COMPUTERS AND BIOMEDICAL RESEARCH 25, 336–350 (1992)

Evaluation of a Meta-1-Based Automatic Indexing Method for Medical Documents

MICHAEL M. WAGNER AND GREGORY F. COOPER

Section of Medical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15261

Received April 2, 1992

This paper describes MetaIndex, an automatic indexing program that creates symbolic representations of documents for the purpose of document retrieval. MetaIndex uses a simple transition network parser to recognize a language that is derived from the set of main concepts in the Unified Medical Language System Metathesaurus (Meta-1). MetaIndex uses a hierarchy of medical concepts, also derived from Meta-1, to represent the content of documents. The goal of this approach is to improve document retrieval performance by better representation of documents. An evaluation method is described, and the performance of MetaIndex on the task of indexing the Slice of Life medical image collection is reported.

1. INTRODUCTION

In this paper, we describe a method for automatically creating representations of documents for use in document retrieval systems. The process of constructing such representations is called *automatic indexing*. The earliest description of an automatic indexing method is attributed to Luhn (*1*). Luhn's basic idea, which is now called the *vector-space* model, was that simple text features (e.g., words or phrases) could be identified automatically in documents and used to represent the documents for indexing and retrieval.

The properties of the vector-space model have been extensively investigated (*2–5*). One of the principal results of this body of research is that there is an upper bound on performance of document retrieval systems using this method.

Partly because of this performance limitation, there has been recent interest in the use of *symbolic representations* of documents in document retrieval systems (*6*). We use the term *symbolic representation* to mean a representation of the objects in some domain (and their interrelationships) that has the property of cognitive correspondence, i.e., the symbols correspond to some natural human conceptualization of the domain. One example of a symbolic representation for indexing is the Medical Subject Heading thesaurus (MESH). MESH subject headings correspond to medical concepts, and the interrelationships between subject headings in the MESH trees correspond to medical taxonomic

0010-4809/92 $5.00

1.	Esophagitis, Candida, Mult Diffuse Plaq., 6 mo painful dysphagia, 40 lb wgt loss, no predisposition
2.	Menstrual Cycle days 6-10 endometrial cells; cervical smears: gynecologic cytology
3.	Mitral Insufficiency, Floppy Mitral valve with Ruptured chordae

FIG. 1. Three image descriptions from the SOL database (the title and comment fields are merged).

relations. Symbolic representations may lead to improved document retrieval performance because they can potentially represent the contents of documents better than the vector-space model. However, a key drawback of symbolic indexing is that it is difficult to automate the process, in the general case (*6*). The obstacles to automatic generation of symbolic representations are significant, being similar to the problems of automatic natural language understanding. These obstacles can broadly be described as problems with ambiguities (e.g., word sense, pronoun reference) that may require vast amounts of general and specific knowledge to resolve. Although currently there is no general solution to the problem of natural language understanding, for selected classes of medical documents it has been possible to create reasonably accurate representations of the content automatically from text (*7–9*). This paper describes a Meta-1 based method that can do this for a medical image collection.

2. THE SLICE OF LIFE MEDICAL IMAGE COLLECTION

The document collection used in this work is the Slice of Life (SOL) medical image collection, a collection of more than 31,000 medical images (*10, 11*) on a videodisc. Each image in SOL is represented by a record in a flat-file database that contains both coded and free text fields. The coded fields indicate the type of image, the stain (if applicable), and the magnification (if applicable). The text fields contain the title of the slide and a brief description (Fig. 1).

The free text descriptions in the SOL database are linguistically simple. The descriptions consist of brief phrases that usually describe one central object (e.g., a heart valve) and some property of the object to which attention is being drawn (e.g., endocarditis). There is very little word sense ambiguity (e.g., foot does not mean 12 in. or the base of a cliff) and expressions involving negation, quantification, or uncertainty are rare. The language does have some computationally problematic features including elliptical usages (missing words, e.g., "pituitary" instead of "pituitary gland"), misspellings, and nonstandard abbreviations.

3. METAINDEX

MetaIndex is a program that processes the free text SOL descriptions to generate symbolic representations of SOL images. Since MetaIndex makes

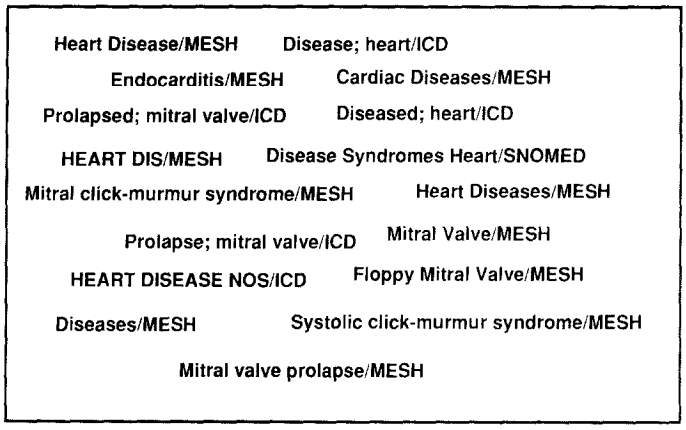

FIG. 2. Seventeen of ~97,700 strings in Meta-1p with source nomenclature. The source of a string follows the string, separated by a slash. Abbreviations: MESH, Medical Subject Headings; ICD, International Classification of Diseases; SNOMED, Systematic Nomenclature of Medicine.

