Journal of Biomedical Informatics **34,** 262–273 (2001)
doi:10.1006/jbin.2001.1025, available online at http://www.idealibrary.com on IDEAL®

Selective Automated Indexing of Findings and Diagnoses in Radiology Reports

William Hersh,*[,1] Mark Mailhot,* Catherine Arnott-Smith,† and Henry Lowe†

**Division of Medical Informatics and Outcomes Research, Oregon Health and Science University, Portland, Oregon 97201; and †Center for Biomedical Informatics, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15213*

Received November 18, 2001; published online January 22, 2002

The recent improvements in capabilities of desktop computers and communications networks give impetus for the development of clinical image repositories that can be used for patient care and medical education. A challenge in the use of these systems is the accurate indexing of images for retrieval performance acceptable to users. This paper describes a series of experiments aiming to adapt the SAPHIRE system, which matches text to concepts in the UMLS Metathesaurus, for the automated indexing of image reports. A series of enhancements to the baseline system resulted in a recall of 63% but a precision of only 30% in detecting concepts. At this level of performance, such a system might be problematic for users in a purely automated indexing environment. However, if the ability to retrieve images in repositories based on content in their reports is desired by clinical users, and no other current systems offer this functionality, then follow-up research questions include whether these imperfect results would be useful in a completely or partially automated indexing environment and/or whether other approaches can improve upon them.

Key Words: natural language processing; automated indexing; SAPHIRE; unified medical language system (UMLS); Metathesaurus; evaluation.

[1]To whom correspondence and reprint requests should be addressed at Division of Medical Informatics and Outcomes Research, Oregon Health and Science University, BICC, 3181 SW Sam Jackson Park Road, Portland, OR 97201. Fax: 503-494-4551. E-mail: hersh@ohsu.edu.

1. INTRODUCTION

Until recently, the availability of online repositories of clinical images was extremely limited. But with the ability of modern desktop computers to display high-quality images, along with large disks to store them on servers and fast networks to allow their widespread viewing, the door has been opened to new applications of clinical image repositories for clinical care and education. For clinical care, these collections allow clinicians to compare a current case against previous instances of the suspected finding or diagnosis, while for education they make the use of images for teaching purposes much easier.

There is, however, a major challenge to the effective use of clinical image repositories, which is that images must be properly indexed for accurate retrieval. Automated indexing of radiology images involves processing either the image itself or some surrogate, such as the clinical report, to identify the content. Analyzing the image itself for indexable features is a process which has been investigated for many years without generalizable success [1–3]. Another approach is to index text associated with the image. This has proved more tenable, as exemplified by the searchable index accompanying the well-known Slice of Life videodisk application [4].

1532-0464/01 $35.00

A great deal of work has focused on the processing of the reports of clinical images in recent years. Two applications, MedLEE [5] and Symtext [6], have advanced to the point of their output being included in operational clinical information systems. These systems are used, however, for a task different from general indexing of the content of images. They are instead focused on the task of recognizing specific clinical conditions or events in the context of the image, such as the presence of a lesion or a diagnosis on a chest radiograph [5, 7–11] or on a neuroradiology image [12], or the rendering of a specific diagnosis [13].

Applications like MedLEE and Symtext are highly successful at finding a very constrained number of findings or diagnoses. In none of the studies cited in the previous paragraph were more than 47 findings identified. These systems have not been developed to code for the thousands of terms in the Unified Medical Language System (UMLS) Metathesaurus. Furthermore, they do not carry out *selective* indexing of the type seen in document indexing (or information retrieval) systems, where only important terms are indexed as opposed to any terms that occur. MedLEE and Symtext are likely able to be modified to perform these tasks, but such applications and evaluations of them have not been published. A preliminary analysis of the issues in adapting MedLEE for coding of SNOMED terms was recently described [14].

The value of selective indexing of radiology reports with an exhaustive vocabulary such as the Metathesaurus is that retrieval of a much wider diversity of images can be facilitated. Such a capability would not only be helpful in clinical care (e.g., the clinician seeking to look at other images of a finding or diagnosis to compare the one just obtained) but also in medical education (e.g., the instructor seeking different images of a finding or diagnosis for demonstration to students). This paper thus reports on the adaptation of a system for recognizing concepts from the Metathesaurus in narrative text to the processing the text of radiology reports.

