A Concept-Based Retrieval System for Thoracic Radiology

Ricky K. Taira, David B. Johnson, Vikas Bhushan, Monica Rivera, Clement Wong, Lujia Huang,
Denise R. Aberle, Melanie Greaves, and Jonathan G. Goldin

Current digital information systems in radiology are insufficient to accommodate the retrieval needs of academicians. Significant efforts are required in retrieving clinical cases for teaching and research. We describe a prototype system that supports intelligent case retrieval based on a combined specification of patient demographics, radiologic findings, and pathologic diagnoses. The documents for these cases can be distributed among multiple heterogeneous data bases. The system features automatic indexing of radiology and pathology reports, a comprehensive lexicon for thoracic radiology, an interface to a hospital information system, radiology information system, and picture archiving and communication systems, and a graphical user interface for query formulation and results visualization. The prototype system was developed within the domain of thoracic radiology involving patients with lung cancer.

Copyright © 1996 by W.B. Saunders Company

KEY WORDS: information retrieval systems, query interfaces, free-text analysis.

RADIOLOGY information systems (RIS), picture archiving and communication systems (PACS), and hospital information systems (HIS) are ill-structured to accommodate the retrieval needs of clinicians and researchers. Significant efforts are required for amassing interesting cases for teaching files and lectures, and for data collection of retrospective research studies (eg, receiver-operating characteristic studies). The large repository of images and descriptive text stored in PACS, RIS, and HIS installations are not effectively used because of the limited means by which cases can be retrieved (usually via patient hospital identification number).^{1,2}

A new generation of intelligent data base systems must emerge to effectively disseminate, integrate, retrieve, correlate, and visualize multimedia medical information.^{3,4} Many PACS installations have several terabytes of image data on-line, yet utilization of data for research and teaching is almost nonexistent. An analogy from the library sciences can be drawn when books and magazine articles could be located only via an ISBN number or exact title or author name. The efficiency of locating books or journal articles today is greatly improved with the additional ability to search by key words, sub-

ject, and date. The ability to find relevant articles from among the several thousands of volumes of books, journal articles, and magazine articles is much improved over what was possible 10 years ago by the availability of front-end search engines such as Melvyl and Medlars. Similar search engines for medical image and textural data are needed.

The goal of this report is to describe a prototype system that supports intelligent information retrieval based on document (image or text) content. The system allows users to retrieve patient cases based on a combined specification of patient demographics, low-level radiologic findings, and their attributes (densities, size, shape, number, etc), intermediate radiologic findings (eg, atelectases, infiltrates, etc), and/or high-level pathology diagnoses (eg, well-differentiated small cell carcinoma). The documents for these cases can be distributed among multiple heterogeneous data bases (ie, PACS, RIS, and HIS).

Concept-based retrieval is especially difficult when the task involves searching for image data or for documents that contain imprecise descriptors. To illustrate some of these difficulties, consider the following query: "Retrieve magnetic resonance images and all medical reports of Korean adolescent patients who demonstrate a malignant spherical mass in the vicinity of the sphenoid sinus that has doubled in size since initial diagnosis." Some of the problems that must be addressed include: (1) How do we express this natural language query to a computer system? Conventional data base query languages are limited in their semantics and expressibility,

From the Departments of Radiological Sciences and Computer Science, and the School of Medicine, University of California, Los Angeles.

Supported in part by National Institutes of Health Grant No. USPHS CA39063 awarded by the National Cancer Institute, and by the National Science Foundation Grant No. IRI9224559.

Address reprint requests to Ricky K. Taira, PhD, Department of Radiological Sciences, University of California, Los Angeles, 10833 Le Conte Ave, Los Angeles, CA 90024-1721.

*Copyright © 1996 by W.B. Saunders Company
0897-1889/96/0901-0004\$3.00/0*

especially concerning fuzzy correlations (eg, *in the vicinity of*), spatial patterns (eg, shapes), and temporal concepts (eg, *size doubled since initial diagnosis*). (2) What methods are used to match cases stored within our information repository (ie, HIS, RIS, and PACS) to the conditions specified in the user query? The system needs to be able to mimic the response of a trained expert in the medical field. That is, the system must be knowledgeable of all subclasses of a specified object (eg, a malignant tumor could be an adenocarcinoma, or any number of other cancerous growths) and reference variations (eg, synonyms, lexical variants, abbreviations, etc). (3) How does the system intelligently relax certain constraints when no exact query solutions exist? For example, for the system to find a solution, it may need to relax the shape of the mass from spherical to elliptical, or relax the constraint of patient race from Korean to Oriental. (4) What is the user's role in performing queries? A flexible user interface is required to allow the user to express the query, refine the query, and navigate through the presented query solutions. The user requires tools to visualize the presented answers in a specific format (eg, pictorial, graphical, tabular, etc).

