Automatic Indexing of Medical Literature Using Phrase Matching -An Exploratory Study

Hayati Abdul* and Christopher Khoo**

Abstract: *This study sought to find out to what extent phrase matching could be used to automatically assign MeSH headings and subheadings to abstracts of journal articles. A phrase matching program was written using Turbo Prolog. The program assigned a MeSH heading if the heading or one of its "see" references was found in the abstract. The program also used a database of manually constructed phrase matching rules to assign subheadings. This study was limited to Category C8 MeSH terms only. The program was run with 200 abstracts taken from MEDLINE. The automatically assigned heading/subheadings were compared with MEDLINE indexing, and indexing problems encountered by the program were identified. Our results suggested that the program would be able to pick up most of the MEDLINE-assigned major headings (central concepts) and would assign few incorrect headings if the program was extended in 2 ways: a) syntactic and/or semantic analysis was incorporated to allow the program to effectively distinguish central concepts from incidental ones; b) the MeSH thesaurus and "see" references were supplemented with a synonyms list. The results also indicated that our phrase matching rules for assigning subheadings needed to be extended in a number of ways.*

INTRODUCTION

THIS reports the first part of a study to explore the feasibility of developing a program to automatically assign MeSH headings and subheadings to abstracts of journal articles.

In the first part of the study, we sought to find out to what extent phrase matching could be used to automatically index journal abstracts, and to identify the problems encountered by the phrase matching approach. In subsequent work, we shall try to overcome these problems using natural language processing techniques.

We limited the study to Category C8 (Respiratory tract diseases) of the MeSH

*Hayati Abdul B.A., Dip. Lib. Inf. Sci. Medical Library, National University of Singapore
**Christopher Khoo B.A., M.Sc. (Lib. & Inf. Sci.) Science Library, National University of Singapore

thesaurus. A phrase matching algorithm was implemented using Turbo Prolog and run on an IBM compatible microcomputer. Category C8 MeSH terms and their "see" references were entered into a database of phrase matching rules. The program was used to assign headings and subheadings to 200 abstracts. The abstracts were the last 200 abstracts in the 1988 MEDLINE on CD-ROM that had one or more Category C8 headings in the major or the minor descriptor field. The heading/subheadings assigned by our indexing program were compared with MEDLINE indexing.

Section 2 discusses the assignment of headings. The assignment of subheadings is discussed in Section 3.

ASSIGNMENT OF HEADINGS

The Algorithm

The indexing program assigns a heading if the heading or any of its "see" references appear in the title or in the abstract. The phrase matching algorithm takes a loose definition of what a phrase is. A phrase (such as a MeSH term) is considered to be present in a record if words in the phrase (excluding stopwords) occur anywhere, in any order within a sentence. Moreover, the words need not be adjacent in the sentence. For example, "neoplasm of the lung" would retrieve the heading *Lung neoplasms.*

The program observes MEDLINE's specificity rule in preferring narrower terms to broader ones. Section 20.3 of the *MEDLARS Indexing Manual Part II* (Charen, 1983) states: "Index a concept under a specific term but not also under a general concept." After completing the phrase matching, the indexing program examines each retrieved MeSH term to see if a narrower term is also retrieved. If so, the broader term is discarded.

The program applies a small stoplist to eliminate "with" and "of" from the MeSH terms. The program also uses a simple stemming algorithm and a small synonyms file. When constructing the stemming rules, care was taken to ensure that they covered most of the words used in the indexing. Words that cannot be handled by the stemming rules are handled by the synonyms file. The synonyms file also handles exceptions to the stemming rules. A word is not stemmed if it is found in the synonyms file.

A trial run on a small sample revealed 2 main problems. Synonyms of "neoplasm" and "pulmonary" substantially affected the effectiveness of the automatic indexing. The following rules were added to the synonyms file:

neoplasm --> cancer
malignant --> cancer
tumor --> cancer
carcinoma --> cancer
melanoma --> cancer
pulmonary --> lung

The trial also revealed problems with very broad headings like *Lung diseases* and *Respiratory tract diseases*. The concept of "disease" was sometimes expressed in an indirect way. The word "disease" often did not appear in the abstract. The words "disease" and "tract" were included in the stoplist to improve retrieval of the broad MeSH terms. This meant, however, that any mention of "lung" would be assumed by the program to refer to a diseased lung.