The system used in these experiments, SAPHIRE, was originally designed to perform concept-based automated indexing of knowledge-based documents, such as journal articles and textbooks [15, 16]. In this work, SAPHIRE was modified to facilitate the selective indexing of image reports with the goal of allowing their accurate retrieval. This work was motivated specifically by the desire to add image retrieval functionality to ChartEngine, a multimedia medical record system developed at the University of Pittsburgh Medical Center. Preliminary work showed that most of concept descriptors needed to describe the content of image reports were present in the UMLS Metathesaurus [17]. Thus the general goal of the work is to develop automated approaches to indexing images using terms from the Metathesaurus discovered by adding enhancements to the SAPHIRE concept-matching system. In particular, iterations of development and evaluation were performed to:

1. Establish a image report collection for experimentation and baseline results;
2. Assess enhancements to improve precision of output;
3. Assess enhancements to improve recall of output.

2. SAPHIRE

SAPHIRE is a computer system designed to take medical text as input and return matching concepts from the UMLS Metathesaurus [15, 16]. It was originally designed for the literature retrieval domain [18], and this application of it here is its first test processing text in the electronic medical record. SAPHIRE takes any amount of text for input (e.g., retrieval system query, document sentence) and returns a ranked list of matching concepts from the UMLS Metathesaurus. It was initially implemented in the single-user Macintosh environment [18, 19] and has been evaluated extensively [20, 21]. The results of these experiments showed that while the concept-matching algorithm used for automated indexing did not confer benefit over single-word automated indexing, it could be useful for assisting users in mapping the free text of queries into terms from controlled vocabularies, such as the UMLS Metathesaurus [22]. Using the UMLS Metathesaurus [23], with its breadth of concepts as well as depth of synonyms, this allows a number of different expressions of medical concepts to be recognized and normalized to a canonical form.

The original SAPHIRE concept-matching algorithm used a strict pattern-matching approach requiring not only that all words in a matching term be present but also that they occur in the word order of the term in the vocabulary. The motivation for this high-precision approach was to avoid false-positive matching. This resulted, however, in missing some true-positive matches, which was shown in failure analyses to cause nonretrieval of relevant documents [20]. We subsequently developed a new algorithm to allow partial matching of concepts and not require exact word order [15], building off of similar approaches used by others [24, 25]. One limitation is that this approach necessarily results in an increased amount of false-positive matching, which we aimed to control with a weighting scheme that would give

highest weight to those concepts matching the input document or query most closely. The indexing algorithm or user performing a query could then decide upon the trade-off between false-positive and false-negative matching, based on how far down the weighted list of concepts they chose to go to include terms. The new algorithm was also implemented as a server to run in modern client-server environments [15].

Before describing the specifics of the algorithm, definition of Metathesaurus-related terms would be helpful. The Metathesaurus is organized into *concepts*, which have a unique identifier (the CUI). Each major synonym form that is not just a simple lexical variant (i.e., plural or word order change) is a *term*, each of which also has a unique identifier (the LUI). There can be one or more LUIs for each CUI. Each lexical variant of each term is a *string* (with a unique identifier SUI), and there can be more than one SUI for each LUI. As an example, consider the concept *atrial fibrillation*, which has terms *atrial fibrillation* and *auricular fibrillation*. The former term has the lexical variants *fibrillation, atrial*, and *atrial fibrillations*.

The MRCON file from the Metathesaurus, which is its central table, contains a row for each string along with its SUI, LUI, and CUI. (Each CUI has a canonical or preferred string; for SAPHIRE's purposes the remaining strings are synonyms.) The MRXW.ENG file, which is the inverted word list table of English words, contains a row for each word that occurs in an SUI/LUI/CUI triple. The SAPHIRE algorithm uses these tables intact, but adds some additional files to allow their rapid access. In particular, B-Tree files are added that allow quick look-up of words, LUIs, and CUIs.

Figure 1 depicts the actual algorithm in pseudocode. The algorithm begins by breaking the input string (which can be a sentence or phrase from a document or a user's query) into individual words. Words are designated as *common* if they occur with a frequency above a specified cutoff in the Metathesaurus. The purpose of designating words as common is to reduce the computational overload for words which are occasionally important in some terms but occur frequently in others, such as the word *A* in *Vitamin A* or *acute* in *acute abdomen*. Since the words *A* and *acute* occur commonly in many other terms, calculating weights for these additional terms adds a large and unnecessary computational burden.

For each word in the input string, a list of Metathesaurus terms in which the word occurs is constructed. The Metathesaurus term lists for common words contain only those terms that also occur in one or more of the non-common words in the input string. Using one of the above examples, if the string were *acute abdomen*, the common word *acute* would only contain the term *acute abdomen* and not the term *acute leukemia*.