extensive use of the information in the Unified Medical Language System Metathesaurus (Meta-1), we give an overview of Meta-1 before describing the MetaIndex algorithm. More complete descriptions of Meta-1 can be found in its documentation (*12*), and in a series of papers by Sherertz, Tuttle, and coworkers (*13–18*).

3.1. The Unified Medical Language System Metathesaurus (Meta-1)

In this project, we used a prerelease version of Meta-1, which we denote as *Meta-1p*, because only the prerelease version was available when we began our investigation. To avoid confusion, we will use the term *Meta-1* in this paper when comments apply to the general Metathesaurus schema, and reserve the term *Meta-1p* for comments about Meta-1p.

For clarity, we use three figures to describe the organization of information in Meta-1. As a starting point, we view Meta-1 as a very large set of unrelated strings of medical language from different source nomenclatures (Fig. 2). These strings are predominantly simple noun phrases, e.g., *mitral valve prolapse*.

This set of strings is clustered into conceptual equivalence classes defined by two types of links (binary relationships) between strings in Meta-1: *lexical variant* and *synonymy* links (Fig. 3). Two strings are *lexical variants* if one is a reordering of the words of the other, e.g., "Huntington's disease" and "disease; Huntington's" or if they differ only in capitalization or other minor typographical ways, e.g., "Huntingtons Disease". Two strings are synonyms if they are explicitly identified as synonyms in one of the Meta-1 source nomenclatures. Meta-1p contains 19,748 lexical variant relations and 11,700 synonymy relations. For convenient reference, each concept is given the name of one of its constituent strings.

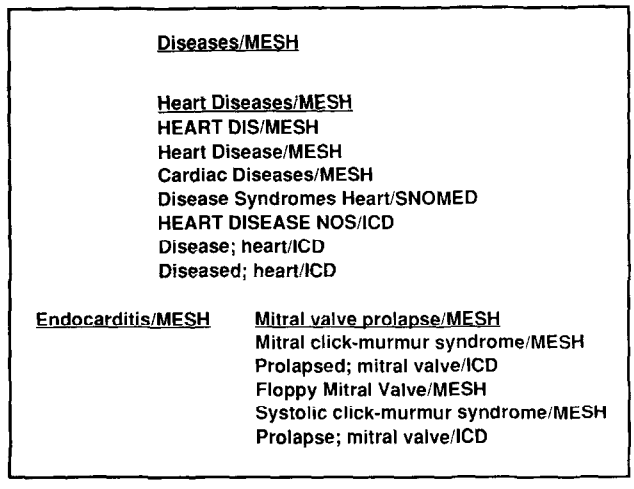

FIG. 3. Four of the ~30,000 concepts in Meta-1p. The string chosen to be the name of the concept is underlined.

The highest epistemological level in Meta-1 is defined by a set of *interconcept* link types (Fig. 4). The most frequently used interconcept link type is the narrower-than type (NT). The NT link type is a standard link type used in the field of information retrieval. It is roughly equivalent to the combination of two standard semantic network relation types, *part-of* and *is-a*. A NT link between two concepts means that all instances of the narrower concept are also instances of the broader concept. For example, given the link *heart diseases NT diseases* and definitions of the concepts as classes of diseases, then if x is a heart disease then x is a disease. In Meta-1, the set of NT links between concepts was created by a team of Meta-1 editors using the set of intersubject heading links in the

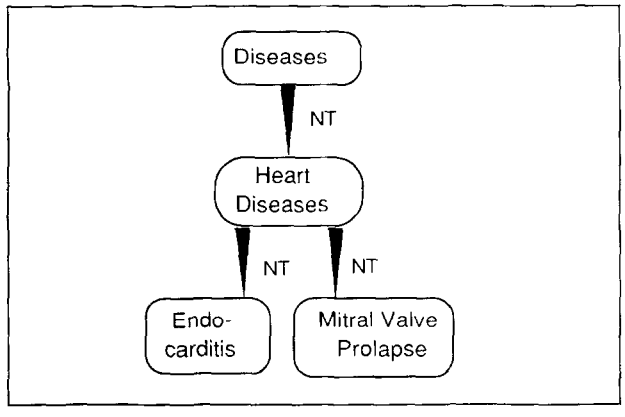

FIG. 4. Interconcept relationships, the highest level of organization of information in Meta-1p.

