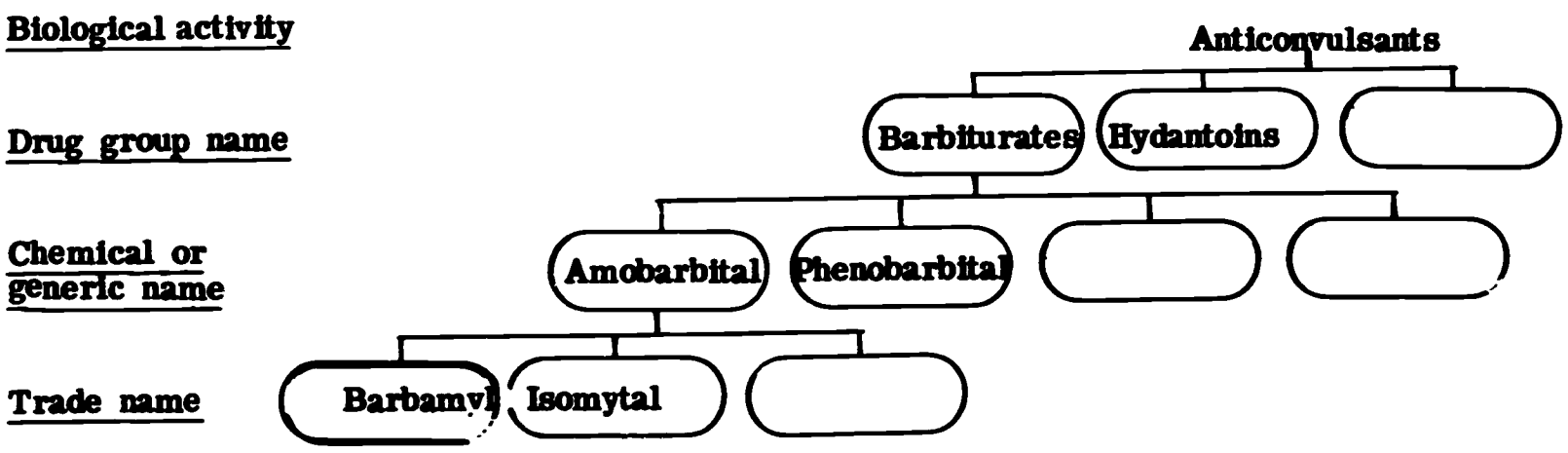
For the selection of the test documents the definition used by the National Library of Medicine was applied, since the documents were selected from the output of a MEDLARS search on anticonvulsants.

When the computer scans the document it takes into consideration the uncontrolled "vocabulary" of the text. When creating the index record for the document the computer is programmed to switch to the controlled vocabulary of the system. Thus, terms included in the stored dictionary used for the selection of index terms can be differentiated by their function and their nature. Differentiated according to their function, the dictionary includes two kinds of terms, those which are compared with the text to be indexed and those which appear in the index record.

According to their nature there are the following types of terms:

1) trade (brand, proprietary) names of individual drugs
2) chemical names of individual drugs
3) generic names of individual drugs
4) names of groups of chemical compounds
5) names of drug groups according to activity
6) terms which are other than names of drugs or chemicals.

The hierarchical relationship which exists between the type of terms included in (1) to (5) is utilized in the indexing algorithm to provide for automatic generic posting. This results in an index record which allows access to the information at as many as four hierarchical levels.



Terms in the dictionary may be single or multiple word terms and their length is not limited. Compilation of the dictionary for the project required the identification of those drugs which belong to the group of anticonvulsants and the selection of those non-drug terms judged to be of indexing value. While a great deal of effort was made to make the list of anticonvulsants as complete as possible and to establish equivalencies between chemical, generic, and trade names, no claim is made that the list is all inclusive; since the primary concern of the project was to demonstrate the feasibility of the indexing method.

The principal functions of the dictionary are:

1) to select index terms on the basis of explicitly defined text characteristics

2) to control the index language of the index records

3) to assign "packages" for inclusion in the index record.

Packages provide for vocabulary control and for access at various hierarchical levels. With each dictionary term a standard package is associated consisting of those terms which will appear in the index record whenever the particular dictionary term appears at least once in the text being indexed.