Once the term lists for each word are created, a master term list is created that contains any term which occurs in one or more individual word lists. If there is a single matching concept (i.e., an exact match), then that concept is the only concept returned, with a weight higher than any partially matching concept could attain. Terms in which less than half of the words occur in the input string are discarded. (Thus, a partial match must have half or more of the words from the term in the input string.) The terms are then weighted based on formula that gives weight to terms that are longest, have the highest proportion of words from the term in the string, and have the words of the term occurring in close proximity to each other.

The algorithm has a number of switches that allow modification of its parameters. The common word cutoff is the Metathesaurus frequency threshold to designate terms as common. It can be set at any level but the default value is 270, which means that any word which occurs more than 270 times in all of the Metathesaurus strings is designated as common. This number was set based upon empirical observation of the algorithm's behavior and results in the most frequent 10% of words in the Metathesaurus being designated as common.

The other switches affect the format of the list of matched terms. The CUI switch causes output to be listed by CUI instead of LUI. (All LUIs for a given CUI are merged, with the weight for the CUI becoming the weight of its highest LUI.) Two other switches control the size of the output list, with one a cutoff for the number of terms displayed and the other a cutoff for the lowest weight allowed. A final switch allows a specific source vocabulary to be set (e.g., MeSH or DxPlain), in which all terms not in that vocabulary are discarded from the returned list.

3. DEVELOPMENT AND EVALUATION METHODOLOGY

Performance of the indexing term assignment process was assessed by using adaptations of the recall and precision measures that are commonly used to evaluate document retrieval. Since the goal of the experiments was to determine concept assignment and not document retrieval, we used the concept-oriented redefinitions of these measures used by Sager *et al.* [26] and Friedman *et al.* [5]. As such, recall was defined as the proportion of concepts in the collection

```
Input:
        string of words
        common word cut-off (default = 270)
        CUI (default = false)
        size cut-off (default = 40)
        weight cut-off (default = 1.0)

Output:
        list of Metathesaurus terms with weights

for each word in string
        word_frequency = number of Metathesaurus concepts that word occurs in
        if word_frequency is greater than the common word cut-off then
                word is common
        else
                create a list of all Meta terms that contain this word
for each word in string that is common
        create a list of all Meta terms that contain this word alone or
                        that contain this word and at least one word that is not common
create a master list of all Meta terms that occur in any of the words
if all terms occur in a single concept then
        return concept with weight = 11
else
        for all Meta terms in the master list
                twis = number of words from this term in string
                wit = number of words in term
                if (wit + 1) div 2 < twis then discard this term
        for all Meta terms in master list
                term_words_present = twis / wit
                log_term_length = log (wit) + 1
                intervening_words = number of words between first and last words of term
                        in string
                log_intervening_words = log (intervening_words + 1) + 1
                weight = term_words_present * log_term_length / log_intervening_words
        sort Meta terms in master list
        if CUI is true then
                eliminate all LUIs which have a common CUI with a higher-ranked LUI
        eliminate all terms below size cut-off
        eliminate all terms below weight cut-off
```

FIG. 1. Pseudocode for the SAPHIRE concept-matching algorithm.

properly assigned (number of terms correctly assigned/total number of correct terms) and precision was defined as the proportion of terms that were assigned correctly (number of terms correctly assigned/number of total terms assigned).

Fifty radiology image reports were randomly selected from a large repository at the University of Pittsburgh. The collection consisted of six image types:

- Chest X-ray—10 reports
- Head CT—10 reports
- Chest CT—10 reports
- Abdominal CT—10 reports
- Head MRI—5 reports
- Bone scan—5 reports

For the purposes of developing a subset of images that could be used to develop and test new algorithms, the collection was divided into training and test sets. The training set consisted of three images each from Chest X-ray, Head CT, Chest CT, and Abdominal CT and two images each from

Head MRI and Bone scan. Only reports from the training set were viewed by the experimental team; the test set was only seen by the indexer.

Indexing was performed by a team member not involved in the development of the indexing algorithm (CAS). The indexer was a medical librarian knowledgeable about the UMLS Metathesaurus and she selected terms likely to represent aspects of the report on which a clinical user would potentially be interested in retrieving. An example of a report and its indexing terms is shown in Fig. 2.

Experiments began by obtaining values for recall and precision with a baseline system. After a failure analysis, an enhanced system was implemented with the aim of improving precision. Another set of experiments was run with this system, followed by another failure analysis. A third system was implemented to enhance recall. A final round of experiments was run using this system, followed by a final failure analysis.