MESH hierarchy to identify candidate relationships. Other link types (e.g., *is-a* and *manifestation-of*) are included on an experimental basis in Meta-1.[1] We can interpret this level of organization as a semantic network which we can use to make inferences. We refer to this as a *domain model.*

3.2. MetaIndex Components

MetaIndex consists of three logical components, a preprocessor, a parser, and a domain model. Because the parser and domain model use information contained in Meta-1, they will be described first.

Parser. The parser recognizes a language similar to the language defined by the set of strings in Meta-1p. These strings are found in the main concept and lexical variant files (i.e., the MRMC and MRLV files) in Meta-1p. In MetaIndex, we consider each of these strings to be a legal expression in a medical language, and that the set of these strings taken together completely defines the language.[2] We justify the adequacy of this language of noun phrases for the natural language processing of SOL descriptions by pointing out that the SOL language is also basically a simple noun phrase language.

Because this language is finite (i.e., every legal expression can be listed), a simple parser for it could be based on a dictionary data structure with a "member" operation. However, we use a slightly more complex parser, for reasons that we will discuss. We organize the set of Meta-1 strings into a graph in which each node represents a string from the Meta-1 language. To this structure we add another set of nodes that represent strings that are not in the language, but are necessary to convert the structure into a simple transition network (Fig. 5A).

Specifically, the transition network is constructed as follows: First, we downcase and remove most punctuation marks from the set of strings in the main concept file. Second, we group the terms from the main concept file into size classes by the number of words in the term. Let n be the size of the largest class. Finally, the network is constructed by adding a pair of parent nodes for each node in size class n. The first parent consists of the first $n - 1$ words and the second parent consists of words 2 through n. For example, the concept *coronary artery disease* generates the parents *coronary artery* and *artery disease*. Links from parents to children are created, and duplicate nodes in size class $n - 1$ are merged. This process is repeated for each smaller size class until $n = 1$. Note that we do not generate all $n - 1$ word parents of a node; the two parents described are sufficient for the word-order independent phrase recognition that we use.

The parser uses a queue-based spreading-activation search algorithm to traverse the transition network structure in a work-order-independent manner. This can best be understood by way of an example. Consider how the SOL

[1] The links are in file MRCTX; 1 in the September 1990 release of Meta-1.

[2] Because of our handling of lexical variants, the actual set is somewhat different. This will be explained.

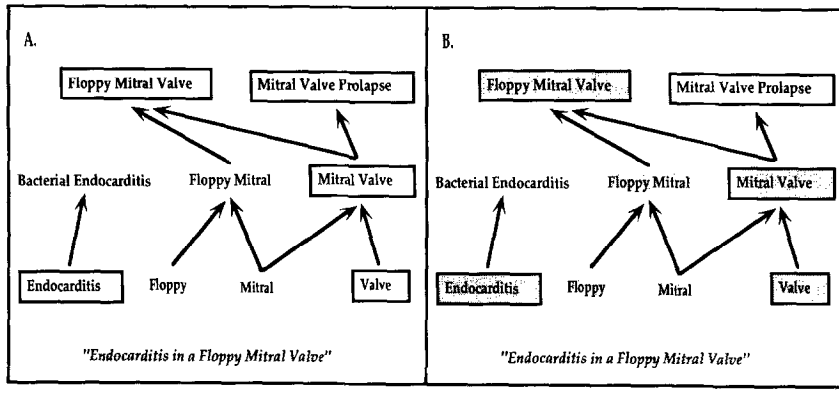

FIG. 5. Transition network before (A) and after (B) parsing the phrase "endocarditis in a floppy mitral valve." Strings in the "Meta-1p language" are in boxes; unboxed strings are not in the language. Shading indicates recognition of a string. Transition conditions are not labeled but can be inferred in a straightforward manner (e.g., transition from "valve" to "mitral valve" requires the word "mitral").

description *endocarditis in a floppy mitral valve* is processed (Fig. 5B). First, each of the individual words in the phrase is presented to the network. Words that do not match nodes in the entry level of the network (the set of single word roots of the transition network) are not recognized and, in effect, eliminated. Those words that are recognized (*endocarditis, floppy, mitral, valve*) cause the corresponding node to be "activated" and placed on a search queue. Nodes are then taken off the queue in turn and their children are visited. Whenever a child is visited, its list of words is marked based on the words in the parent node. For example, when the node *floppy mitral valve* is visited from the node *floppy mitral,* the words *floppy* and *mitral* are marked. When all of a child's words are marked, it is itself "activated" and put on the search queue. The goal of this algorithm—recognition of legal expressions in the language—is achieved whenever a node representing a Meta-1p expression is "activated." This approach is similar to that described by Shoval (*19*).

Note that this parser does not use the information about lexical variants contained in the Meta-1p MRLV files. Instead, word-order variants are recognized because the algorithm that traverses the transition network is word-order insensitive. Thus the input *A C B* will ultimately activate the *ABC* node. We chose this design because unusual word orderings, perhaps not in the MRLV file, are common in SOL. Other lexical variants are recognized because we remove punctuation marks and downcase capital letters before building the transition network and before indexing documents.