Results

The results of running the program with 200 abstracts are given below. Subheadings are ignored in the analysis, and headings from only Category C8 are included. If a heading appears in both a major descriptor (printed Index Medicus heading) and a minor descriptor (non-Index Medicus heading), the minor descriptor is ignored. (This occurs when 2 subheadings are assigned to a heading, and one of them is designated a minor subheading.)

Major descriptors (Central concepts/Printed Index Medicus headings)
192 headings from Category C8 were assigned by MEDLINE.
131 (68%) were picked up by the indexing program.

Minor descriptors (Non-Index Medicus headings)
85 headings were assigned by MEDLINE.
40 (47%) were picked up by the indexing program.

Major and minor descriptors combined
62% of MEDLINE-assigned headings were picked up by the indexing program.

Automatic indexing
268 headings were assigned by the indexing program.
169 (63%) coincided with MEDLINE-assigned major or minor headings.

We can use Hooper's measure of indexing consistency to measure how close our machine indexing is to MEDLINE's manual indexing. For 2 indexers, indexer A and indexer B, indexing the same article, the indexing consistency (in %) is given by

$$\frac{100\ \text{X}\ \text{No. of headings common to both indexing}}{\text{No. of headings assigned by A} + \text{No. of headings assigned by B} - \text{No. of headings common to both}}$$

If we consider the indexing program and MEDLINE as two indexers, then the average indexing consistency over 200 articles is 56%.

In contrast, human indexers have been found to have an indexing consistency ranging from only 39% to 48%. Funk et al. (1983), using 760 articles inadvertently indexed twice by MEDLINE, found an average indexing consistency of 48.2%. Lancaster (1968, p. 178-180), in his study involving 16 articles and 3 indexers, found an average indexing consistency of 46.1%. Leonard (1975), using 100 articles indexed 5 times, obtained an indexing consistency of 48.2%. (Both Lancaster's and Leonard's results include checktags.) Marcetich and Schuyler (1981), using 50 articles and 4 indexers, obtained a figure of 43% for computer-assisted indexing and 39% for routine indexing.

In view of the rather low indexing consistency among human indexers, the 56% agreement between our automatic indexing and MEDLINE indexing seems too good to be true! It must be noted that we limited our sample to records that MEDLINE had already assigned a Category C8 heading, and we excluded non-Category C8 headings when computing Hooper's measure. Moreover, the kinds of errors that machine indexers make are probably different from human errors.

We did a small test to see if the indexing program would incorrectly assign headings to articles not dealing with respiratory tract diseases. The program was run with 100 records that were not assigned a Category C8 heading by MEDLINE. The program assigned a heading to 4 abstracts. *Influenza* was assigned to 2 abstracts that dealt with an influenza virus. *Lung diseases* was assigned to 2 other abstracts because lung disease was mentioned.

Reasons Why Some MEDLINE-Assigned Headings Were Not Retrieved

Reasons	No. of Major Descriptors	No. of Minor Descriptors
Broader term not assigned	21	9
Synonyms not in "see" reference	16	7
Problem with multi-concept heading	10	2
Concept not mentioned in the abstract	8	25
Problem with very broad headings	3	1
Stemming problem	3	1

Figure 1. Reasons MEDLINE-assigned headings were not retrieved.

Figure 1 gives a summary of the reasons why MEDLINE-assigned headings were not picked up by the indexing program. The reasons are discussed below.

Broader term not assigned.

Of the major headings not picked up by the indexing program, about 30% were not assigned because narrower terms were assigned by the indexing program. Following the specificity rule, the program does not assign a broad heading if a narrower heading is picked up. Sometimes, however, the narrower concept is only a minor aspect of the article. The broad heading may, in such cases, be the appropriate heading to assign. The program cannot distinguish between concepts that are central to the article and concepts that are incidental.

To complicate matters, the broad and the narrower concepts can sometimes both be important in an article. MEDLINE assigns both a broad and a specific heading if the broad and the specific concepts are both substantially discussed. Section 20.4 of the MEDLARS Indexing Manual Part II states: "Index an article on both a general concept and a specific one if the article is on both." Section 23.6.1 also specifies that if several specific terms under a broader term are treated by the article, the indexer may want to assign the broader term as a major descriptor and the specific terms as minor descriptors.

Synonyms not in MeSH "see" references

The program uses a very small synonyms list to handle very common synonyms. We intend to extend the synonyms file by entering synonyms from a medical dictionary.