Effort was devoted to adhering to generally accepted principles of evaluating natural language processors in the clinical domain [27]. Bias was minimized by allowing the developers only to see the training set and not the test set of documents. The developers also knew the recall and precision results of the test set from previous experiments, which allowed directed improvement of the system. The reference standard was developed by a medical librarian who was not a developer of the system. The construction of the image report collection and the evaluation methods are described above. Analysis of failures was performed throughout the process and limitations of the study are described under Conclusions.

4. BASELINE ALGORITHM AND RESULTS

As noted above, the overall goal of the indexing process was to define UMLS Metathesaurus terms that were suitable for facilitating retrieval. In particular, the SAPHIRE system was used to identify terms by designating phrases and attempting to recognize important Metathesaurus terms within them. Later enhancements would attempt to promote the designation of important terms and remove unimportant terms in the training set, with the goal of achieving similar results in the test set.

4.1. Methods

Figure 3 lists the pseudocode for the baseline algorithm. In summary, the text was cleaned by removing punctuation, extra spaces, and extra (meaningless) words. Each sentence was then converted to a series of phrases separated by common (also called "stop" or "barrier") words. This approach has been used by others in identifying appropriate-sized phrases to pass to concept-matching algorithms for the matching of controlled vocabulary terms [28, 29]. Each phrase was passed to SAPHIRE to obtain a list of matching Metathesaurus concepts. Starting with the highest ranking term, the words of the input phrase and term were compared, with all words from the term present in the phrase removed from the phrase. The iteration continued down the output list to identify additional terms that had words in the phrase, stopping when the end of the list was reached or all words from the phrase could be associated with a term. Terms would only be selected for indexing when they were an exact match; i.e., the words in the phrase and the concept string were identical. After each phrase of each sentence was processed, a concept profile was created for each report, as shown by example in Fig. 2. After the concept profile was generated, recall and precision were calculated for each report by a Perl program.

4.2. Results

Table 1 shows recall and precision for the training and test data of the baseline system. Recall was substantially higher than precision.

4.3. Failure Analysis

The goal of each failure analysis was to develop categories of errors that would guide improvements to the system. To keep us from "overcorrecting" the system, failure analysis was only done on the training data.

The recall analysis focused on the 50 concepts which should have been identified by the system but were not. There were two broad categories of problems: concepts were either returned by SAPHIRE but not ranked high enough to be designated by the algorithm (i.e., other concepts that were not appropriate were ranked higher) or the concepts were not returned at all. A total of 30 concepts fell into the former category, while 20 were in the latter. An example of a concept returned but not ranked high enough was the text *right calvarium* not leading to an adequate ranking for the concept *Calvaria*. An example of a concept not returned at all was the concept *CT of Head* not matching due to the intervening barrier word.

The precision analysis focused on the 617 concepts that were returned by the system but should not have been so.

Report

CT SCAN OF THE ABDOMEN WITH LOW OSMOLALITY CONTRAST.
HISTORY: METASTATIC RENAL CELL CARCINOMA.
NOW PRESENTS WITH ACUTE RIGHT ABDOMINAL PAIN.
Serial unenhanced images of the liver were obtained.
After the administration of intravenous contrast, biphasic liver imaging was performed and carried out to the level of the pubic symphysis.
Comparison is made to a prior biphasic CT.
Compared with the prior study, there has been increase in size and number of lung base metastatic lesions.
A lesion seen abuting the left hemidiaphragm measures 30 cm x 22 cm (series 2, image 6).
Several less than 10 cm hypodense lesions are seen again in the liver and most likely represents cyst.
A larger less well defined hypodense lesion is seen in segment 6 of the liver.
This measures 20 cm x 20 cm and has also increased in size (series 3, image 48).
An additional approximately 10 cm hypodense area is seen adjacent to the falciform ligament and likely represents an area of focal fat, however, an additional metastatic lesion cannot be excluded.
Patient is status post left nephrectomy, and left adrenalectomy.
An omental mass that is anterior to the stomach has increased in size from the prior study and is more heterogeneous in appearance.
It measures 35 cm x 30 cm (series 3, image 48).
A 10 cm hypodense mass in the right kidney is unchanged from the prior study.
Linear filling defect in the distal abdominal aorta is unchanged and consistent with chronic dissection.
The pancreas, gallbladder, spleen and right adrenal gland appear normal.
No free air or fluid is seen.
IMPRESSION:.
1.
INCREASE IN SIZE AND NUMBER OF LUNG BASE METASTASES.
2.
INCREASED SIZE IN OMENTAL MASS ANTERIOR TO THE STOMACH.
3.
INCREASED SIZE OF HYPODENSE LIVER LESION IN SEGMENT 6 LIKELY REPRESENTING METASTATIC DISEASE.
SEVERAL SMALLER HYPODENSE LIVER LESIONS ARE UNCHANGED AND LIKELY REPRESENT CYSTS.
4.
UNCHANGED CHRONIC DISTAL AORTIC DISSECTION.
5.
UNCHANGED SMALL HYPODENSE RIGHT RENAL PARENCHYMAL MASS.
J30.
My signature below is attestation that I have interpreted this/these examination(s) and agree with the findings as noted above.
END OF IMPRESSION:.