SYSTEM OVERVIEW

Figure 1 shows the overall system architecture of our prototype system. We have previously reported on the Data Access Layer that allows document retrieval (PACS images and

Fig 1. The prototype case retrieval system is composed of (1) a data access layer to HIS, RIS, and PACS data using patient hospital ID; (2) an indexing system for radiology and pathology reports; (3) an intelligent query processing engine; and (4) a graphical interface for the expression of high-level research queries.

radiology, pathology, admission, discharge, medical, and laboratory reports) from HIS, RIS, and PACS.⁵⁻⁷ The Data Access Layer has limited query processing capabilities and supports only read access to HIS, RIS, and PACS given a patient hospital identification number. To extend the query processing capabilities of the system, we developed three supporting and cooperative software subsystems: (1) a free-text indexing system for radiology and pathology reports to summarize and structure the important findings in these documents, (2) an intelligent query processing engine that uses a lexical knowledge base, and (3) a user interface layer that supports the expression of high-level research type queries.

REPORT INDEXING

The fundamental deficiency in present radiology information retrieval systems is the lack of structured abstracts that summarize the important observations, impressions, and conclusions associated with a radiology procedure. Each procedure currently has a highly variable free-text report associated with it that contains descriptions of image features, anatomical structures, and pathological processes. Summarizing and structuring the semantic content within these reports is needed to create indexable abstracts that can be used to quickly locate cases by image content. An analysis of free-text radiology reports provides an indirect approach for summarizing the content of corresponding radiologic images. The correspondence between the report and the associated radiology images is made using the patient hospital identification number, date and time of examination, and radiology procedure description.

Analysis of natural language reports is a difficult task. The main problem is the inherent complexity associated with the definition of a general and sufficiently complete model for representation and comprehension of human language. The problem can be considerably simplified by restricting the communicative information to a specified domain (eg, reports from thoracic radiology) confined to a focused set of topics (eg, chest lesions and lymph node involvement).

We developed free-text analysis modules for both pathology and radiology reports. For our

prototype system, we focused on a specific target population, lung cancer patients seen at the UCLA Boyer Clinic. These patient's all have periodic computed tomography (CT) examinations and some projectional imaging procedures. Most had biopsies for conformation of pathology. Pathology reports were restricted to procedures involving analysis of lung mass biopsy samples. Radiology reports were restricted to procedures from thoracic projectional and CT examinations. In total, 30 patients with confirmed lung cancers were identified. For these 30 patients, we retrieved 153 pathology reports and 178 thoracic radiology reports. Our goal was to automatically generate a comprehensive index summarizing the important findings transcribed in these reports. There was no departure in the style or vocabulary used from the traditional mechanism for radiology/pathology reporting. We briefly describe the algorithm below (see Figs 2 and 3). Further details can be found in similar free-text analysis systems described in the literature.⁸⁻¹²

Report Filtering

To locate radiology reports that match the target population and findings, filtering programs were constructed for radiology and pathology reports (Fig 2). (We describe only the more difficult radiology reports in this article.) Reports from our RIS were sent to a preprocessing filter that consists of three stages. The first stage passes only reports that originate from procedures in which the system is qualified to analyze (ie, reports from thoracic CT and projectional

