Informatics in Radiology

RADTF: A Semantic Search–enabled, Natural Language Processor–generated Radiology Teaching File

Bao H. Do, MD • Andrew Wu, MD • Sandip Biswal, MD • Aya Kamaya, MD • Daniel L. Rubin, MD

Storing and retrieving radiology cases is an important activity for education and clinical research, but this process can be time-consuming. In the process of structuring reports and images into organized teaching files, incidental pathologic conditions not pertinent to the primary teaching point can be omitted, as when a user saves images of an aortic dissection case but disregards the incidental osteoid osteoma. An alternate strategy for identifying teaching cases is text search of reports in radiology information systems (RIS), but retrieved reports are unstructured, teaching-related content is not highlighted, and patient identifying information is not removed. Furthermore, searching unstructured reports requires sophisticated retrieval methods to achieve useful results. An open-source, RadLex®–compatible teaching file solution called RADTF, which uses natural language processing (NLP) methods to process radiology reports, was developed to create a searchable teaching resource from the RIS and the picture archiving and communication system (PACS). The NLP system extracts and de-identifies teaching-relevant statements from full reports to generate a stand-alone database, thus converting existing RIS archives into an on-demand source of teaching material. Using RADTF, the authors generated a semantic search–enabled, Web-based radiology archive containing over 700,000 cases with millions of images. RADTF combines a compact representation of the teaching-relevant content in radiology reports and a versatile search engine with the scale of the entire RIS-PACS collection of case material.

©RSNA, 2010 • radiographics.rsna.org

Abbreviations: ACL = anterior cruciate ligament, DICOM = Digital Imaging and Communications in Medicine, IT = information technology, NLP = natural language processing, PACS = picture archiving and communication system, RIS = radiology information system

RadioGraphics 2010; 30:2039–2048 • Published online 10.1148/rg.307105083 • Content Codes: ED IN

1From the Department of Radiology, Stanford University Hospitals and Clinics, 300 Pasteur Dr, Room H1307, Stanford, CA 94305 (B.H.D., S.B., A.K., D.L.R.); and Department of Radiology, Kaiser Permanente Downey Medical Center, Downey, Calif (A.W.). Presented as an education exhibit at the 2009 RSNA Annual Meeting. Received April 6, 2010; revision requested June 2 and received July 7; accepted July 12. S.B. is a co-founder of SiteOne Therapeutics and a member of the GE Healthcare Tiger Team; D.L.R. receives research support from GE Healthcare; all other authors have no financial relationships to disclose. Address correspondence to B.H.D. (e-mail: bando@stanford.edu).
Introduction

Archiving and searching radiology reports for findings and diagnoses of interest is a crucial activity for education and clinical research, but manually identifying teaching cases and recording the pathologic conditions they illustrate are time-consuming tasks. Furthermore, creation of organized, structured teaching files may omit imaging and report data that could be useful in other teaching scenarios. Accidental pathologic conditions not pertinent in demonstrating the primary teaching point may be discarded (eg, when a user indexes a case of aortic dissection but not images of an incidental osteoid osteoma). An alternate strategy for identifying teaching cases is text search of reports in radiology information systems (RIS); such a search makes the reports and their images in picture archiving and communication systems (PACS) accessible. Several such systems have been developed to date, including MoSearch (1), Render (2), and Illuminate (Softek, Kansas City, Kansas). The collection of radiology reports and images in existing RIS and PACS can be a useful resource for teaching material.

However, studies retrieved are the full reports and images and are not as compact as manually indexed teaching files. They contain information extraneous to teaching—such as normal findings, descriptions of technique (section thickness, postprocessing steps, contrast material dose, and timing), and patient identifying information—which is not appropriate to present when this information is accessed for teaching. Furthermore, text search of unstructured reports by using traditional keyword matching techniques may retrieve irrelevant results due to the nature of radiology reports, which often contain negative, “hedge,” and uncertainty or conditional statements that lead to poor search precision by automated indexing systems (3,4). For example, if reports containing the term appendicitis are retrieved by using keyword matching, reports containing the following phrases would be retrieved: “consistent with appendicitis,” “no evidence of appendicitis,” and “cannot exclude appendicitis.” However, only the first match is relevant.