The package for dictionary terms which are chemical or generic names of drugs usually includes the following: preferred chemical name, preferred generic name, chemical group name and the Chemical Abstracts Registry Number. For dictionary terms which are trade names the package includes the trade name in addition to the above. For non-drug terms the package consists of the preferred synonym or word form. For example, if pentobarbitone sodium would appear in the document text, the following corresponding package would be recorded in the index record:

pentobarbital sodium

sodium 5-ethyl-5-(1-methylbutyl)barbiturate

barbiturates

57330

The same package would be generated if sodium 5-ethyl-5-(1-methylbutyl)barbiturate appeared in the text. If, however, Nembutal appeared in the text, it would be added to the index record because it is a trade name.

Inclusion of trade names in packages provides for very specific information access

points to facilitate the retrieval of documents about a particular pharmaceutical product. This practice also allows for the system to act as a data retrieval system, since the tracings in the output record include the chemical compound which corresponds to the particular trade name that is indexed.

## Weights and Links

As indicated earlier, the indexing algorithm was designed to compute and to assign weights to index terms and to generate links between index terms and modifiers automatically.

Weighting means the assignment of a value to a term to indicate the relative importance of that term in the subject description of the document, in a query, or in a user profile. Thus, an index term which represents a central theme in the document gets a high weighting and one which represents only a marginal element in the subject gets a low weighting. Links indicate particular connection between terms where the lack of such a link may create ambiguity. Both weights and links increase the specificity of terms. They are precision devices because, by increasing the specificity of the term, non-relevant documents are not retrieved.

Research related to weights and links has been largely centered around systems in which they are assigned by the human indexer on the basis of the indexer's intellectual judgment. The algorithm developed in Project MEDICO for the automatic assignment of weights is based on the assumption that the relative frequency of occurrence of terms in text can serve as an indication of the importance of the subjects the terms represent. Relative, rather than absolute frequency, is used to compensate for differences in length among articles. The computer calculates the number of occurrences per thousand text words and converts the resulting figure into a weight in the following manner: If the frequency of the term per thousand text words is less than or equal to 1, the document is assigned a weight of 1. If the frequency of the term per thousand text words is greater than 1 and less than 3, the term is assigned a weight of 2. Finally, if the frequency of the term per thousand text words is greater than or equal to 3, a weight of 3 is assigned to it.

The automatic generation of links is based on the assumption that co-occurrence within a sentence is a satisfactory indication that the terms belong together within the context of the document. Because the test documents were medical articles, and because the emphasis was

on drug-related information, links were created primarily between names of drugs or chemicals and terms which could act as modifiers, such as therapy, toxicity, administration, dosage, etc.

The following are examples of sentences from which valid links between drug names and modifiers were automatically generated.

"In the review of English literature of the past ten years the author has found no other reference to the occurrence of hyperglycemia following the administration of diphenylhydatoin in humans."

"It was therefore estimated that he received a total dosage of at least 800 mg. of diphenylhydantoin in a 24 hour period or approximately 70 to 80 mg. per kilogram."

The first sentence generated the index term diphenylhydantoin/administration and the second generated the term diphenylhydantoin/dosage.

## New Terms and Dictionary Updating

Inherent in dictionary-based indexing is the risk of missing unknown new information, information too new to be included in the dictionary. The problem is to devise some method by which a new anticonvulsant, for example, could be indexed when it is first reported in the literature. "Catching" of new indexable information is also important from the point of view of dictionary updating. While it is possible to select new terms to be added to the dictionary manually, it is desirable to make the process at least semi-automatic.

One approach to the design of algorithms which will enable the computer to find drug names in text is to determine some common characteristics which names of drugs have, characteristics which will sufficiently distinguish drug names from other terms. Some characteristics which have been identified in case of drug names are such things as their length; an alternating string pattern of numbers, letters, and dashes; the capitalization of registered trade names followed by a capital R (words which begin and end with an upper case letter); the presence of such words as ethyl, methyl, etc., or the presence of Greek letters in chemical names.

For example, the indexing program selects terms on the basis of their length to complement the dictionary method. Strings of characters exceeding 18 characters and not

contained in the dictionary are put out for visual inspection. Those terms which are judged to be useful index terms are used in two ways:

1) they are added to the index record of the document which generated them
2) they are used to update the dictionary.

