

Indexing Large Archives of Pathology Images Using the Unified Medical Language System (UMLS)

**G. William Moore, M.D., Ph.D., David S. Brenner, M.D.,
Jules J. Berman, Ph.D., M.D.**

Running title: Image Indexing with UMLS

Abstract

The value of any large image archive resides in the ability to select and retrieve images based on features of interest contained in the images. We here show that images can be automatically encoded from descriptive text (image-legends), into concept codes of the Unified Medical Language System (UMLS), a technique that permits powerful image categorization and retrieval, and is generalizable to image archives of any size. A collection of 5,465 pathology image legends was encoded into UMLS terms, via our computer translation program that parses and maps plain-text image-legends into lists of UMLS terms. Each image-legend yielded an average of 15 index-terms, ranging in frequency from five terms in the least-indexed legend to 58 terms in the most-indexed legend. The resulting UMLS index can be used to retrieve images, even when a chosen query term is not included in the image legend.

Key words: index, UMLS, nomenclature, histopathology, retrieval, image, archive, query, database, semantic, vocabulary

Introduction

Without image indexing, large image archives quickly become image cemeteries, collections of inaccessible, buried images. How can images be archived so that they can be searched and retrieved by image features that relate to medical concepts? Neophytes to the field of image-capture may wrongly assume that mne-

monics incorporated in the image filename (e.g., assigning the filename *adrenca.jpg* to the image of an adrenal carcinoma) can serve as adequate reminders of the image content when image retrieval is desired. Others put their faith in off-the-shelf software products that display images in small "thumbnail" versions that can quickly be scanned. Unfortunately, when the images in an archive number in the thousands or millions, the problem of image retrieval becomes very difficult. Indexing pathology images has particular importance, because pathology images contain levels of information beyond a single diagnostic term.

Consider an image of a tuberculous granuloma occurring in the lung. For discussion sake, let us assume that the granuloma is caseous, and that giant cells are present. The thoughtful archivist indexes the image under the title "Pulmonary TB," and assumes that anyone wanting to review the image can find it under this plainly adequate descriptive terminology, which includes both the disease process and the name of the involved organ. Two years later, a pathologist wishes to review all the images of tuberculosis involving the lung. A simple query system tries to match the pathologist's query against the terms contributed by the archivist (such as "Pulmonary TB"). The pathologist enters "granuloma," as the query term, but the system fails to match the image because "granuloma" is not included in the index term "Pulmonary TB". The pathologist enters "caseous," "caseating," "tuberculosis", "tubercle," and "granulomatous": no matches. Frustrated, the pathologist tries to cross-reference against organ, and enters "lung": no match, and "lungs" :no match, since the search engine index only contains "Pulmonary". At this point, the pathologist gives up in frustration.

This selected example illustrates that indexing under a single diagnostic term may eliminate any chance of retrieving the image using alternate terminology. The example also illustrates that indexing the image under a diagnostic term ignores the variety of pathologic concepts (e.g., caseous necrosis, granuloma, giant cells) that are valid attributes of the image.

The ability to retrieve information from a large data collection is one of the most fundamental and important problems facing any archivist. In fact, it can fairly be argued that there is no value in archiving images if they cannot be retrieved. In this report, we demonstrate that pathology images can be indexed using UMLS (Unified Medical Language System), a coded listing of about 1.5 million concept names, that subsumes over forty standardized medical vocabularies. These include HL7 (Health Level 7 Vocabulary), CPT2000 (Physician's Current Procedural Terminology), ICD (International Classification of Diseases), SNOMED (Systematized Nomenclature of Human and Veterinary Medicine) and LOINC (Logical Observation Identifiers, Names and Codes). Extensive information describing UMLS is available at:

<http://www.nlm.nih.gov/>

One of the medical vocabularies contained in UMLS is SNOMED, a coded glossary of over 150,000 terms, used primarily by pathologists to index anatomic pathology reports. We have previously shown that descriptive free text can be automatically computer-translated into SNOMED. We employ those same computer techniques to encode images under all the pathologic concepts represented in the image legend-text. (1-5) We have used a non-copyrighted collection of over 6,000 pathology images, previously published by the Armed Forces Institute of Pathology (AFIP), as a working illustration.