This parser implementation has several advantages over a simple dictionary look-up. First, it ignores word order, thus enlarging the set of word-order lexical variants recognized. Second, it can potentially be used to resolve elliptical usages in the SOL database. For example, the word *Hashimoto's* only occurs

in one sense in Meta-1—meaning Hashimoto's thyroiditis. If the word *Hashimoto's* alone was found in the text of a SOL database record, it would be reasonable to infer that Hashimoto's thyroiditis was the intended meaning. Structurally, this corresponds to a network in which the node representing "Hashimoto's" has only one child, and that child is "Hashimoto's thyroiditis." Although this heuristic method seems promising, to date it has not been implemented.

Domain model. The second MetaIndex component, the domain model, is taken directly from Meta-1 as the set of all Meta-1 concepts with their *NT, is-a,* or *part-of* interconcept relationships (Fig. 4). This domain model is a very simple model of medicine, representing only a fraction of the relationships that exist between medical concepts. For example, *endocarditis* and *mitral valve* are related by the fact that endocarditis is a pathologic process that can involve the mitral valves, but this relationship is not encoded in the domain model.

We choose to represent images with concepts drawn from this domain model because the medical knowledge encoded in the domain model is potentially useful for retrieval. For example, consider the Meta-1 concepts *ruptured ectopic pregnancy* and *abruptio placenta,* which are narrower than the Meta-1 concept *pregnancy complications.* This domain model could be used to support the inference that images indexed with the more specific terms, are also of interest to a searcher that used the more general term *pregnancy complications.*

The mapping from parser to domain model. We have described the parser and the domain model. Next we show the connection between the two. Automatically deriving a representation in a domain model from the output of a parser can be a difficult problem. Consider the SOL description *endocarditis in a floppy mitral valve.* One cannot conclude that the image should be represented by the terms *endocarditis* and *mitral valve prolapse* without assuming that the text descriptions of SOL images are declarations about the subject of the image. In indexing SOL, we categorically make this assumption. We assume that every expression recognized by the parser is a description of a medical concept that is relevant to the image. In MetaIndex, this assumption corresponds to a node in the parser being linked to a node in the domain model if and only if the text string represented by the node is a member of the conceptual class that is named by the concept in the domain model. Figure 6 shows this parser to domain model mapping. The parser node representing the text string *floppy mitral valve* is linked to the domain model concept *mitral valve prolapse* because *floppy mitral valve* is a member of the *mitral valve prolapse* equivalence class of strings.

Preprocessor. MetaIndex handles plurals, inflections, abbreviations, and some elliptical usages by preprocessing SOL descriptions before they are presented to the parser. The preprocessor uses substitution tables to expand abbreviations and complete ellipses. We developed substitution tables for ellipses and abbreviation by inspecting the input and output of MetaIndex on a random 10% sample of SOL descriptions. Candidate substitutions for inclusion in substitution tables (e.g., *thyroid disease* for *thyroid*) were identified. Keyword-in-context listings of every occurrence of a candidate substitution were reviewed

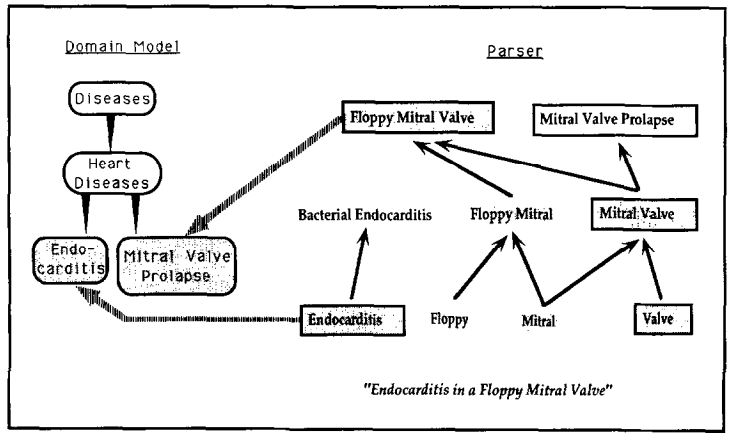

FIG. 6. Connection between parser and domain model. The arrows from the language model to the domain model indicate that the presence of the string in the SOL free text is sufficient evidence to conclude that the image is about that concept. Shading indicates that a node has been "activated." The output of the parser is the set of shaded, boxed nodes, e.g., floppy mitral valve, mitral valve, endocarditis, valve. The activated concepts in the domain model are endocarditis and mitral valve prolapse. Not shown are the two additional activated concepts, valve and mitral valve.

to ensure that the term was only used in one sense in SOL. If a candidate term was only used in one sense in SOL, it was added to a substitution table. Thirty-four ellipses and 30 abbreviations were identified and handled in this manner. Note that these elliptical usages differed from those previously discussed because they only became unambiguous in the context of the SOL collection. For example, *thyroid* is a highly ambiguous term in Meta-1p (it has 15 children in the network, e.g., *thyroid storm* and *thyroid effects*). However, in SOL *thyroid* only denotes the thyroid gland. Therefore, this elliptical completion was permitted.