procedures). The header information of the report file, which contains the procedure description of the imaging modality, the anatomic body region, referring radiology section, and the imaging protocol, is used to perform this initial decision-making task. Nonqualified reports are not processed. Qualified reports are then sent to the second preprocessing filter stage that identifies the various sections of the report and passes only the "Findings" portion of the report text. This is performed by identification of section heading labels and use of a priori knowledge of formatting conventions. If no "Findings" section can be found, the entire report body is analyzed. The "Findings" section is used instead of the "Impressions/Conclusion/Summary" sections because the "Impression/Conclusion/Summary" sections are typically more telegraphic in its explanatory style and thus often more ambiguous (assumed descriptive words are often left out, eg, right lobe). Also, the "Impression/Conclusion/Summary" sections of the report contain redundant information with respect to the "Findings" section. The third filter stage takes the free text of the "Findings" section of the qualified reports and parses the text into paragraphs (finding enumerations) and sentences. Each sentence is given a report address consisting of the name of the file, the section label, the paragraph number, and the sentence number. Each sentence is then collated into target specific *subreports*. Each of the component subreports contain sentences that either directly or indirectly reference (via pronouns) a specific predefined target radiology finding. In our case, three types of subreports were generated: (1) masses, (2) lymphadenopathy, and (3) other. Each subreport focuses around a single central theme (eg, lung masses in projectional chest radiographs) and may have overlapping sentences with other subreports if a sentence has multiple target findings. This third preprocessing step can be described as a type of "topicological transform" (decomposition) of the original text report. The final preprocessing operation is to direct each subreport to a corresponding specialized natural language processing (NLP) agent whose topic of expertise matches the findings topic of the subreport. (For example, the mass subreport is directed to the mass NLP agent.)

Fig 2. The overall processing stages of our text analysis system. It contains three preprocessing stage filters and finding specific natural language processing agents.

To summarize the preprocessing text analysis steps, multiple subreports are created from a given report to simplify the domain scope of each subreport. This is performed to reduce the amount of ambiguities within a subreport that are common in unrestricted free text. The task beyond the preprocessing steps reduces to developing separate NLP modules for each focused topic domain.

Lexical Analysis

Development of the "mass" NLP agent involved modeling the microworld consisting of a radiologist, a radiograph, and a presented mass on the radiograph. An understanding of the concepts, nomenclature, a priori knowledge assumed within the microworld, and communicative goals were modeled and explained below. Our NLP agent assumes that given this microworld, standards for a radiologist's behavior (communication methods, sense perception, and abstract thought) are followed.¹⁰

Figure 3 illustrates the major processing steps of our NLP agents. The first step involves a lexical analysis that transforms as many words and word phrases from meaningless text strings to meaningful concepts. We used two electronic lexicons. The first is the publicly available Unified Medical Language Sources (UMLS) metathesaurus, version 1.4 developed by the National Library of Medicine.¹³⁻¹⁵ The UMLS metathesaurus contains over 300,000 medical terms derived from multiple controlled vocabularies, including the International Classification of Diseases—9th Revision (ICD-9), American College of Radiology (ACR) codes, Medical

Subject Headings (MeSH), Current Procedural Terminology (CPT), and Systematized Nomenclature of Medicine (SNOMED). It serves an excellent lexical source for general medical concepts, but is deficient in its entries for radiology because of the lack of spatial, textural, density, and shape-related visual concepts.¹⁶

The second lexical source is one that we have developed for the domain of thoracic radiology. It supplements the UMLS metathesaurus in the following areas: (1) radiology findings related to chest pathology, (2) thoracic anatomy, (3) spatial, compositional, and temporal concept descriptors, and (4) lead-in and lead-out phrases to findings, anatomy, and pathology. The UCLA Thoracic Radiology Lexicon was constructed by: (1) identifying the terms and common phrases to be included in the Lexicon (glossary construction) and (2) organizing these entries into a hierarchy of related concepts (thesaurus construction).

Words and word phrases included in the Lexicon were gathered from two distinct sources. The first was from published sources including the indices of thoracic radiology textbooks,^{17,18} radiology review manuals,¹⁹ radiology word compilations,²⁰ and published thoracic radiology glossaries.^{21,22} The terms from these published sources ensure the generality of the concepts covered by the Lexicon. However, it does not include all the string representations for these concepts. The second source of terms to include in the Lexicon must come from actual dictated thoracic radiology reports from different radiologists. The collection of words and word phrases from actual reports insures that the system works at a practical level and that most of the string representations for the basic concepts related to the target findings are included. For example, the concept "well-defined"—a border attribute for a mass—has over 40 string representations (eg, well-circumscribed, well-demarcated, well-marginated, sharply defined, etc). The glossary step of the development of the Lexicon connects all lexical variants, abbreviations, and synonymous phrases to their common concept identification. Figure 4 shows the structure of the UCLA Thoracic Glossary. Its structure is similar to the structure of Meta-1.4 of the UMLS knowledge sources.²³ It provides a structured method for representing unique concepts

Fig 3. Processing stages of a natural language processing agent.