## III. SEARCHING THE MEDICO SYSTEM

The immediate output of the indexing program is the MEDICO index file. It was essential to design this file to be satisfactory for search purposes and at the same time to be capable of producing a printout with a format that is convenient to use.

The MEDICO index file is a direct file of document records on magnetic tape arranged in document accession number order. The principal characteristic which distinguishes this file from many other document retrieval files is that its content and format is automatically generated by the computer from natural language text.

The record for each document in the MEDICO index file consists of the following elements:

author

title

bibliographic citation

total number of text words

index terms without modifiers, with their respective weights and Chemical Abstracts Registry Numbers

index terms with modifiers, with their respective weights

The fact that each record shows all index terms which were assigned to the document is useful because the terms taken together add up to a rudimentary abstract. While they are not substitutes for good informative abstracts prepared by good human abstractors, they do provide a certain degree of informativeness for the user.

Figure 1 shows a sample printout of an index record generated from full text, and Figure 2 shows the index record for the same document generated from reduced text.

Since the MEDICO file is a direct file, each record stands for a single document as opposed to an inverted file in which each record stands for a single index term. Inherent in the process of searching a direct file for documents specified by subject is the need to make a complete scan of the entire file for each query to be processed. The capability for simultaneous searches, processing several queries in a single pass of the tape, can compensate for this limitation.

GILBERT JC, ORTIZ WR, MILLICHAP JG
THE EFFECTS OF ANTICONVULSANT DRUGS ON THE PERMEABILITY OF BRAIN CELLS TO D-XYLOSE.
J NEUROCHEM 13,247-55, APR 66

NO. OF WORDS = 3338
(3) ANTICONVULSANTS

50066 (3) 5-ETHYL-5-PHENYLBARBITURIC ACID, PHENOBARBITAL, BARBITURATES

695534 (3) DIMETHADIONE, 5,5-DIMETHYLOXAZOLIDINE-2,4-DIONE, OXAZOLIDINEDIONES

57410 (3) DIPHENYLHYDANTOIN, 5,5-DIPHENYL-2,4-IMIDAZOLIDINEDIONE, HYDANTOINS

127480 (1) TRIMETHADIONE, 3,5,5-TRIMETHYL-2,4-OXAZOLIDINEDIONE, OXAZOLIDINEDIONES

ANTICONVULSANTS/ EFFECT (1), ACTIVITY (1)
PHENOBARBITONE/ EFFECT (2), THERAPY (1), ADMINISTRATION (1), ACTIVITY (1)
DIPHENYLHYDANTOIN/ ACTIVITY (1), EFFECT (2)
DIMETHADIONE/ ACTIVITY (1)
TRIMETHADIONE/ THERAPY (1)

END OF ARTICLE

**FIGURE 1**

GILBERT JC, CRTIZ WR, MILLICHAP JG
THE EFFECTS OF ANTICONVULSANT DRUGS ON THE PERMEABILITY OF BRAIN CELLS TO D-XYLOSE.
J NEUROCHEM 13,247-55, APR 66

NO. OF WORDS = 149
SC066 (3) 5-ETHYL-5-PHENYLBARBITURIC ACID, PHENOBARBITAL, BARBITURATES

695534 (3) DIMETHADIONE, 5,5-DIMETHYLOXAZOLIDINE-2,4-DIONE, OXAZOLIDINEDIONES

57410 (3) DIPHENYLHYDANTOIN, 5,5-DIPHENYL-2,4-IMIDAZOLIDINEDIONE, HYDANTOINS

(3) ANTICONVULSANTS

DIPHENYLHYDANTOIN/ EFFECT (3)
ANTICONVULSANTS/ EFFECT (3)

END OF ARTICLE

FIGURE 2

Searching is essentially the reverse of indexing, and the preparation of a search instruction involves procedures and sources of errors that are very similar to those encountered in indexing. The objective of searching is to identify those documents whose content is relevant to the query.

The output of a search may be viewed as the result of the relevance judgment of the system. Theoretically, the closer this resembles the relevance judgment of the user the better the system performs. In practice, however, the problem is not quite as clearcut; and factors influencing both system and user judgment need to be taken into consideration.