Problem with multi-concept headings

Many MeSH terms, like *Lung neoplasms,* are pre-coordination of 2 concepts. The 2 concepts may not appear in the same sentence. "*Lung*" and "*neoplasm*" may occur in different sentences in the same abstract. Moreover, 1 or both of the component concepts may themselves be broad MeSH terms. Lung and Neoplasms are MeSH headings in their own right. In such cases, a narrower term (e.g. "ENDOMETRIOSIS") may occur instead of the broad term (Neoplasms). The program has to be able to recognise that "ENDOMETRIOSIS" is a neoplasm. If the whole MeSH thesaurus is available to the program, it can check whether a Category C4 (Neoplasms) heading has been picked up. A rule can be constructed to assign *Lung neoplasms* if the word "lung" occurs and a C4 heading is picked up.

Not mentioned in the abstract

Our results support the notion that most (96%) of the major headings (central concepts) can be determined from the title and the abstract.

Problem with very broad headings

Concepts near the top of the MeSH tree, like *Lung diseases* and *Respiratory tract diseases* can be difficult to pick out because they are "soft" concepts. They can be expressed in many ways and the MeSH term often does not occur. Moreover, these concepts are sometimes expressed indirectly or merely implied.

Examples:

TITLE	DESCRIPTOR IMPLIED
Endoscopy of the airway in infants and children.	Lung diseases/diagnosis
Asbestos exposure - cigarette smoking interactions among shipyard workers	Respiratory tract diseases/etiology

Reasons Incorrect Headings Were Assigned By The Indexing Program

Figure 2 gives a summary of the reasons incorrect headings were assigned by the indexing program.

Reasons	No. of Headings
Concept mentioned but not central to the article	55
Problem with stopword	15
Narrower term not picked up	14
Stemming problem	7
Wrong association between words in different parts of a sentence	5
Nearly similar headings not having BT-NT relation	3

Figure 2. Reasons incorrect headings were assigned by the program.

Concept mentioned but not central to the article.

55 (56%) of the headings incorrectly assigned by the program were concepts that were mentioned in the abstracts but were really incidental to the articles. 11 of these headings were immediately narrower terms of headings assigned by MEDLINE.

Problem with stopword

The stopwords "disease" and "tract" caused some headings to be incorrectly retrieved.

Examples:

PHRASES	HEADINGS INCORRECTLY ASSIGNED
bronchial washings	Bronchial diseases
pulmonary artery shunt	Lung diseases ("Lung" is a synonym of "pulmonary")
pulmonary interstitial emphysema	Pulmonary fibrosis (which has a "see" reference from *Lung disease, Interstitial*)

Narrower term not picked up

14% of the incorrectly assigned headings were assigned because the narrower terms assigned by MEDLINE were not picked up by the program. So the broader terms were assigned instead.

Stemming problem

Stemming causes certain assumptions to be made by the program. For example, "asbestos" always retrieves the heading *Asbestosis*, and "H. influenzae" retrieves

Influenza. The assumptions do not always hold.

Wrong association between words in different parts of a sentence.

A heading is retrieved even if its component words appear in different parts of a sentence and are only distantly related to one another. Consider the following sentence:

"We studied lungs at autopsy from 40 patients with cystic fibrosis ..."

In addition to *Cystic fibrosis*, the program also assigned *Pulmonary fibrosis* because the words "fibrosis" and "lungs" (synonym of "pulmonary") occurred in the sentence.

Nearly similar headings not having Broader Term-Narrower Term relation

A MeSH term (e.g. *Pneumonia*) can occur within another term (e.g. *Pneumonia, Mycoplasma*). The indexing program relies on the Broader Term-Narrower Term (BT-NT) relation to eliminate one of the terms if both are retrieved. *Pneumonia* is never assigned together with *Pneumonia, Mycoplasma* because the program eliminates the broader term Pneumonia. Some pairs of nearly similar headings do not have a BT-NT relation:

Respiratory distress syndrome and *Respiratory distress syndrome, Adult*
Rhinitis and *Rhinitis, Allergic, Perennial*

Whenever *Respiratory distress syndrome, Adult* is assigned, the program also assigns *Respiratory distress syndrome.*

ASSIGNMENT OF SUBHEADINGS

The Algorithm

The program uses a database of phrase matching rules for assigning subheadings. Rules for assigning subheadings are of the following types:

a. If phrase A occurs at least N times, then assign subheading SH.