4. EVALUATION OF METAINDEX

The ideal evaluation for an indexing method would be a field test in which the performance of a retrieval system using the index is compared against the current best retrieval system. However, there is also a role for less expensive evaluations in the course of the development of indexing methods. In this study, we evaluated the MetaIndex method relative to manual indexing. Such a method cannot predict or estimate absolute field performance of a retrieval system incorporating the indexing method, but can identify problems with the indexing method that might affect field performance.

4.1. Methods

A study set of 126 SOL images was selected at random from the collection of 31,500 images. For each image, the free text fields from the image's record in the SOL database were concatenated with the contents of the "organ" field

TABLE 1

CAUSES OF INACCURATE INDEXES (N = 35)

Cause	N	%
Wrong term sense	20	57
Term part of longer index term	10	29
Word-order effect	2	6
Missed negation	1	3
Coder error	2	6

and the resulting string was used as input to MetaIndex, which generated a set of indexes for the record. The "organ" field was included because it contained text that also described the subject of the image.

Five physicians with specialty training in Internal Medicine (2), Pediatrics (1), Neurosurgery (1), or Pathology (1) were recruited to serve as judges. Each judge was given detailed instructions and asked to judge the MetaIndex indexes of 25–35 image records for accuracy and completeness. For each image, the judge was given the SOL database description and the MetaIndex-generated terms. Note that the judges were provided with the SOL database description of the image, not the image itself. The judges were asked to perform two tasks. In the first task, s/he decided whether each index was acceptable relative to the image description. In the second task, the judge was asked to write in any additional indices that s/he felt were missed by MetaIndex. No time limit was placed on the judging.

To assess the reliability of the method, we measured interjudge variability on a set of 10 description/MetaIndex-index pairs that were judged by all five judges.

4.2. Results

On average, a SOL image was described by eight words of text. MetaIndex generated 407 indexing terms for the 126 image sample (3.3 terms/image).

Accuracy of indexing. Of a total of 407 MetaIndex generated terms, 372 (91%[3]) were judged acceptable. For this task, we define interjudge agreement as the proportion of terms generated by MetaIndex for which at least four of the five judges agreed on the acceptability of the term. The interjudge agreement was 94% (34/38).

We analyzed the 35 unacceptable indexes to assign causes to the errors (Table 1). Most (57%) errors resulted from our assumption that the presence of a text string was sufficient evidence for a concept ("wrong term sense"). For example, the index *patients* was generated from the description *Huntington's disease patient with old cva.* A smaller number of errors resulted from properties of the algorithm that could be changed easily. For example, for the description *multiple*

[3] The 95% confidence interval is 85–94%.

TABLE 2

CAUSES OF MISSED INDEXES (N = 207)

Refractory	N	%	Tractable	N	%
Sophisticated inference	52	25	Missing concept	66	32
Abbreviations	7	3	Missing is-a link	46	22
Ellipses	7	3	Missing part-of link	16	8
Word variants	2	1	Coder error	4	2
Misspellings	3	1	Program error	4	2
Totals	71	34		136	66

sclerosis, the indexes *sclerosis* and *multiple sclerosis* were generated. This type of error is counted in the category "term part of longer index term." Two errors resulted from ignoring word order. For example, the index *cell cycle* was generated for the description *menstrual* cycle *days 6–10 endometrial* cells; *cervical smears; gynecologic cytology*. Ignoring negation in the SOL descriptions caused only one error.

Completeness of indexing. To the set of 372 MetaIndex terms that they judged acceptable, the judges added an additional 207 index terms for a total of 579 terms. Thus, the algorithm identified 372/579 (64%[4]) of the total set of indexes that were acceptable to, or volunteered by, the judges.

Interjudge agreement was poor for the task of listing missing indexes. For the 10 image descriptions indexed by all five judges, only 6% (3/53) of the additional indexes were suggested by at least four of five judges. Note that most of these judge-generated indexes seemed to be good indexes. Therefore, 64% should be interpreted as an upper bound on the completeness of the indexing of the algorithm. This methodological problem is discussed in Section 6.