Methods

Images and accompanying legend-texts were extracted from the electronic ver-

sions of AFIP fascicles [www.afip.org]. The AFIP fascicles are a U.S. government work, are not copyrighted, and have unrestricted use. All 6,241 legend-texts from the CD-ROM versions (Electronic Fascicles) of the Tumor Fascicles of the Armed Forces Institute of Pathology (AFIP) were assembled as a single ASCII file. The legend-texts were examined individually, for any indication that the corresponding image had been reprinted from a copyrighted source, and these texts and images were deleted. The remaining 5,465 images were compressed 1:10 as JPEG files, and loaded into The Johns Hopkins Autopsy Resource (JHAR), on March 11, 1997. The image search engine and all of the PERL source code for the search engine, are available at:

<http://www.netautopsy.org/iadbimag.htm>

In 1986, the U.S. National Library of Medicine began the construction of UMLS. The purpose of the UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources. The UMLS project develops machine-readable Knowledge Sources, that can be used by a variety of applications programs, to overcome retrieval problems caused by differences in terminology and the scattering of relevant information across many databases. Investigators worldwide may apply cost-free to become registered users of the UMLS.

This report employs the UMLS Metathesaurus, which is organized by concept or meaning. Each concept has attributes that help to define its meaning, e.g., the semantic types or categories to which it belongs, position in hierarchical contexts from various source vocabularies, and a text definition. The year 2000 Metathesaurus contains 1.5 million concept names, and is compiled from more than 90 source vocabularies. Each concept number (CUI = concept unique identifier) consists of C followed by seven decimal digits. Different concept-names, which are regarded as synonyms by the Metathesaurus, may

have the same CUI.

For example, excluding upper-case-lower-case repeats, there are 40 terms that map to C0017155, the CUI for HYPERTROPHIC GASTRITIS, as follows:

adenopapillomatosis gastrica
 chronic hypertrophic gastritis
 disease, menetrier
 disease, menetrier's
 gastric hyperplasia
 gastric mucosal hyperplasia
 gastric mucosal hypertrophy
 gastritides, hypertrophic
 gastritis giant hypertrophic
 gastritis hypertroph gigantica
 gastritis hypertrophic
 gastritis hypertrophic gigantica
 gastritis hypertrophica gigant
 gastritis hypertrophica gigantea
 gastritis hypertrophica gigantica
 gastritis, giant hypertrophic
 gastritis, hypertrophic
 gastropathy mucous cell type
 giant hypertrophic gastritis
 giant rugal gastritis
 giant rugal hypertrophy
 giant rugal hypertrophy of stomach
 giant rugal hypertrophy stom
 hyperplastic gastropathy
 hyperplastic gastropathy of mucous cell
 type
 hypertrophic bulbous gastritis
 hypertrophic gastritides
 hypertrophic gastritis
 hypertrophic gastropathy
 hypertrophic prolif gastritis
 hypertrophic proliferative gastritis
 hypertrophy, gastric mucosa
 hypertrophy, giant rugal
 massive hypertrophic gastritis
 menetrier disease
 menetrier's disease
 polypoid swell gastric mucosa
 polypoid swelling of gastric mucous mem-
 brane
 prolif chr hypertr gastritis
 proliferative chronic hypertrophic gastritis

In addition to text-terms contained in the official UMLS database, we assigned additional synonyms to our UMLS database. Such additional terms in our database

serve to recognize legend-text or database queries which might contain these alternate terms. Compound terms that contained subconcepts were indexed redundantly. Thus, CELLULAR BLUE NEVUS was redundantly indexed as: CELLULAR BLUE NEVUS (C0334448), BLUE NEVUS (C0206736), CELL (C0007634), BLUE (C0332584), and NEVUS (C0027960).