Indexing Concepts (UMLS CUI and preferred name)

C0023884	Liver
C0034015	Pubic Symphysis
C0010709	Cysts
C0230240	Falciform ligament
C0022646	Kidney
C0003484	Aorta, Abdominal
C0012737	Dissection
C0340643	Dissection of aorta
C0220651	Metastasis to lung
C0011980	Diaphragm
C0262613	RENAL MASS
C0202839	Computerized tomography of abdomen
C0278613	metastatic disease
C0028977	Omentum

FIG. 2. Example report and indexing terms.

```
For each report
    Clean the text of the report
        Remove punctuation, extra spaces, and extra words
        Identify individual sentences
    For each sentence of the scan
        Break the scan up into phrases based on stop words
    For each phrase
        Send phrase to SAPHIRE
        For all terms in ranked list of matches
            Add top-ranking term to output if a SAPHIRE exact match
            Discard words from input that matched
            Continue until all words exhausted or bottom of list is reached
    For all concepts matched by all of the phrases
        Omit redundant concepts
```

FIG. 3. Pseudocode for baseline algorithm.

Many of these concepts were present in the text and were appropriately identified by the algorithm. But since the goal of our work was to develop an algorithm to selectively index the most important concepts, we needed to determine approaches that would eliminate the concepts present in the text for which we did not want to index. After analyzing the output, it became clear that a couple approaches showed promise.

The first observation was that many of the desired concepts fell into a constrained set of UMLS semantic types. It was also noted that many undesired concepts were in the remaining types. Table 2 shows, by semantic type, the number of inappropriately retrieved concepts, the number of appropriately retrieved concepts, and the total number of indexed concepts. This table led to the development of a semantic filter to be described in the next section.

A second observation was that 67 concepts were retrieved inappropriately because the concept was negated. A related problem was that concepts were described as being "normal." Making matters even more difficult was that many of these negations and abnormalities occurred across barrier words, in particular *and* and *or.*

TABLE 1
Results for Baseline, Precision-Improving, and Recall-Improving Runs

	Found	Possible	Returned	Precision	Recall
Training data					
Baseline run	88	138	705	0.12	0.64
Precision-improving run	87	138	358	0.24	0.63
Recall-improving run	102	138	291	0.35	0.74
Test data					
Baseline run	228	361	1660	0.14	0.63
Precision-improving run	203	361	898	0.23	0.56
Recall-improving run	227	361	765	0.30	0.63

5. PRECISION-IMPROVING ENHANCEMENTS

Because precision was so substantially lower than recall, and an operational version of this system would likely be unacceptable to users because of the low precision, the first set of enhancements aimed at improving precision. This was done by implementing improvements in the algorithm to address the problems identified in the failure analysis. In particular, a semantic type filter and negation recognizer were implemented, and the history portion of the report was not processed whenever it could be recognized as such. One other change that was actually a recall-improvement enhancement was also made, which was the modification of certain common phrases to ensure appropriate concept recognition by SAPHIRE.

5.1. Methods

The first of the three enhancements was the establishment of a semantic type filter that would only allow terms of certain semantic types to be included as indexing terms. This was based on the observation that some types of terms would be unlikely to be chosen as indexing terms by virtue of the fact that all terms of their semantic type would be inappropriate for indexing of radiologic images. These were terms that had semantic types which always led to incorrectly retrieved concepts as listed in Table 3, such as *Classification* or *Substance*. It should be noted that even though some terms had semantic types that were predominantly assigned incorrectly, they were types that were likely to appear in radiology reports, and we did not want to filter them out via this mechanism. The list of semantic types in the filter was determined by manual review of the semantic types of the training data.