We also analyzed these 207 indexes to assign a cause for the errors (Table 2). We divided the missed terms into two categories, *refractory* and *tractable*, based on whether MetaIndex, with straightforward modifications, could be expected to generate the missed term. Within the two major categories, we created subcategories that reflected the type of knowledge we believe physicians used to generate the index from the SOL description. In 34% (71/207) of the indexes that were not generated by MetaIndex, the missed index was considered refractory to algorithmic identification. Of these 71 *refractory* terms, 3 were missed due to misspellings in SOL (e.g., the word *opstructive* in the description *obstructive lung disease*). Seven were unrecognized abbreviations (e.g., *nplsm* for *neoplasm*). Seven were elliptical usages that were either missed by the sampling method used to build preprocessor substitution tables, or were ambiguous in the SOL database (e.g., *cortex* meaning either adrenal, auditory, or cerebral cortex). Two missed terms were word variants (e.g., *arterial* instead of *arteries*). The remaining 52 *refractory* terms were missed because they could

[4] The 95% confidence interval is 56–72%.

not be readily identified algorithmically. These terms were the product of sophisticated inference on the part of the physician. For example, from the text "endometrium, 7–10 days, normal" a physician produced the indexes *proliferative endometrium* and *endometrial dating technique*.

Sixty-six percent (136/207) of the indexes that were not generated by MetaIndex were considered amenable to algorithmic generation. Sixty-six indexes were literally present in the SOL text and could have been identified if the string had been part of the Meta-1 language. An additional 62 terms would have been identified by simple one-step or chained deductions from the text through *is-a* (46), or *part-of* (16) inferences, had the necessary interconcept links been present in Meta-1p.[5] The method used to make these determinations involved writing down the chain of deductions starting with the phrase in the SOL description and checking that each link in the chain was a simple *is-a* or *part-of* inference or a synonomy transformation. Note that some of these 62 terms would also have required the presence of additional concepts in Meta-1p, as well as the presence of the aforementioned links. We chose to classify such terms into the appropriate "missing-link" category, not the missing concept category, because we believe that adding interconcept links to Meta-1p will be more difficult than adding concepts. The remaining 4% of the errors were considered tractable because they were either caused by a single correctable error in the transition network, or were indexing terms suggested by the judges that did not seem to be appropriate to us.

5. Related Work

Use of MESH and Meta-1 for automatic indexing. In 1968, the MESH thesaurus (the principal Meta-1 source) was used experimentally in the SMART system to index MEDLARS documents automatically (*4*). More recently, Meta-1 has been used in this manner in the SAPHIRE and CLARIT systems (*20, 21*). Vries (*22*) uses thesaurus *narrower-than* links to identify the most specific indexes for documents (the standard approach is to use thesaurus transformations to find *broader-than* indexes).

Natural language understanding (NLU) and automatic indexing. NLU requires many types of knowledge about language and the domain of discourse to constrain the interpretation of natural language statements. A review of NLU work related to automatic indexing and NLU can be organized by the types of knowledge used in the systems.

Morphologic knowledge is knowledge about the substructure of words (e.g., roots and affixes). Morphologic knowledge is sometimes useful in NLU because the meaning of a medical word, such as *appendicitis*, can be determined from the meanings of its root and suffix. Approaches using morphology that are relevant to the indexing of medical documents are described in (*23–25*). Mor-

[5] A subsequent analysis using the interconcept relations in Meta-1 showed that 17/62 would be identified.

phology is also the basis of stemming, a common procedure in automatic indexing.

Syntactic knowledge is usually encoded as a set of grammatical types (e.g., noun), a listing of the types of individual words, and a grammar which defines the allowable relationships between grammatical types. Syntax can constrain the set of possible meanings for words with multiple senses. Syntax can also be used to identify potentially useful clusterings of words, e.g., noun phrases, for indexing. Syntactic parsing has been used to recognize noun phrase features in the vector-space systems SMART and CLARIT.

Semantic knowledge, in some NLU systems, may be encoded as a set of semantic types, a listing of the types of individual words, and a grammar which defines the allowed combinations of semantic types. Semantic knowledge alone is often effective for NLU in technical domains because syntactic knowledge is not needed to constrain the meanings of words. For example, the word *foot* has a single meaning in most medical settings. Canfield (*7, 26*) used a semantic approach to the problem of understanding echocardiogram reports. Semantic approaches to NLU problems in technical domains have become sufficiently common that the term *sublanguage analysis* has been coined to refer to the process (*27*).

Systems often use more than one of the above types of knowledge. For example, Sager's Medical Narrative Processor uses *semantic types* as *selectional restrictions* in a basically syntactic parser (*9*).

Our approach can be viewed as a semantic approach in which each Meta-1 equivalence class (concept) is a semantic type. Each Meta-1 string is assigned the semantic type of the equivalence class of which it is a member. In our application, there is a single grammatical rule that defines a legal SOL description as any combination of semantic types. Within this framework, it is easy to see that this approach might be extended by adding semantic combination rules to limit how types can combine in SOL descriptions.

6. Discussion

Performance of MetaIndex. The evaluation of MetaIndex used the set of indexes either generated by or accepted by a single medically sophisticated person as a "gold standard." We evaluated the indexes generated by MetaIndex relative to this standard. We believe that the use of physician-generated indexes is a better "gold standard" than, for example, indexes produced by professional indexers, because physicians may tend to use the same terms for indexing as other physicians would later use for searching. This would correlate better with retrieval system performance. As was stated earlier, this type of evaluation cannot predict absolute field performance of a retrieval system incorporating the generated index, but can identify ways to improve performance of the indexing system.