In order to match UMLS concepts to corresponding word sequences in the AFIP image-legend file, we analyzed the image-legend file by the BARRIER WORD METHOD (2). In the barrier word method, natural-language medical text is regarded as a sequence of medical concept phrases linked together with grammatical objects. The grammatical objects, or BARRIER WORDS, consist of numerals, punctuation, single letters, articles, prepositions, and common verbs and modifiers. For example, consider the following image legend. The software algorithm parses through the text, extracting the medical concept phrases that lie between the barrier words. For example, consider the following image legend:

LENTIGINOUS COMPOUND
 NEVUS this LESION is an EARLY
 COMPOUND NEVUS, because a
 NEST has MIGRATED from the
 EPIDERMIS into the DERMIS
 (LOWER RIGHT of c) elsewhere,
 the HISTOLOGY is that of a
 SIMPLE LENTIGO.

In this example, the barrier words are displayed in lower case and the parsed medical concept phrases are displayed in UPPER CASE. In general, a single medical concept phrase in a legend-text, or a sequence of medical concept phrases uninterrupted by barrier words, should point to one or more UMLS concepts, as shown in Table I.

Finally, there are ambiguities in UMLS, in which the same or similar English words have a different meaning, depending upon context. For example, IRIS (C0022077) as part of the eye has a different meaning from IRIS (C0331686) as a flower.

In the AFIP image-legend file, one can confidently predict that all occurrences of the word IRIS intend the Topography-meaning, not the Living-Organism-meaning, so that IRIS (C0331686) as a flower is retired. Despite these efforts, there are legend-text terms that will be incorrectly translated unless an appropriate context is present in the legend-text, such as the word ADNEXA without a nearby word that implies either SKIN ADNEXA (C0221943), UTERINE ADNEXA (C0001575), or OCULAR ADNEXA (C0229243). In this case, the indexing system incorporates ALL THREE meanings, so that a certain number of uterine and ocular adnexa will be indexed and retrieved when the user actually desires skin adnexa, etc..

A survey of 100 consecutive image-legends was examined manually, in order to determine the efficacy of the computerized indexing system. For each image-legend, the terms that were captured by the translation program were compared to the free-text description. A medically significant term, present in the image-legend but not captured by the program, was considered as a FALSE NEGATIVE. Conversely, a term captured by the program but not present in the text was considered as a FALSE POSITIVE. By design, the program captured no false positives. The false negative rate, FNR, is calculated by the formula: $FNR = FN / (TP + FN)$, where FN is the number of false negatives and TP is the number of true positives.

In the JHAR website, an image can be retrieved by any combination of the terms included in the UMLS index for the image. In addition, the image can be retrieved by any term that maps to any of the terms included in the UMLS index for the image (i.e., synonyms will match). In addition, the UMLS vocabulary was enhanced by the authors' creation of an override list of terms that includes added terms, and also imposes a contextual priority over the UMLS codes.

Results

Using this synonym dictionary from ordi-

nary English terms into UMLS, each of the 5,465 separate legend-texts was pointed to its corresponding UMLS-codes. Among 5,465 image-legends, there was a total of 85,253 index-terms captured by the system. The image-legends yielded an average of $15.6 = 85,253 / 5,465$ index-terms per legend, standard deviation 5.93, ranging in frequency from five terms in the least-indexed legend to 58 terms in the most-indexed legend.

In a manual quality assurance review of 100 consecutive image-legends, a total of 2,950 index-terms were captured by the system in the 100 image-legends, and there were 84 terms that should have been indexed by the system, but were not, for a false-negative rate of $2.8\% = 84 / (2950 + 84)$. These false-negatives consisted predominantly of synonyms used in the image-legends, which had not yet been incorporated into the program's synonym list. There were no false positives.