Examples:
If "biopsy" occurs at least 2 times, then assign subheading PA (Pathology).
If "death rate" occurs at least 1 time, then assign subheading MO (Mortality).

b. If at least N of the phrases in the following set [A,B,C,...] occur, then assign subheading SH.

Examples:
If 2 of the following phrases ["physiology", "pathology"] occur, then assign subheading PP (Physiopathology).
If any of the following phrases ["preoperative", "postoperative", "surgery", "surgical", "resection", "operative"] occur, then assign subheading SU (Surgery).

c. If at least N of the assigned headings are in any of the following MeSH sub-trees [tree no. A, tree no. B, ...], then assign subheading SH.

Example:
If at least 1 of the headings are in the MeSH sub-tree D26.394, then assign subheading DT (Drug therapy).

Rules of the type C above could not be used in this study because we were limited to Category C8 headings.

The rules were formulated by one of the authors (Hayati) with the help of a small sample of abstracts for each subheading. The program was run with a trial sample of 50 abstracts and the rules refined. Although the rule base was incomplete and needed much more refinement, we felt it was good enough to help us identify the main problems of the phrase matching approach. We plan to refine the rules in the next part of the project using word frequency and word association analyses.

The program assigns the subheadings retrieved to all the headings. The indexing program is at present unable to determine which subheading should be attached to which heading.

Results

In the analysis, each heading/subheading combination is considered as one descriptor. Descriptors containing the subheadings AN, BS, CY, SC, SE and UL are ignored. Most of these subheadings are applicable only to Category C4 (Neoplasms) headings. A few Category C4 headings (e.g. "Lung neoplasms") also appear in Category C8.

The results are summarized below:
190 major heading/subheading combinations were assigned by MEDLINE.
77 (41%) of these were retrieved by the indexing program.
60 of the descriptors were not retrieved because the headings were not assigned by the indexing program.

486 descriptors (major and minor) were assigned by MEDLINE.
146 (30%) were retrieved by the indexing program.
104 of the descriptors were not retrieved because the headings were not assigned by the indexing program.

785 descriptors were assigned by the indexing program.
145 (18%) coincided with MEDLINE indexing.
320 of the automatically assigned descriptors were incorrect due to incorrect headings.

Of the remaining 320 incorrectly assigned descriptors, 15 were due to the subheading being attached to the inappropriate heading.

A large number of the heading/subheading combinations assigned by the indexing program were incorrect because of incorrect headings. Since the program assigns the subheadings to all the headings, one incorrect heading can result in many incorrect heading/subheading combinations.

It is interesting to note that very few of the indexing errors were due to the subheading being attached to the wrong heading. This suggests that it is not necessary for the indexing program to be able to determine which subheading belongs with which heading.

The average indexing consistency between MEDLINE and our indexing program was only 13%. Funk et al. (1983) obtained a figure of 33.8% for human indexers, Lancaster (1968) obtained 34.4% (including checktags) and Leonard (1975) had 36.5% (including checktags).

Indexing Problems

100 of the abstracts were visually analyzed to identify the main indexing problems.

Problem with "soft" concepts

Many of the concepts expressed by the subheadings are "soft" concepts in the sense that they can be expressed in many different ways. The concepts are sometimes indirectly expressed or merely implied.

Examples:
"Growth failure in bronchopulmonary dysplasia." (Bronchopulmonary dysplasia/ PHYSIOPATHOLOGY)

"The development of independence in adolescents with cystic fibrosis." (Cystic fibrosis/PSYCHOLOGY)

"The reliability of passive smoking histories reported in a case-control study of lung cancer." (Lung neoplasms/ETIOLOGY)

"Prediction of the duration of hospitalization in patients with respiratory syncytial virus infection." (Respiratory tract infections/THERAPY)

It is difficult to build an exhaustive list of phrases that will reliably retrieve each subheading.

Varying contexts of words

When formulating the phrase matching rules, we had to include words that were likely to imply the subheading. For example, we devised a rule that assigned *Drug therapy* if the phrase "dose response" occurred. However, we found that the phrase did not always refer to drugs. In one abstract, it referred to exposure to cigarette smoke; in another, to exposure to radiation. Additional words can be added to the rules to restrict the contexts. It is, however, difficult to anticipate all the possible contexts.