The second modification was to develop automated means for recognizing negation. The rationale for this was that some reports state that a finding or diagnosis was not present. SAPHIRE does not detect negation; i.e., the phrase "no infiltrate" would lead to the term "infiltrate" being detected. An algorithm for negation was developed and tuned for optimal performance with the training data. The algorithm

TABLE 2
Semantic Types of Concepts Correctly and Incorrectly Retrieved by Baseline Run

Semantic types	Incorrectly retrieved concepts	Correctly retrieved concepts	Total number of indexed concepts
Acquired abnormality	2		2
Acquired abnormality/Disease or syndrome			2
Acquired abnormality/Finding/Disease or syndrome		1	1
Acquired abnormality/Tissue/Finding			1
Anatomical abnormality	3	4	5
Biomedical occupation or discipline	3		
Body location or region	27	13	21
Body part, organ, or organ component	84	30	41
Body space or junction	4	6	6
Body substance	1		
Body system	1		
Cell or molecular dysfunction	1		
Classification	1		
Congenital abnormality	6		
Congenital abnormality/Intellectual product	1		
Diagnostic procedure	18	6	19
Disease or syndrome	19	8	9
Disease or syndrome/Sign or symptom	2		
Experimental model of disease/Functional concept/Sign or symptom	1	2	2
Finding	56	2	3
Finding/Disease or syndrome		1	1
Finding/Pathologic function	3		
Finding/Qualitative concept	1		
Functional concept	10		2
Health care activity	13		
Immunologic factor	2		
Injury or poisoning	8		
Injury or poisoning/Disease or syndrome	2		
Intellectual product	48		
Laboratory or test result	2		
Laboratory procedure	1		
Medical device	4		
Mental or behavioral dysfunction	1		
Mental process	3		
Mental process/Organ or tissue function	1		
Natural phenomenon or process	3		
Neoplastic process	27	8	9
Occupation or discipline	14		
Occupational activity	3		
Organ or tissue function		1	1
Organic chemical	10		
Organic chemical/Antibiotic	1		
Organic chemical/Indicator, reagent, or diagnostic aid	1		
Organic chemical/Pharmacologic substance	1		
Organism attribute	5	1	1
Organism function			
Organism function/organ or tissue function	1		
Organism function/social behavior/individual behavior	1		
Pathologic function	2		1
Pathologic function/sign or symptom			1
Patient or disabled group	5		
Physiologic function		2	4
Population group	1		
Qualitative concept	60	1	3

TABLE 2—*Continued*

Semantic types	Incorrectly retrieved concepts	Correctly retrieved concepts	Total number of indexed concepts
Qualitative concept/functional concept	9		
Qualitative concept/spatial concept	1		
Quantitative concept	39		
Quantitative concept/spatial concept	2		
Regulation or law	1		
Research activity/quantitative concept	2		
Sign or symptom	8	1	1
Spatial concept	47		
Steroid	1		
Steroid/pharmacologic substance	1		
Substance	1		
Temporal concept	22		
Therapeutic or preventive procedure	18		1
Tissue	1	1	2
Total	617	88	138

implemented negation routines similar to those developed for comparable systems, such as CAPIS [24].

A final modification was to limit processing to non-history portions of the report when they could be recognized. This was usually identified by a tag such as HISTORY:.

TABLE 3
Semantic Types Used in Filter

Acquired abnormality
Anatomical abnormality
Anatomical structure
Body location or region
Body part, organ, or organ component
Body space or junction
Congenital abnormality
Diagnostic procedure
Disease or syndrome
Embryonic structure
Finding
Fully formed anatomical structure
Functional concept
Indicator, reagent, or diagnostic aid
Medical device
Neoplastic process
Organism attribute
Organism function
Pathologic function
Physiologic function
Therapeutic or preventive procedure
Tissue

5.2. Results

Table 1 shows the results for the precision-improving run as well. The major effect of these enhancements was to reduce the size of the output for each scan. In the training set, precision nearly doubled while recall was virtually unchanged. In the test set, however, precision only improved by 64.6% while recall actually fell by 11.0%.

5.3. Failure Analysis

The failure analysis after these experiments not surprisingly showed little change in the recall errors, as virtually the same appropriate concepts were retrieved. For precision errors, it showed that the semantic filter and negation algorithms had a positive impact but were imperfect. Further analysis showed that 126 inappropriately retrieved concepts were in "historical phrases," e.g., *s/p mastectomy.* It was also noted that certain terms were common and of little value from the standpoint of concepts in the reports, such as *Measured* and *Approximate*.