A listing of every concept-number used in the image database and its frequency in the database is instructive. The concept PAPILLARY (C0205312) occurs 166 times. This would include any image in which the word papillary appears in its legend-text, whether it refers to thyroid neoplasm, breast neoplasm, urinary tract neoplasm, or any image associated with the concept PAPILLARY. The concept IRREGULAR (C0205271) occurs 136 times. Again, the word IRREGULAR is a general concept, and may refer to a nuclear feature, tumor boundary, cellular distribution, etc.

The fifty most frequent UMLS Concept-Names found in the AFIP legend texts are as follows: MEDICINE, PATHOLOGY, EDUCATION, PHOTOGRAPH, CELL, PICTURE, NEOPLASM, LUNG, DISEASE, PATIENT, FOCUS, SMALL, LARGE, APPEARANCE, BREAST, SEE, NEGATIVE, STAIN, PRESENT, ARISE FROM, EYE, SAME, PATTERN, CYTOPLASM, SOME, INFILTRATE, KIDNEY, PROMINENT, ATYPIA, LEFT, THYROID GLAND, UTERUS, RIGHT, CANCER, MANY, SPECIMEN, TYPE, BONE,

SMEAR, BONE MARROW, GIEMSA STAIN, HEMATOXYLIN, AFFECT, CERVIX, TUMOR CELL, EOSIN, ELEVATE, VESSEL, LYMPHOCYTE, OLD. These most frequent codes represent repeating descriptive themes found in many legend-texts, and tend not to be very specific as indexing terms. In general, modifiers occur at higher frequency than diagnoses. More specific terms are encountered with decreasing frequency. For example, there are nine images associated with the concept, MALIGNANT MENINGIOMA (C0259785) and with the concept, ROSAI DORFMAN DISEASE (C0019625). There are eight images associated with DERMOID CYST (C0011649), and eight with CHONDROBLASTOMA (C0008441).

Discussion

Before the advent of practical forms of electronic image storage, pathology images were stored as kodachromes, or as glossy prints, or could be found in published pathology atlases. Paper atlases tend to have a relatively small number of images (hundreds), because of space and cost limitations. Currently, there are numerous collections of pathology images archived on the Internet, ranging in size from dozens of images to thousands of images. Although small-scale efforts have been described (6,7), to our knowledge, there is no example of a public pathology archive site that retrieves large numbers (greater than 100,000) of well-coded histopathology images. This is particularly unfortunate when one considers that every pathology image on the Internet, regardless of its server location, could be immediately accessible from a single site if the complete URL (web address or Uniform Resource Locator) of the image were known.

When image datasets are coded using a standard vocabulary, such as UMLS, students and researchers receive the obvious advantage of being able to easily retrieve images based on semantically equivalent concept look-ups. Another important attribute of encoded pathology-related data-

sets is the ability to 'bin' entities so that the frequencies of occurrence of every UMLS concept can be determined and compared with the frequencies of occurrence of all other entities in the dataset. The global analysis of indexed data (as opposed to merely retrieving a particular data element from a dataset) is a feature of coded datasets that is not provided by natural language search algorithms. Indexing software algorithms that collect all natural words or phrases without assigning like terms to common codes cannot 'bin' concepts and therefore do not support meaningful enumeration of pathologic entities. Researchers now envision a growing role of UMLS-encoded medical text for retrieving and evaluating medical textual content (5,8,9).

This study discusses encoding algorithms for a relatively small number of pathology cases (over 5,000). It is important to appreciate that the algorithms discussed are computationally efficient, extensible to other related projects in which medical free-text is mapped to standard nomenclature, and scalable to include any number of reports. In a related study recently undertaken by two of the authors [GWM and JJB], a large textual dataset of hundreds of thousands of pathology reports and pathology text was parsed to extract all of the contained medical concept phrases. Approximately 418,000 phrases were matched against the entire UMLS concept file (about 1.5 million concepts), an algorithm that required iterative comparisons between each of the 418,000 parsed text phrases against each of the 1.5 million UMLS phrases. The entire algorithm took about 90 seconds to complete, providing a completely coded dataset that covered 16 years of pathology reports collected from a major medical institution. The same procedure, if humanly possible, may well have usurped years of a pathologist's time.