We believe that reports and images in RIS and PACS can provide a resource for teaching material, provided that indexing and search engines can (a) extract the relevant content that could be useful for teaching, (b) identify the context of findings and diagnoses by filtering negations and hedges, and (c) remove patient identifying information. Therefore, we have developed an open-source teaching file solution that is compatible with RadLex® (5) and that uses natural language processing (NLP) methods to extract, de-identify, store, and search teaching-relevant material within unstructured radiology reports to create a teaching file resource from existing RIS and PACS. NLP methods can be applied to unstructured radiology reports to remove extraneous information and to identify key content useful for teaching.

The generated database is separate from the RIS and PACS, but the extensible design can launch Web-based and full PACS viewers for viewing studies in Digital Imaging and Communications in Medicine (DICOM) quality without the need for a separate image repository or export step. Both single query (eg, appendicitis) and multiple-term query (eg, fracture of the scaphoid, infarct of the basal ganglia, rupture of the anterior cruciate ligament [ACL]) searches can be performed. A custom-designed, previously validated NLP system ranks results by identifying contextual cues to eliminate negation and hedging (4).

Our teaching file solution combines a compact representation of the teaching-relevant content in reports while maintaining the scale of the entire RIS-PACS collection of case material. In addition, use of actual PACS studies as teaching material may confer the advantage of demonstrating complex pathologic conditions best appreciated only with cross-sectional imaging (eg, sigmoid volvulus or bucket-handle meniscal tear). In this article, we discuss the architecture, basic features, technical details, security, and unique features of our system, which is called RADTF and is provided freely as open-source (Open Source Initiative, San Francisco, Calif) (http://www.opensource.org). Individuals and institutions may request code directly from the authors and may modify under the GNU General Public License.
An overview of the RADTF system is shown in Figure 1. RADTF comprises three core functional components: (a) the importer, (b) the Web server and database, and (c) the search engine and viewer.

RADTF requires a Web- and PHP-configured MySQL database server (Sun Microsystems, Santa Clara, Calif). Our implementation uses the Apache Web server (Apache Software Foundation), which can be hosted on multiple platforms including Windows (Microsoft, Redmond, Wash), Mac Os X (Apple, Cupertino, Calif), and Linux. Implementing code was written by using the Hypertext Preprocessor (PHP) scripting language (PHP Group). Hardware requirements are minimal; RADTF has been configured on a commercial “netbook” with the following technical specifications: Atom 1.6-GHz processor (Intel, Santa Clara, Calif), 16-GB hard drive, 1-GB random-access memory (RAM), Windows XP operating system.

The data importer and search engine components of RADTF employ NLP. NLP methods are computer algorithms that automate data extraction from unstructured text. The data importer automatically extracts and de-identifies relevant statements from the full, unstructured source reports. The semantic search engine retrieves cases from the database and contextually ranks results on the basis of recognized negative and hedge expressions at the statement level. Detailed descriptions of each component are provided in the Technical Details section.
Basic Features

User Interface
Users perform queries via a Web browser (Fig 2). Searches can be filtered by age range, sex, modality (magnetic resonance [MR] imaging, computed tomography [CT], ultrasonography, interventional procedures, plain radiography, fluoroscopy, mammography, and nuclear medicine), and search strategy (semantic or exact string). Users can perform a search using “exact string” or exact terms matching (eg, “gallbladder perforation”). However, by default RADTF performs semantic searches, or searches based on concepts in which the order of the search terms does not matter. For example, a search for “perforated gallbladder” will match results for both “gallbladder is perforated” and “a perforated gallbladder is noted” as well as “perforatedgallbladder” (example of a transcription punctuation error). Semantic search is discussed further in the Technical Details section.

Search Results
RADTF returns the results of a query as a list of cases (Fig 3). Each result includes an excerpt containing the highlighted query terms or concepts and the patient’s age and sex. Results are ranked by relevancy on the basis of recognized negative and hedge expressions. Reports that do not contain a negative or conditional expression (eg, “consistent with ACL tear”) are ranked highest. Results include not only statements that contain the exact query terms string “ACL tear” but also any statements that contain the concepts “ACL” and “tear,” allowing more user flexibility in query generation. The typical query time length varies depending on the size of the database and complexity of the query term, from approximately 0.8 second (“acl tear,” MR imaging database with 222,712 sentences) to 3 seconds (“pneumothorax,” radiography database with 1,475,124 sentences).