Stemming problem

Stemming sometimes removes important semantic content or changes the meanings of words. For example, "secondary" (as in "secondary cancer sites") when stemmed becomes "second." The phrase matching rules should be extended so as to be able to specify that a particular word not be stemmed.

Lack of syntactic analysis

Syntactic analysis is sometimes required to decide whether a subheading should be assigned. For example, we have a rule that assigns the subheading *Etiology* if the word "cause" is present. Applying this rule, the indexing program assigned *Lung neoplasms/ETIOLOGY* to the following sentence:

> "Lung cancer is rapidly becoming the leading CAUSE of cancer mortality among women."

Without syntactic analysis, an indexing program cannot identify which is the subject and which the object of the word "cause."

Exceptions to phrase matching rules

The phrase matching rules need to be extended to be able to express exceptions to the rules. For example, we have a rule that assigns the subheading Complication if the word "complication" occurs. Unfortunately, it also assigns the subheading when the phrase "no complication" occurs. An extended rule to handle this problem might take this form:

> Assign subheading *Complication* if the phrase "complication" appears, unless the phrase "no complication" is also present.

Central versus incidental concepts

The phrase matching program cannot distinguish between subheading concepts that are central to the article and concepts that are incidental.

Some subheadings are implied by certain MeSH sub-trees. The following are some examples.

MeSH Sub-trees	**Subheading Implied**
D8 (Enzymes, coenzymes, enzyme inhibitors)	Enzymology
D26.394 (Misc. drugs & agents) with "therapy"	Drug therapy
B3 (Bacteria)	Microbiology
B4 (Viruses)	Microbiology
B5.354 (Fungi)	Microbiology
D12.644 (Peptides)	Metabolism

Since we only used Category C8 terms in our study, the program could not make use of terms in other MeSH categories to help in the assignment of subheadings.

Subheading trees not used

The program currently does not "tree" the subheadings. The subheading trees (MEDLARS indexing manual, part II, p.89) can be used to assign a more general subheading in place of many overlapping subheadings. For example, *Therapy* can be assigned to an abstract dealing with both Drug Therapy and Radiotherapy.

CONCLUSIONS

The automatic indexing program used in this study assigns MeSH headings by matching words in the abstracts with MeSH terms and "see" references. Our results suggest that such a program will be able to pick up most of the MEDLINE-assigned headings designated as central concepts if the program is extended in 2 ways:

a. Some kind of syntactic and/or semantic analysis is incorporated to allow the program to effectively distinguish important concepts from incidental ones.

b. The MeSH thesaurus and "see" references are supplemented with a synonyms list.

The assignment of subheadings presented many problems. The study has shown that our phrase matching rules need to be extended in a number of ways. Syntactic analysis can be incorporated into the program to reduce the number of incorrectly assigned subheadings. We also intend to carry out word frequency and word association analyses to refine the rules.

Though our indexing program can be substantially improved in the assignment of subheadings, it will be difficult to refine the program to the performance level of human indexers. However, a less-than-perfect machine indexer may be acceptable if there is some means by which the program can evaluate its own indexing. Abstracts that present indexing difficulties can be flagged for human review. This is one of the problems we shall address in the next part of the project.

ACKNOWLEDGEMENTS

We are grateful to the National University of Singapore, the Library Association of Singapore, and the Lee Foundation for their support which enabled Hayati to present this paper at the Joint Conference of the Health Libraries Sections of the Australian Library and Information Association and New Zealand Library Association, 12-16 November 1989, Auckland, New Zealand.

REFERENCES

Charen T. MEDLARS indexing manual, part II. Bethesda, MD : National Library of Medicine, 1983. (NTIS PB84-104280)

Funk ME, Reid CA, McGoogan LS. "Indexing consistency in MEDLINE." Bulletin of the Medical Library Association Apr 1983; 71(2) : 176-83.

Hooper RS. Indexer consistency tests: origin, measurement, results, and utilization. Bethesda, MD : IBM Corporation, 1965. (TR95-56)

Lancaster FW. Evaluation of the MEDLARS demand search service. Bethesda, MD : National Library of Medicine, 1968.

Leonard LE. Inter-indexer consistency and retrieval effectiveness: measurement of relationships. Ph.D. thesis, University of Illinois at Urbana-Champaign, 1975.

Marcetich J, Schuyler P. "The use of AID to promote indexing consistency at the National Library of Medicine." Paper presented at the 81st Annual Meeting of the Medical Library Association, Montreal, Canada, June 1981.