Using the above criteria, we found that most (91%) of the MetaIndex-generated indexes were good indexes. Most of the indexes that were considered

errors were caused by our assumption that the presence of a text string was sufficient evidence to conclude that a concept was present. We believe that this rate of false indexes is acceptable in many automatic indexing applications.

Because of methodological problems identified by the high interjudge variability in adding index terms, we are unable to state the completeness of indexing relative to a thorough manual indexing. The high interjudge variability suggests that each image could be indexed by many more terms than a single physician would produce. Therefore, we believe that a thorough manual indexing procedure would assign more indexes to the images than the total we observed. A different problem stems from the fact that we provided the judges with the MetaIndex-generated indices and let them mark them as appropriate or not. We do not know whether physicians would have generated all of these terms spontaneously.

Nevertheless, a comparison of the MetaIndex indexing with the incomplete human indexing still showed that the algorithm did not generate many desirable indexes. An important result of our study is the identification of the principal determinants of the completeness of indexing of the MetaIndex algorithm. Completeness of indexing depended heavily on the information in Meta-1p. In particular, we showed that approximately 60% of the missed indexes could have been generated using the MetaIndex approach, if the existing Metathesaurus schema had been fully populated with the necessary concepts and interconcept relationships. Since many other information retrieval models use the knowledge encoded in concepts and interconcept links to improve system performance, correction of the deficiencies that we identified in Meta-1p may improve the performance of other information retrieval systems that use the information in Meta-1p.

Generalization of the method. The performance of NLU methods depend on the domain of discourse and the linguistic characteristics of the natural language. The accuracy that we observed for MetaIndex probably depends on the simple semantics of SOL descriptions. The semantics make it possible for a simple phrase recognition procedure and a mapping from text to concepts (defined by the synonomy relations in Meta-1) to generate accurate indexes.

NLU performance also depends on the depth of understanding required. This can also be thought of as the expressiveness of the representation language that is being used to represent the documents. The performance of MetaIndex would have been worse if the judges had suggested coordinate indexing terms, e.g., *endocarditis, treatment of.*

Knowledge-based approaches tend to be brittle, meaning that it is difficult to apply them to new domains or languages. The current version of MetaIndex is no exception. MetaIndex is applicable to other document collections provided they have the following characteristics: (1) the domain is biomedicine, (2) the semantics are simple (i.e., the presence of a phrase signifies that the document is about the phrase), and (3) the representation requirements are coarse, i.e., at the level of MESH subject headings. To apply this approach to a nonmedical domain one would need a different thesaurus. To extend this method to docu-

ments with greater linguistic complexity would require a different parser and a different method for mapping from parser output to the symbolic representation. The kinds of syntactic knowledge required to parse a more complex language are not in Meta-1 at the current time. To map from parser output to a semantic representation is usually a much harder problem than it is for the SOL document collection. A promising approach to this problem is learning the mapping between text features and conceptual representation from examples (*28*).

7. Summary and Conclusions

The MetaIndex program demonstrates a method for using the information in the UMLS Metathesaurus for the automatic generation of symbolic representations of documents. Relative to manual indexing, MetaIndex generated a relatively accurate, but incomplete indexing for the Slice Of Life collection of medical images. The majority of the missed indexes could have been generated by the algorithm if the Metathesaurus schema, consisting of medical terms, concepts, and the interrelationships between concepts, had been completely populated with the needed concepts and interconcept relationships. Other document retrieval systems make use of term, concept, and interconcept relations for indexing. Thus, efforts to extend the coverage of the current Metathesaurus are likely to improve significantly those systems that use the Metathesaurus to index medical documents.

Acknowledgments

We thank Randolph A. Miller, Johanna Moore, Steven Small, and Dario Giuse for helpful comments on earlier drafts of this paper. Suzanne Stensaas supplied the Slice of Life Database. We also thank Stuart Weinberg and the other test subjects. This work was supported by the National Library of Medicine Training Grant T15 LM 07059.

References

1. Luhn, H. A new method of recording and searching information. *Am. Doc.* **4,** 14 (1953).
2. Cleverdon, C., Mills, J., and Keen, M. "Factors Determining the Performance of Indexing Systems, Aslib–Cranfield Research Project." Cranfield College of Aeronautics, England, 1966.
3. Salton, G. "Automatic Information Organization and Retrieval." McGraw–Hill, New York, 1968.
4. Salton, G. "The SMART Retrieval System: Experiments in Automatic Document Processing." Prentice–Hall, Englewood Cliffs, NJ, 1971.
5. Salton, G., and McGill, M. "Introduction to Modern Information Retrieval." McGraw–Hill, New York, 1983.
6. Croft, W. Automatic indexing. *In* "Proceedings, American Society of Indexers, New York, 1988," pp. 85–100.
7. Canfield, K., Bray, B., Huff, S., and Warner, H. Database capture of natural language echocardiographic reports: A unified medical language system approach. *In* "Proceedings, 13th Annual Symposium for Computer Applications in Medical Care," pp. 559–563. IEEE Comp. Soc. Press, Washington, DC, 1989.
8. Haug, P., Ranum, D., and Frederick, P. Computerized extraction of coded findings from free-text radiologic reports. *Radiology* **174,** 543 (1990).