This study shows: (1) That an image retrieval system utilizing all the pathology concepts contained within the image is achievable; (2) That images can be automatically UMLS encoded from pre-exist-

TABLE I
UMLS TRANSLATION OF SAMPLE LEGEND-TEXT

LEGEND NAME	UMLS CODE	UMLS NAME
LENTIGINOUS	C0023321	Lentigo
COMPOUND NEVUS	C0259781	Compound Nevus
LESION	C0012634	Lesion
EARLY	C0205085	Early
NEST	C0205234	Focal
MIGRATED	C0232902	Migration
EPIDERMIS	C0014520	Epidermis
DERMIS	C0011646	Dermis
LOWER	C0205104	Inferior
RIGHT	C0205090	Right
HISTOLOGY	C0019638	Histologic
SIMPLE LENTIGO	C0302255	Lentigo Simplex

ing text that describes the images; (3) That automatic UMLS encoding can utilize the entire UMLS nomenclature. Since coding is automatic, the process can be extended to any number of archived images. Readers are encouraged to visit the Johns Hopkins Autopsy Resource, and use the image archive.

References

1. Berman, J.J., Moore, G.W. SNOMED-encoded surgical pathology databases: A tool for epidemiologic investigation. *Mod Pathol.* 9:944-950, 1996.
2. Moore, G.W., Berman, J.J. Automatic SNOMED coding. *Proc Annu Symp Comput Appl Med Care.* 18:225-229. 1994.
3. Moore, G.W., Berman, J.J. Performance analysis of manual and automated systematized nomenclature of medicine (SNOMED) coding. *Am J Clin Pathol.* 101:253-256, 1994.
4. Berman, J.J., Moore, G.W., Donnelly, W.H., Massey, J.K., Craig, B. A SNOMED analysis of three years accessioned cases (40,124) of a surgical pathology department: implications for pathology-based demographic studies. *Proc Annu Symp Comput Appl Med Care.* 1994;18:188-192, 1994.
5. Moore, G.W., Berman, J.J. Anatomic Pathology Data Mining. In: Cios KJ, ed, *Medical Data Mining and Knowledge Discovery.* Springer-Verlag, Berlin/Heidelberg, 2000
6. Eysenbach, G., Bauer, J., Sager, A., Bittorf, A., Simon, M., Diepgen, T. An international dermatological image atlas on the WWW: practical use for undergraduate and continuing medical education, patient education and epidemiological research. *Medinfo 9 Pt 2:*788-92, 1998.
7. Jaulent M., Le Bozec C., Cao Y., Zapletal E., Degoulet P. A property concept frame representation for flexible image content retrieval in histopathology databases. *Proc AMIA Symp.* 20(Suppl):379-83, 2000.
8. Lowe, H.J., Antipov, I., Hersh, W., Smith, C.A. Towards knowledge-based retrieval of medical images. The role of semantic indexing, image content representation, and knowledge-based retrieval. *Proc AMIA Symp* 882-886, 1998.
9. Goldberg HS, Hsu C, Law V, Safran C Validation of clinical problems using a UMLS-based semantic parser. *Proc AMIA Symp* 805-809, 1998.

G. William Moore, MD, PhD. [1,2,3] , David S. Brenner, MD. [2] , Jules J. Berman, PhD, MD. [1,2,3,4]. Pathology and Laboratory Medicine Service, Veterans Affairs Maryland Health Care System, Baltimore, Maryland [1]; Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland [2]; Department of Pathology, The Johns Hopkins Medical Institutions, Baltimore, Maryland [3]; and Resources Development Branch, Cancer Diagnosis Program, National Cancer Institute, National Institutes of Health, Bethesda, Maryland [4].

Address reprint inquiries to: Dr. Jules J. Berman email: bermanj@mail.nih.gov