In addition to utilizing the primary access points in the index record, the MEDICO search program is designed to make possible complex Boolean searches using the connectives AND, OR, and NOT. Any data element can be combined with any other included in the subject description part of the index record, involving the capability for many more possible combinations than would be needed in practice.

While the Boolean expression corresponding to the query is formulated by the human searcher, normalization of the search terms to make the query compatible with the index record is accomplished automatically by the computer. The index file is searched sequentially to find terms as prescribed in the Boolean expression. Several queries can be searched simultaneously and the output relating to each query can be printed out as a separate unit. The following are examples of queries which were used in testing the system.

What is the use of 3, 5, 5-trimethyloxazolidine-2, 4 -dione as an anticonvulsant?

What kind of toxic side effects can be expected when Tegretol is administered?

What dose of trimethadione is used in the treatment of epileptics?

## IV. EVALUATION OF THE INDEXING ALGORITHM

A major part of the evaluation work was concerned with the statistical evaluation of the output of the MEDICO automatic indexing algorithm. The statistical tests were intended to examine the validity of the assumptions which were the bases of the indexing algorithm with primary emphasis on the methods which were developed for the computation of weights and the generation of links. Also included in the tests was a comparison between output generated from the full text of the article and from the processing of the abstracts or summaries of the same documents.

Evaluation of the output of the indexing involved essentially a comparison between the judgment of the human indexer and that of the indexing algorithm to see the proportion of agreement between the manual and the machine method. The documents were examined by a human indexer to check the correctness of the weights appearing in the output of the automatic indexing program when intellectual judgment rather than an automatic method was used. The links in the same output were checked to determine whether they were correct in the context of the document.

When the method of assigning weights was used, the automatic and manual methods were said to agree when both assigned the same weight to a particular index term. A comparison of the weights assigned to terms using the MEDICO and manual procedures gave the same value for 71 percent of the terms generated by either procedure. A moderate increase in agreement (78 percent) was observed when only terms having a weight of 3 by at least one of the methods were considered. Ninety-eight percent of the weights assigned by the two methods was found to be either in agreement or to differ by a weight of 1. These findings tend to suggest that allowing only two weights in the system instead of three would perhaps increase the proportion of agreement. However, more research is needed to examine the validity of this assumption.

In the evaluation of the linking procedure, the purpose of the test was to (1) determine the proportion of relevant links, (2) determine whether writing style has an effect on the

proportion of relevant links, and (3) consider criteria other than co-occurrence within a sentence in defining a link.

Seventy-two percent of the links generated by a full text scan of the articles by the MEDICO procedure were relevant. While writing style does not appear to have an effect on the proportion of agreements on weights for the two methods, the percentage of relevant links observed was found to be dependent on the author's writing style. The proportion of relevant links decreased as the average length of the sentences increased. This suggests that it may be desirable to change the definition of a link from co-occurrence within a sentence to co-occurrence within two punctuation marks. A preliminary check of the 15 articles studied showed that the number of relevant links would increase to 84 percent as a result of such a re-definition. Further research is needed to study this assumption in detail.

Evaluation also included a comparison between the output generated from full text with output from abstracts or summaries of the same articles.

This type of comparison is important because there is reason to believe that the difference in cost between the two methods may be considerable. Data relative to the effectiveness of the two methods should give some indication of the degree of improvement that can be expected for the additional expense involved in full text processing.

The purpose of the comparison between the two kinds of output was to evaluate the following questions:

(1) Does an abstract provide a better index than a summary paragraph?

(2) Do the terms which have a weight of 3 in the full text index appear more frequently in the reduced text index than terms which have a weight of 2 or 1?

(3) Can we consider the link distance or some function of the average link distance in order to segregate relevant and irrelevant links?

Since no statistically significant difference was found in the two forms of reduced text, abstracts and summaries were considered together in the evaluation.

Because the average number of words in the reduced text was 127, it was impossible to rank the importance of terms indexed by assigning weights. It must be assumed that any term mentioned in the reduced text is important. This means that terms having a weight of 3 in the full text index should appear more frequently in the reduced text index than terms which have

a weight of 2 or 1.