Fig 4. Organization of strings, terms, and concepts within the system. There are many string and term representations for a single medical concept.

by a number of string level representations. Currently, the lexicon contains over 4,000 entries (both single- and multiple-word phrases). Each word or word phrase in the lexicon naturally falls into distinct conceptual classes (ie, semantic classes). The classification of words and word phrases was determined by reviewing the classification schemes proposed by other working groups in medical natural language processing^{8,12} and through analysis by our own radiologists. UMLS contains some 134 different semantic types to cover all of medicine. Sager et al⁸ defined 31 semantic categories. We currently have 101 concept categories modeled as a semantic network to cover the encountered terms within our domain.

Pattern Analysis and Frame Construction

The second step in processing radiology free text involved understanding the relationship between the concepts being described within a sentence. The initial lexical analysis step provides either grammatic or semantic labels of the identified words or word phrases. In the second step, words and word phrases were progressively arranged into bigger related blocks. The aggregation of terms was based on proximity, conceptual dependencies, and syntactic and semantic patterns.^{12,24} Radiology syntactic and semantic pattern templates were stored in a knowledge base and used to interpret the most frequently encountered sentence structures. This radiology grammar was modeled within the knowledge base by an extensible object-oriented system of states and connected processes. This state-based system is similar although more powerful than a finite-state automata processor.^{25,26} For each syntactic/semantic unit, a state

Fig 5. Simple example illustrating the construction of conceptual links and sentence interpretation from structure. A knowledge-base maintains templates for common sentence structures and the relationship between sentence parts.

change takes place followed by a process that may extract and link the unit to other units, add the unit to an existing frame (see below) or pass the unit on for further processing. Figure 5 illustrates this linking process with a simple example.

The third step in processing radiology free text is to create a formal data model for each target finding instance. For each radiology finding, a knowledge frame is maintained that models the possible attributes and possible values associated with the finding. Rules for instantiating and populating findings frames were developed based on proximity of attributes to central findings and conceptual dependency rules. A description of the use of frames in radiology text analysis can be found in Haug et al.¹¹ Thus far, we have modeled attributes of masses (see Fig 6), and involved lymph nodes (lymphadenopathy). The completed frames were

Fig 6. Knowledge frame and slots for the radiology finding "mass."

then used to create the document index managed by a relational data base management system (Sybase; Sybase Inc, Emeryville, CA).

THREE-LEVEL DATA MODEL

An important component of any concept-based retrieval system is the overall data model used to abstract and infer knowledge about objects of interest within the data base. Our prototype system uses a three-level model (Fig 7). The first level is the Image Stack Level that stores the raw image data within our PACS. The images are organized in a dynamic stacked logical data structure.²⁷

The second level, the Semantic Object Description Level, provides a formalized description of the important findings in the images stored in the Image Stack Level. These descriptions are derived from the text analysis steps of corresponding radiology reports as described above. This level models finding instances. That is, each radiology report will result in a data model dependent on the findings of the radiograph under examination. An object-oriented semantic data model approach is used to model these finding instance descriptions.²⁸

The top level of our data model, the Knowledge Representation Level, models medical concepts at the class (generic level) as opposed to the instance level. This level functionally captures the thesaurus information within our lexicon where relationships between conceptually dependent entries within the lexicon are identified. These relationships are needed by the NLP agents and query processing system (see below) to infer additional information about a given concept (eg, right lower lobe of the lung, is part of the right lung, which is part of the thorax). The system currently uses tradi-

tional object-oriented data modeling constructs including: (1) "Is-a" relationships: This construct builds generalization/specialization hierarchies that relate an object type to more generic types. Subclasses inherit characteristics of their parent superclasses. For example, the histologic classification of lung cancer is modeled with this construct (small cell carcinoma "Is-a" type of malignant epithelial cancer). (2) "Part-of" relationships: This construct builds aggregation hierarchies that model the composition of a complex object into its constituent parts. These constructs are useful for describing anatomy and complex features (eg, nodule appearance is described collectively as a function of brightness, uniformity, texture, shape, etc). More advanced temporal, evolutionary, and spatial data modeling constructs are also being developed.²⁸

SYSTEM OPERATION

To test the operation of our system, we compiled a test bed of cases from lung cancer patients seen at the UCLA Boyer Clinic. We located 30 patients with confirmed lung cancers. For these 30 patients, we retrieved 153 pathology reports and 178 thoracic radiology reports with references to our target findings of "masses" or "abnormal lymph nodes." All pathology, medical, admissions, and laboratory reports and CT and computed radiography images for these patients were retrieved using our integrated HIS/RIS/PACS interface.⁶

Query Formulation

The user interface and query processing layers allow users to construct research queries and browse the information system. The interface is concerned with the expressibility of English-like queries into a formal query language. We used the PICQUERY + multimedia interface design proposed by Cardenas et al,²⁹ customizing its presentation for thoracic-radiology lung cancer applications. The prototype system, allows users to specify three classes of query constraints: Patient demographics, disease pathology, and radiological findings. Figure 8 shows the main menu screen.