Analysis at this stage focused on looking for errors affecting recall. It was noticed, for example, that certain concepts were not matched because the words in the report and the concept differed in an ending "s." It was also found that some concepts were not being indexed because they did not qualify as an exact match for SAPHIRE, even though they were the correct concept.

6. RECALL-IMPROVING ENHANCEMENTS

The goal of the final set of experiments focused on improving recall. These included the addition of stemming the final "s" in all words, choosing the top-ranking concept to be added even if it did not qualify as an exact match, and modifying some phrases to become recognizable by SAPHIRE, such as CT scan of the head being changed to CT head so that the concept *CT of Head* would be matched by SAPHIRE.

One measure aimed at improving precision was also added that was discovered from the failure analysis of the precision-enhancing system. Toward this end, a variety of enhancements were made to the system based on observations from both failure analyses that showed why concepts were not being retrieved.

6.1. Methods

Recall-enhancing improvements added for this step included:

- "s" stemming of all words in reports and concepts;
- Modifying some common phrases to enhance retrieval, e.g., *CT scan of head*;
- Not applying the exact-match requirement to top-ranking concept.

Some precision-enhancing improvements were also made at this step, including:

- Removing certain stop concepts, such as *Measured* and *Approximate*;
- Eliminating concepts in historical phrases, e.g., *s/p mastectomy.*

6.2. Results

As seen in Table 1, the enhancements in this step improved not only recall but also precision. As with the second experiment, the gains were larger in the training data than the test data. Based on the results with the test data, we conclude that the highest performance we could obtain was a recall of 63% and a precision of 30%.

6.3. Failure Analysis

The failure analysis of the final run shows that despite the improvement seen from some modifications of the algorithm, many of the same types of problems continue to exist. Recall errors still occur from phrases that either do not result in the correct term being ranked adequately high in SAPHIRE's output (e.g., *9th rib approximately* does not yield the correct concept *Ribs* as its highest rank) or at all (e.g., *pulmonary coin lesion* does not return *Pulmonary Nodule*). Likewise, precision errors persist because concepts occur in the report which have not been designated as indexing concepts and are not eliminated by the algorithm means developed so far.

7. CONCLUSIONS

The main goal of these experiments was to identify means to selectively recognize indexing concepts in radiology reports to facilitate retrieval. Despite considerable improvements over the baseline, performance would probably be inadequate to use the system in a purely automated indexing process. Recall errors will likely persist because the natural language utterances of dictating radiologists cannot be perfectly mapped to controlled vocabulary concepts. Likewise, precision errors will continue because the means to only select concepts in the report deemed important to humans are difficult to encode algorithmically.

It is possible that other approaches could improve our results. One area worthy of investigation would be the incorporation of natural language processing techniques known to be effective for other applications of radiology report text processing. For example, the more sophisticated phrase generation approaches of SymText might replace the barrier word methods used in these experiments. Likewise, the semantic grammars of MedLEE might improve the matching of controlled vocabulary concepts. Or perhaps probabilistic approaches of mapping words or phrases into concepts would provide gains. These experiments set out a baseline of performance upon which other systems can aim to improve.

Another area where the task in general might be improved is to consider the use of other vocabularies (e.g., SNOMED) and/or other approaches that use a more complex information model than the simple list of terms employed here. While the simple list of terms approach might facilitate retrieval of images by users, it does not allow for the complexity of desired findings and diagnoses to be represented. While the Metathesaurus provides an exhaustive coverage of terminology, any information model to represent a domain with more complexity must be built on top of it.

Future work must also include building a larger image report collection. Continued work on these relatively small

number of documents will eventually result in overtraining on them. A larger image report collection will also allow a richer diversity of concepts in general. Another requirement for a larger image report collection will be the use of more than one indexer so that a more diverse view of concepts that should be indexed is attained. It will likely also be beneficial to include clinicians among the indexers to obtain their perspective on features important for indexing.

It may be that a system such as the one we have developed would more appropriately serve as an assistant to a manual indexing process. It could well be that no amount of purely automated concept recognition will achieve the quality of indexing necessary for a system acceptable to users. Toward that end, we plan to conduct experiments with real indexers and users using this system embedded in an actual image-retrieval setting. This will allow a larger range of experiments, such as the utility of the system as an assistant in the manual indexing process and the comparison of the system with a variety of other approaches, from simple key word matching to more sophisticated natural language processing approaches.

Our results in this paper show that this important task is also a very difficult one. Some of the further pathways described above can potentially improve performance. If we believe there is value in the ability to retrieve images from clinical repositories, then further research in indexing images is essential. Purely manual indexing would be untenable given the massive numbers of images that are created at each medical institution each day. Therefore some sort of automation of the process is essential. Further research must address the best means to do so in an efficient and cost-effective manner.