A comparison of the index terms generated from full text with those which were generated from reduced text confirmed this hypothesis. The test showed that the proportion of terms indexed from reduced text is greater for those terms which had high weights in the full text analysis. Eighty-six percent of terms having a weight of 3, 46 percent of the terms having a weight of 2, and 11 percent of the terms having a weight of 1 in the full text analysis were also generated from reduced text.

The average of the relevant link distance in reduced text was 3.4 and the average irrelevant link distance was 8.9 words. The irrelevant link distances were quite large when compared to the lengths of the relevant links (11, 18, 11, 14). This would indicate that a significant reduction in the frequency of irrelevant links might occur by defining a link as the co-occurrence within a predetermined distance. For example, suppose we consider only links of 10 words or less in distance. This criterion would yield 23 links of which 20 are relevant. The upper limit of 10 was obtained by considering the average of the relevant link distances (3.4) plus 3 standard deviations of the relevant link distances (2.16).

It should be pointed out that this observation about link distance is based on a small sample size (27). Previous data from the full text analysis indicate that the average of the relevant link distances is dependent on the author's writing style. However, it might happen that all authors write in a more concise form in a summary paragraph or an abstract. Future investigations with larger sample sizes should shed more light on this observation.

## V. EVALUATION OF THE OUTPUT OF THE SYSTEM

The evaluation of the output of the MEDICO index file was concerned with the system's ability to respond to specific queries. Considerable limitations were placed on this phase of the evaluation work by the small size of the index file and by the limited scope of subject matter. Both of these factors made it necessary to use artificial questions which in turn introduced further limitations. The test results which are presented here should be interpreted in the context of the limitation just outlined and should not be considered as figures which can form the bases of broad generalizations.

While for the reasons just explained the usefulness of the quantitative results given in this section of the report is somewhat limited, an examination of the qualitative results do provide some useful insights for future research.

Most of the precision failures were due to the inclusion of terms in the article's index record which represent peripheral subjects, resulting in high exhaustivity in indexing. This situation is to some extent inherent in the indexing method. However, it can be controlled through the use of weights in indexing and through the proper utilization of weights in the search formulation to achieve the optimum level of generality for a given query. In contrast with these precision failures, faulty search strategies, causing recall failures, involved search formulations which were too specific. An indexing problem causing recall failures was due to the lack of the actual occurrence of a term in text, which, however, was implied.

On the whole, search strategy emerged as a very important factor governing the performance of the system and the capability to allow for a flexible search strategy emerged as an important strength of the system. The ability to vary the search strategy through the use of links, weights, and search logic illustrates the flexibility of the system. It is possible to search using any version of a drug's name (chemical, generic, or trade name) because of the use of packages and the "normalization" of terms. Query 2, for example, causes five articles to be retrieved although only one of the sample articles in the file contained the

chemical name used in the question. All articles mentioning that drug in any form of its name were retrieved.

Questions involving the linking of terms are also automatically "normalized" by the system. Query 3 calls for articles on "barbiturates used in the treatment of convulsive disorders." Any article on barbiturates which linked anticonvulsants with the words: treat, treated, treatment, therapy, therapeutic, or therapeutic effect, can be retrieved.

Because of the use of packages, requests for articles on a class of drugs (barbiturates, hydantoins, etc.) will yield articles on any drug in the class, although the class may not be mentioned in the article. This capacity also enables the searcher to use a logical not strategy (such as in Query 1) and thus eliminate large numbers of non-relevant documents on retrieval.

The measures used in the evaluation were the recall ratio and the precision ratio. While these ratios have been used extensively for the measurement of system performance, it should be remembered that their application requires considerable subjective judgment on the part of the evaluator to determine the relevancy of a document to a query.

The recall ratio equals the number of relevant documents retrieved over the number of known relevant documents in the collection times 100. The precision ratio equals relevant documents retrieved over the total number of documents retrieved times 100. When the two measures are applied together they should indicate what proportion of the total number of relevant documents in the collection has been retrieved and at what cost, in terms of noise, a particular performance was achieved.

Twelve queries were used to search the two MEDICO index files generated from the indexing of full text and of reduced text, respectively. The average ratios were calculated by averaging the appropriate individual ratios, and the relevance judgments were made by a single individual.