9. SAGER, N., FRIEDMAN, C., AND LYMAN, M. "Medical Language Processing: Computer Management of Narrative Data." Addison–Wesley, Reading, MA, 1987.
10. STENSAAS, S. Slice of Life IV Database. (Available from Spencer S. Eccles Health Sciences Library, University of Utah, Salt Lake City, UT.)
11. STENSAAS, S. Slice of Life IV Videodisk. (Available from Spencer S. Eccles Health Sciences Library, University of Utah, Salt Lake City, UT.)
12. "UMLS Knowledge Sources Experimental Edition." National Library of Medicine, 1990.
13. SHERERTZ, D., TUTTLE, M., BLOIS, M., AND ERLBAUM, M. Intervocabulary mapping within the UMLS: The role of lexical matching. *In* "Proceedings, 12th Annual Symposium for Computer Applications in Medical Care," pp. 201–206. IEEE Comp. Soc. Press, Washington, DC, 1988.
14. SHERERTZ, D., AND TUTTLE, M. Lexical mapping in the UMLS metathesaurus. *In* "Proceedings, 13th Annual Symposium for Computer Applications in Medical Care," pp. 494–499. IEEE Comp. Soc. Press, Washington, DC, 1989.
15. SHERETZ, D., OLSON, N., TUTTLE, M., AND ERLBAUM, M. Source inversion and matching in the UMLS metathesaurus. *In* "Proceedings, 14th Annual Symposium for Computer Applications in Medical Care," pp. 141–145. IEEE Comp. Soc. Press, Washington, DC, 1990.
16. TUTTLE, M., *et al.* Toward a biomedical thesaurus: Building the foundation of the UMLS. *In* "Proceedings, 12th Annual Symposium for Computer Applications in Medical Care," pp. 191–195. IEEE Comp. Soc. Press, Washington, DC, 1988.
17. TUTTLE, M., *et al.* Implementing Meta-1: The first version of the UMLS metathesaurus. *In* "Proceedings, 13th Annual Symposium for Computer Applications in Medical Care," pp. 483–487. IEEE Comp Soc Press, Washington, DC, 1989.
18. TUTTLE, M., *et al.* Using Meta-1: The first version of the UMLS metathesaurus. *In* "Proceedings, 14th Annual Symposium for Computer Applications in Medical Care," pp. 131–135. IEEE Comp. Soc. Press, Washington, DC, 1990.
19. SHOVAL, P. An expert consultation system for a retrieval data-base with a semantic-network of concepts. Ph.D. dissertation, University of Pittsburgh, 1981.
20. HERSH, W., PATTISON-GORDON, E., GREENES, R., AND EVANS, D. Adaptation of Meta-1 for SAPHIRE, A general purpose information retrieval system. *In* "Proceedings, 14th Annual Symposium for Computer Applications in Medical Care," pp. 156–160. IEEE Comp. Soc. Press, Washington, DC, 1990.
21. EVANS, D., *et al.* Automatic indexing using selective NLP and first-order thesauri. *In* "Proceedings, RIAO 91, Barcelona, 1991."
22. VRIES, J., *et al.* An automated indexing system utilizing semantic net expansion. *Comput. Biomed. Res.* **25,** 153–167.
23. MOORE, G., *et al.* Word root translation of 45,564 autopsy reports into MeSH Titles. *In* "Proceedings, 11th Annual Symposium for Computer Applications in Medical Care," pp. 128–132. IEEE Comp. Soc. Press, Washington, DC, 1987.
24. NORTON, L., AND PACAK, M. Morphosemantic analysis of compound word forms in medical language. *Methods Inf. Med.* **22,** 29 (1983).
25. PACAK, M., NORTON, L., AND DUNHAM, G. Morphosemantic analysis of -ITIS forms in medical language. *Methods Inf. Med.* **19,** 99 (1980).
26. CANFIELD, K., BRAY, B., AND HUFF, S. Representation and database design for clinical information. *In* "Proceedings, 14th Annual Symposium for Computer Applications in Medical Care," pp. 350–353. IEEE Comp. Soc. Press, Washington, DC, 1990.
27. GRISHMAN, R., AND KITTREDGE, R. "Analyzing Language in Restricted Domains: Sublanguage Description and Processing." Erlbaum, Hillsdale, NJ, 1986.
28. FUNG, R., CRAWFORD, S., APPELBAUM, L., AND TONG, R. An architecture for probabilistic concept-based information retrieval. *In* "Proceedings, 13th International Conference on Research and Development in Information Retrieval, Brussels, 1990," pp. 392–404.