The patient class filter allows the user to specify the demographic profile of patients whose data are to be retrieved (Fig 9). Icons can

Fig 7. Three-level data model for medical images.

Fig 8. Main user window for thoracic radiology research retrieval system.

be selected to visualize the relationship between values for a particular object attribute. For example, if the patient ethnicity icon is selected, a graphical representation showing how different ethnic groups are related is presented as a directed graph (eg, Orientals include Chinese, Japanese, Korean, etc). A user may select general class values (eg, ethnicity = "Oriental") or specific values (eg, ethnicity = "Chinese"). Fields requiring temporal values (eg, age) can be entered using single time points (eg, 18 months), a time period (eg, from 18 months to 24 months), or a more abstract value such as a developmental phase (eg, adolescent).

The pathology class filter allows the user to further specify pathology constraints to the query. In the main menu, when the "Pathology

of Interest" icon is selected, a custom pathology specific form is presented. The lung cancer specification form is shown in Fig 10. The upper portion of this form provides the user with a hierarchical set of lists corresponding to histological classifications of lung cancers. This listing was obtained from data extracted from the ACR Index of Radiologic Diagnoses³⁰ and from publications on lung cancer histologic classification.³¹ The histology class filter allows the investigator to be as general or as specific in qualifying the histology characteristics for the case to be retrieved. The user can be very specific (eg, oat-cell small cell carcinoma) or very general (eg, malignant epithelial tumors). Subclasses of general concepts are automatically known via the thesaurus information within the lexicon. Radio buttons are provided for the user to facilitate the specification of the cell differentiation classification for the lung tumor. Additionally, radio buttons are provided to allow the user to specify the location of the primary tumor site and location(s) of metastatic sites. When multiple values for an attribute are allowed (eg, tumor location within left lower lobe and right upper lobe) the user must specify whether to perform an "OR" or "AND" search constraint.

When the "Radiology Finding" icon is selected from the main menu, a selectable custom radiology findings form is presented. The specification of a radiological findings can including various attributes including: (1) Object composition concepts—gray level, texture, heterogeneity, homogeneity, CT number; (2) spatial concepts—size, shape, location, distance; (3) temporal concepts—changes (eg, anything new, anything resolved, size, position), direction of change (eg, growing, shrinking), and rate of change (eg, fast, slow); (4) temporal relationships (eg, process A "co-occurs with" process B); and (5) spatial relationships (eg, mass "adjacent to" aortopulmonary window). Figure 11 shows the user forms for describing lesions and involved lymph node radiographic findings within the context of lung cancer patients. The finding "lesion" has the attributes, "size," "uniformity," "density," "calcification pattern," "external architecture," "clinical trend" and "location." Under each of these categories are possible values from which the user can select to constrain the query search.

Fig 9. Patient demographic specification window.

The screenshot shows a software window titled "Chest Pathology". Inside, there are three columns of tumor types. The first column lists "Secondary Lung Tumors", "Unclassified Lung Tumors", "Mesothelial Tumors", "Epithelial Tumors" (which is highlighted), "Soft Tissue Tumors of the Chest", "Lung Tumor-Like Lesions", and "Miscellaneous lung Tumors". The second column lists "Malignant Epithelial Tumors", "Dysplasia", and "Benign Epithelial Tumors". The third column lists "Small cell lung carcinoma", "Bronchogenic carcinoma [by location]", "Squamous cell lung carcinoma", "Other epithelial tumors", and "Nonsmall cell lung carcinoma". Below these columns is a "Pathology Class" section with six radio button options: "Well Differentiated", "Moderately Differentiated", "Poorly Differentiated", "Well-to-Moderated Differentiated", "Moderate-to-poorly Differentiated", and "Any". At the bottom, there are two sections: "Primary Site" and "Metastatic Distant Site". The "Primary Site" section has a "Lung" sub-section with checkboxes for "Left upper lobe", "Right upper lobe", "Left lower lobe", and "Right lower lobe", and another sub-section with checkboxes for "Mediastinum", "Other", and "Any". The "Metastatic Distant Site" section has checkboxes for "Brain", "Adrenal", "Skin", "Bone", "Lung", "Other", "Liver", "Pleura", "None", and "Any". There are "And" and "Or" radio buttons between the two site sections. At the bottom right are "Cancel" and "Okay" buttons.