## Index File Generated From Full Text Indexing

The average recall and precision ratios which follow are based on a set of twelve queries which were searched in a file generated from the automatic indexing of the full text

of 30 articles.

| | |
|---|---|
| Over-all precision ratio | 78 percent |
| Over-all recall ratio | 76 percent |

Recall failures occurred in six and precision failures occurred in four of the output of the twelve searches which were analyzed. There were eleven articles which were relevant but not retrieved. Out of the eleven unretrieved articles three were not retrieved because of inadequate indexing and eight were missed because of faulty search strategy. Indexing problems which caused recall failures were due to the following factors:

-- lack of occurrence of a specific term in the text, although implied;

-- lack of a link because of the non-occurrence of a term in a given sentence; and

-- the presence of an incorrect weight.

Faulty search strategies generally involved search formulations which were too specific, demanding too many parameters to satisfy the query.

Precision failures involved 21 articles. Out of these 21 non-relevant articles nine were retrieved because of inadequate indexing and twelve were retrieved because of faulty search strategy.

Indexing problems causing precision failures were of the following kind:

-- mention of a term in text which is only peripheral to the subject of the article causing high exhaustivity in indexing;

-- the presence of an incorrect link; and

-- the level of generality of the terms involved.

Faulty search strategies causing precision failures generally involved search formulations which were too broad, resulting in high recall.

## Recall Failures

| Cause of failure | Number of missed articles | Percent of total missed articles | Number of searches involved | Percentage of total searches |
|---|---|---|---|---|
| Indexing | 3 | 27 | 3 | 25 |
| Searching | 8 | 73 | 3 | 25 |
| Total | 11 | | 6 | 50 |

## Precision Failures

| Cause of failure | Number of articles involved | Percent of total articles | Number of searches involved | Percent of total number of searches |
|---|---|---|---|---|
| Indexing | 9 | 43 | 2 | 16.6 |
| Searching | 12 | 57 | 2 | 16.6 |
| Total | 21 | | 4 | 33.3 |

## Comparison of Full Text and Reduced Text

The performance tests included a comparison between index files generated from the full text and from the reduced text of the same documents. The same twelve queries were searched in two smaller files corresponding to fifteen documents. The following figures were obtained:

Summary of Average Recall and
Precision Ratios for 12 Searches

| | Index file from full text | Index file from reduced text |
|---|---|---|
| | (percent) | (percent) |
| Average precision ratio | 81.8 | 50 |
| Average recall ratio | 82.5 | 40.3 |

## Data Retrieval

In addition to the twelve queries used in the evaluation tests, several queries were used to move the system's ability to retrieve data as distinguished from the retrieval of documents. For example, the response to the query:

"What is the chemical name and Chemical Abstracts registry number of the succinimides which are used as anticonvulsants?" is a list of records containing the chemical names and Chemical Abstracts registry numbers of drugs of the succinimide group. These names and numbers are not necessarily contained in the articles to which the various records refer, but the information is contained in the immediate output of the system. In this instance the system acts as a data retrieval system, and there is no need to consult any documents to find the answer to a query.

The same is true of the query, "What is the chemical name of Sulthiame?" because the chemical name will be contained in the index record of an article on Sulthiame, although the article itself may not include a mention of the chemical name.

## List of Queries

### Query 1

Drugs which are active as anticonvulsants but which are not of the barbiturate, hydantoin, or succinimide family.

### Query 2

The use of 3, 5, 5-trimethyloxazolidine-2, 4-dione as an anticonvulsant.

### Query 3

Barbiturates, with the exception of amobarbital, which are used in the treatment of convulsive disorders.

### Query 4

What hydantoins, other than phethenylate, can be used in anticonvulsant therapy?

### Query 5

What kind of toxic side effects can be expected when Tegretol is administered?

### Query 6

The effectiveness of diphenylhydantoin as a therapeutic agent and the dosage that is recommended.

### Query 7

The dosage of trimethadione used in the treatment of epileptics.

### Query 8

The administration of diphenylhydantoin sodium in the treatment of trigeminal neuralgia.

### Query 9

Articles which have as a central topic the use of Primidone as an anticonvulsant. (Weight 3 for both Primidone and anticonvulsant)

### Query 10

Same as Query 9 with a weight of 2 for anticonvulsants.

### Query 11

Articles dealing with oxazolidinediones.

### Query 12

Articles on barbiturates.