ACKNOWLEDGMENT

This project was supported by National Library of Medicine Contract N01-LM-4-3507.

REFERENCES

1. Yasnoff WA. Bacus JW. Scene-segmentation algorithm development using error measures. Anal Quant Cytol 1984; 6:45–58.
2. Robinson GP, *et al.* Medical image collection indexing. Comput Med Imaging Graph 1996; 20:209–17.
3. Chu WW, Johnson DB, Kangarloo H. A medical digital library to support scenario and user-tailored information retrieval. IEEE Trans Inform Technol Biomed 2000; 4:97-107.
4. Stensaas S. Animating the curriculum: integrating multimedia into teaching. Bull Med Libr Assoc 1994; 82:133–9.
5. Friedman C, et al., A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994; 1:161–74.
6. Haug P, *et al.* A natural language understanding system combining syntactic and semantic techniques. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care; Washington, DC: Hanley Belfus, 1994:247–51.
7. Jain NL, *et al.* Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proceedings of the 19th Annual AMIA Fall Symposium; Washington, DC: Hanley-Belfus, 1996: 542–6.
8. Jain NL, Friedman VC. Identification of findings suspicious for breast cancer based on natural language processing of mammorgram reports. Proceedings of the 1997 Annual AMIA Fall Symposium; Nashville, TN. 1997:829–33.
9. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferrability and sources of physician disagreement. Methods Inform Med 1998; 37:1–7.
10. Fiszman M, Haug PJ, Frederick PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. Proceedings of the 1998 Annual AMIA Fall Symposium; Orlando, FL: 1998:860–4.
11. Fiszman M, *et al.* Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Assoc 2000;7:593–604.
12. Elkins JS, *et al.* Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000; 33:1–10.
13. Hripcsak G, *et al.* Automated tuberculosis detection. J Am Med Inform Assoc 1997; 4:376–81.
14. Lussier YA, Shagina L, and Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proceedings of the 2001 Annual AMIA Symposium; Washington, DC: 2001, in press.
15. Hersh WR and Leone TJ. The SAPHIRE server: a new algorithm and implementation. Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care. New Orleans, LA: Hanley-Belfus, 1995:858–62.
16. Hersh WR, Donohoe LC. SAPHIRE International: a tool for cross-language information retrieval. Proceedings of the Annual AMIA Fall Symposium. Orlando, FL: Hanley-Belfus, 1998:673–7.
17. Lowe HJ, *et al.* Representing images in the multimedia electronic medical record combining semantic indexing and image content-based representation to support knowledge-based retrieval of medical images. Methods Inform Med 1999; 38:303–7.
18. Hersh WR, Greenes RA. SAPHIRE: an information retrieval environment featuring concept-matching, automatic indexing, and probabilistic retrieval. Comput Biomed Res 1990; 23:405–20.
19. Hersh WR, *et al.* Adaptation of Meta-1 for SAPHIRE, a general purpose information retrieval system. Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care. 1990; 156–60.
20. Hersh WR, *et al.* A performance and failure analysis of SAPHIRE with a MEDLINE test collection. J Am Med Inform Assoc 1994; 1:51–60.

21. Hersh WR, Hickam DH. Information retrieval in medicine: the SAPHIRE experience. J Am Soc Inform Sci 1995; 46:743–7.

22. Hersh WR, Hickam DH, Leone TJ. Word, concepts, or both: optimal indexing units for automated information retrieval. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care. Baltimore: McGraw-Hill, 1992: 644–8.

23. Humphreys BL, *et al.* The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998; 5:1–11.

24. Lin R, *et al.* A free-text processing system to capture physical findings: canonical phrase identification system (CAPIS). Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care. Washington, DC: McGraw-Hill, 1991: 843–7.

25. Vries JK, *et al.* An expert system for indexing and retrieving medical information. Technical Report, University of Pittsburgh School of Medicine, 1986.

26. Sager N, *et al.* Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994; 1:142–60.

27. Friedman C and Hripcsak G, Evaluating natural language processors in the clinical domain. Methods Inform Med 1998; 37:334–44.

28. Tersmette KWF, *et al.* Barrier word method for detecting molecular biology multiple word terms. Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care. Washington, DC: IEEE, 1988:207–11.

29. Moore GW, Berman JJ, Performance analysis of manual and automated systemized nomenclature of medicine (SNOMED) coding. Am J Clin Pathol 1994; 101:253–6.