Fig 10. Pathology specification window. The screen is customized for lung cancer.

The user interface to the query processing system was written in SmallTalk ObjectWorks (ParcPlace, Irvine, CA).

Query Processing

Once the user has specified the patient demographics, disease pathology, and radiology constraints for the patient case to be retrieved, a PICQUERY + table²⁹ is presented in the main menu so that the user may see a summary of the entire query. The PICQUERY + table consists of five columns including the following: (1) the name of an object attribute (eg, "Lesion Size"), (2) the relational correlator (eg, "greater than"), (3) the attribute value (eg, "3 cm"), (4) the logical operator (eg, AND, OR), which allows construction of compound complex predicates, and (5) the predicate group identification number that is needed for the definition of parenthesized clauses in complex predicates (see ref 29 for further details). If the table summary is acceptable, the user selects the "FIND" button

on the Main Menu to launch the query. The PICQUERY + table is written to a UNIX file and the query processing routine is spawned to locate candidate cases within the index that match the query predicate conditions.

The query processing program is written in the C programming language and uses the Sybase DB-Library C-interface routines to access the UCLA Lexicon. The query processing begins by parsing the input PICQUERY + table predicate by predicate and constructing a structured query language (SQL) command to query the index table populated by the text-analysis step.

In the cases of pathology and radiology finding predicates, the system substitutes the attribute value provided by the user with a unique concept identification (ID) number stored within the lexicon. Using this concept ID number, all synonymous terms, lexical variations, and subclass concepts for the given concept can be identified using the UCLA Thoracic Lexicon.

Radiographic Findings

Lesion Size

☐ Diameter < 3cm

☒ Diameter > 3cm

☐ Any

Uniformity

☐ Homogeneous

☐ Inhomogenous

☒ Any

Density

☐ Soft tissue

☐ Fluid or fluid-like

☐ Fat

☐ Air

☒ Any

Lesion Location

☐ Lung, unspecified

☐ Lung, left lobe

☐ Lung, left upper lobe

☐ Lung, left lower lobe

☐ Main Bronchus

☐ Peripheral

☐ Lung, right lobe

☒ Lung, right upper lobe

☒ Lung, right middle lobe

☐ Lung, right lower lobe

☐ Segmental

☐ Mediastinum, incl. trachea

☐ Any

Lobar

☐ <2cm below carina

☐ >2cm below carina

Calcifications

☐ Non-calcified

☐ Homogeneous

☐ Diffusely speckled

☐ Ring-like, laminated

☐ Central

☐ Popcorn

☐ Eccentric

☒ Any

External Architecture

☐ Smooth

☐ Spiculated

☐ Lobulated

☒ Poorly defined

☐ Well-marginated

☐ Any

Clinical Trend

☒ Growing

☐ Shrinking

☐ Stable

☐ Any

Thoracic Lymph Nodes

☒ Supraclavicular

☐ Upper paratracheal

☐ Lower paratracheal

☒ Aortopulmonary

☐ Anterior (Prevascular)

☐ Subcarinal

☐ Paraesophageal

☐ Tracheobronchial

☐ Peribronchial

☐ Intrapulmonary

☐ Diaphragmatic

☐ Subdiaphragmatic

☐ Axillary

☐ Retrocrural

☐ Any

Site of Invasion

☐ Apices

☐ Chest Wall

☐ Mediastinum

☐ Diaphragm

☐ None, unknown

☒ Any

T-Class

☐ X

☐ 0

☐ IS

☒ 1

☐ 2

☐ 3

☐ 4

☐ Any

N-Class

☐ 0

☒ 1

☐ 2

☐ 3

☐ 4

☐ Any

M-Class

☒ 0

☐ 1

☐ Any

Cancel

OK

Fig 11. Radiology Findings specification window. The screen is customized for patient's with lung cancer who may have masses and involved lymph nodes.

The assignment of concept IDs within the lexicon follows a structured approach based on an outline classification paradigm. This convention is the key to fast location of subclass concepts. As an example, the concept “malignant epithelial tumor” has a concept ID in the pathology section of the lexicon of 01.01.04.00.00. A subclass of this concept, “small cell carcinoma” is assigned the concept ID 01.01.04.01.00. A subclass of small cell carcinoma, “oat-cell carcinoma” is assigned the concept ID 01.01.04.01.01. Using this hierarchical coding convention, the zero-truncated portion of the concept ID defines a node within a concept abstraction hierarchy (eg, 01.01.04 is the truncated ID string for the concept “malignant epithelial tumor”). The “Like” qualifier in the

SQL language is then used to search for all instances of cases in the data base that conceptually match the user defined concept. The result in submitting the partial SQL command “Where HISTOLOGY like ‘01.01.04’,” results in the return of patient ID numbers that have reference to any tumor of histologic class “malignant epithelial tumor.” In this way, predicates that include general values (eg, malignant epithelial tumors) return results that include all children and their children of the parent concept (eg, small cell carcinoma, oat-cell carcinoma, etc).

Presentation of Results

Cases that match the query constraints are presented to the user in the form of a list of

Fig 13. The matched patient case can be reviewed in full from our HIS, RIS, PACS interface.

oped in radiology to describe complex phenomenon. These descriptors are natural for a radiologist to use during the construction of case retrieval queries. One drawback of the NLP approach is that it may produce an overly abundant set of indices for a particular image.

SUMMARY

We present a prototype system that performs automated indexing of radiology and pathology reports from lung cancer patients using natural language processing agents. The advantages of this system include the following: (1) Automatic indexing of free-text reports. No manual indexing is required. (2) The text analysis system can be integrated almost invisibly into the standard routine of a radiology department. This provides a mechanism for maintaining thousands of abstracts for clinical reports and images to be used for research-type retrieval of cases. (3) The system includes a comprehensive lexicon

and explanation system allowing concept-based matching as compared to simple string based match retrievals. (4) The lexicon and text analysis software are expandable given its simple structure. Other domains outside lung cancer can be included. (5) The system is integrated with an HIS, RIS, PACS retrieval interface.

Currently we are working on improving the system in several important areas: (1) Developing a query interface that is more pictorial and interactive, (2) modeling of temporal and evolutionary data objects,²⁵ and (3) including more intelligent frame based methods of text analysis.

ACKNOWLEDGMENT

We thank Profs Wesley W. Chu and Alfonso F. Cardenas of the UCLA Computer Science Department for their contributions to the data modeling and knowledge representation aspects of this project, and John D. Dionisio for his assistance with the VisualWorks SmallTalk training.

REFERENCES

1. Shortliffe EH: Medical informatics and clinical decision making: The science and the pragmatics. *Med Decis Making* 11:S2-S14, 1991 (suppl)
2. Zink S, Jaffe CC: Medical imaging databases: A National Institutes of Health workshop. *Invest Radiol* 28:366-372, 1993
3. Silberschatz A, Stonebraker M, Ullman J: Database systems: Achievements and opportunities. *Comm ACM*, 34:110-120, 1991
4. Chang S-K, Hsu A: Image information systems: Where do we go from here? *IEEE Trans Knowledge Data Eng* 4:431-442, 1992
5. Taira RK, Breant CM, Stepczyk FM, et al: CASE and object-oriented tools for rapid integration of HIS, RIS, and PACS. *Proc SPIE: Medical Imaging PACS: Design and Evaluation* 2165:185-199, 1994
6. Taira RK, Breant CM, Chan HM, et al: Clinical results of HIS, RIS, PACS integration using data integration CASE tools. *Proc SPIE: Medical Imaging PACS: Design and Evaluation: Engineering and Clinical Issues* 2435:406-413, 1995
7. Taira RK, Breant CM, Chan HM, et al: Architectural design and tools to support the transparent access to HIS, RIS, and PACS. *J Digit Imaging* 9:1-10, 1996
8. Sager N, Friedman C, Lyman MS: *Medical Language Processing: Computer Management of Narrative Data*. Reading, MA, Addison-Wesley, 1987
9. Berrut C, Cinquin P: Natural language understanding of medical reports, in Scherrer JR, Cote RA, Mandil SH (eds): *Computerized Natural Medical Language Processing for Knowledge Engineering*. The Netherlands, Elsevier, 1989, pp 129-137
10. Scherrer JR, Cote RA, Mandil SH: *Computerized Natural Medical Language Processing for Knowledge Representation*. Amsterdam, The Netherlands, Elsevier 1989
11. Haug PJ, Ranum DL, Frederick PR: Computerized extraction of coded findings from free-text radiologic reports. *Radiology* 174:543-548, 1990
12. Do Amaral MB, Satomura Y: Structuring medical information into a language-independent database. *Med Inf* 19:269-282, 1994
13. Humphreys BL, Lindberg DAB, Hole WT: Assessing and enhancing the value of the UMLS knowledge sources, in Clayton PD (ed): *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care*. New York, NY, McGraw-Hill, 1991, pp 78-82
14. Humphreys BL, Linberg DA: The UMLS project: Making the conceptual connection between users and the information they need. *Bull Med Libr Assoc* 81:170-177, 1993
15. Lindberg DAB, Humphreys BL, McCray AT: The Unified Medical Language System, in van Bommel JH, McCray AT (eds): *Yearbook of Medical Informatics*. Stuttgart, Germany, Schattauer Verlag, 1993, pp 41-51
16. Friedman C: The UMLS coverage of clinical radiology, in Frisse, ME (ed): *Proceedings of the Annual Symposium Computer Applications in Medical Care*. Baltimore, MD, Nov McGraw-Hill, 1993, pp 309-313
17. Fraser RG, Pare JAP, Pare PD, et al: *Diagnosis of Disease of the Chest*, vols I and II, chapters 1 and 8. Philadelphia, PA, Saunders, 1991, pp 3-314, 1327-1699
18. Lange S: *Radiology of Chest Diseases*. New York, NY, Thieme, 1990
19. Dahnert W: *The Radiology Review Manual* (ed 2). Baltimore, MD, Williams & Wilkins, 1993, pp 257-334
20. Didyk R (ed): *Stedman's Radiology Words*. Baltimore, MD, Williams & Wilkins, 1993
21. Tuddenham WJ: Glossary of Terms for Thoracic Radiology: Recommendations of the Nomenclature Committee of the Fleischner Society. *Am J Roentgenol* 143:509-517, 1984
22. Webb WR, Muller NL, Naidich DP: *High Resolution CT of the Lung*. New York, NY, Raven, 1992, pp 157-159
23. UMLS Knowledge Sources Documentation (4th Experimental Edition). Bethesda, MD, U.S. Department of Health and Human Services, National Institutes of Health, National Library of Medicine, April 1993
24. Wermter S: *Hybrid Connectionist Natural Language Processing*. New York, NY, Chapman & Hall, 1995 pp 34-55
25. Appelt DE, Hobbs JR, Bear J, et al: FASTUS: A finite-state processor for information extraction from real-world text, in Bajcsy R (ed): *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, Chambéry, France, 1993, pp 1172-1178
26. Blank GD: A Finite and Real-time Processor for Natural Language. *Comm ACM* 32:1174-1189, 1989
27. Joseph T, Cardenas AF: PICQUERY: A high level query language for pictorial database management. *IEEE Trans Software Engin SE* 14:630-638, 1988
28. Chu WW, Ieong IT, Taira RK: A semantic modeling approach for image retrieval by content. *The VLDB Journal—Special Issue on Spatial Database Systems*. 3:445-477, 1994
29. Cardenas AF, Ieong IT, Taira RK, et al: The knowledge-based object-oriented PICQUERY+ language. *IEEE Trans Knowled Data Engin* 5:644-657, 1993
30. American College of Radiology: *Index for Radiological Diagnoses* (ed 3). Reston, VA, American College of Radiology, 1986
31. World Health Organization: *Histological Typing of Lung Tumours* (ed 2). *Am J Clin Pathol* 77:123-136, 1982
32. Kofakis P, Karmirantzou A, Davaklis Y, et al: Image archiving by content: An object-oriented approach. *Proc SPIE* 1234:275-286, 1990
33. Tagare HD, Gindi GR, Duncan JS, et al: A geometric indexing scheme for an image library, in Lemke HU, Rhodes ML, Jaffe CC, et al (eds): *Proc CAR*. New York, NY, Springer-Verlag, 1991, pp 513-518
34. Orphanoudakis SC, Chronaki C, Kostomanolakis S: 12C: A system for the indexing, storage, and retrieval of medical images by content. *Med Inf* 19:109-122, 1994