Viewing Cases
Users can view studies in full DICOM quality. Each search result is linked to the original PACS study via a unique study accession number, which is used to launch the specific study in compatible Web-based or full DICOM viewers (Fig 4). By linking to original PACS studies, users can review all the studies for the selected patient such as follow-up studies that may demonstrate disease evolution or complication (eg, small bowel perforation in the setting of treated ileal lymphoma), thus providing additional teaching points. Users can annotate studies with arrows and text, mark specific images as “key” or index images to highlight disease, and save presentation...
Figure 4. Viewing cases. Results are linked to the original study by means of a unique accession number. Users can review cases in compatible Web-based DICOM viewers or directly in PACS (not shown).

Figure 5. DDX mode. DDX mode enables users to conduct multiple queries simultaneously for comparing studies. For example, a user entertaining the diagnosis of hypersensitivity pneumonitis can review types of interstitial lung diseases by entering a list of differential considerations in a text box. NLP is employed to retrieve the desired cases. The NLP algorithm eliminates matches that mention the query in a negated or hedged context. NSIP = nonspecific interstitial pneumonia, UIP = usual interstitial pneumonia.

Differential Diagnosis Mode
Differential diagnosis (DDX) mode enables users to conduct multiple queries simultaneously for comparison, a task potentially useful for decision support. For example, a user entertaining the diagnosis of hypersensitivity pneumonitis may want to review other types of interstitial lung diseases.
Quiz mode. Radiologists can create a PACS-based quiz or self-learning module by entering a list of diseases in a text box. For example, if the user enters a list of liver lesions, quiz mode performs a batch search and presents the search results as a list of unknown cases by randomizing the order of the query list and suppressing the diagnosis, displaying only the patient age and sex and the link to the PACS images. Users can view the study in PACS, then refer to the actual report for the findings and “answer.”

Manually Authoring Cases
In addition to storing teaching material automatically generated by the NLP, RADTF also allows users to manually author cases directly within the PACS workstation. However, this feature is PACS vendor–dependent and is enabled only with PACS clients that can launch Web browsers and pass relevant study and patient data using a URL GET method. An auto-suggest function is implemented and suggests possible disease terms from RadLex in real time as the user enters the disease entity (Fig 6). Manually authored cases are stored in a separate database from the auto-generated database, although the “report” (or diagnosis and discussion, if pertinent) undergoes the same NLP query analysis. Accession numbers
are automatically populated from the PACS to the Web-based manual author form by using a hypertext transfer protocol (HTTP) URL GET method; the stored accession numbers allow users to launch studies in the same manner as used in the auto-generated database.

**Technical Details**

This section discusses technical details of the following RADTF components: (a) importer, (b) Web server and database, and (c) search engine (Fig 1).

**Importer**

The RADTF importer is a PHP script that processes full radiology reports (the reports can be a flat text file containing an aggregate of radiology reports, which can be generated by the information technology [IT] section of the radiology department). Importing is a semiautomated, one-time process performed at installation (and can be run subsequently to update the generated database to add new reports). Processing time varies depending on the complexity of the report, but generally ranges between 15 and 30 minutes for 50,000–100,000 reports. NLP is performed both at importation and during searches. The importer extracts teaching file–relevant statements from radiology reports using the following NLP steps, or a heuristic approach:

1. Since statements expressing teaching-relevant findings and interpretations are usually contained in the “impressions” section of a radiology report, the importer first identifies and parses the impressions section.
2. Individual statements (sentences or clauses) are extracted by using punctuation cues (specifically semicolons and periods, which are distinguished from decimals by applying a regular expression filter for digits).
3. Nonclinical statements (eg, “I have personally read the above report” or “CT portion of PET/CT exam was performed solely for anatomic localization”) are eliminated. This step is performed by using regular expressions based on a lexicon of standard phrases derived from reviewing sample reports.
4. De-identification is performed to eliminate statements containing patient medical record numbers and patient or physician names. Patient and physician names are eliminated (a) directly by using string matching (eliminating sentences containing the patient’s and radiologist’s names) and (b) indirectly by using contextual cues (eliminating sentences containing a communications or identification concept and using regular expressions such as “findings were discussed with Dr Physician” and “Name: Mr Patient”).

After processing each report, the importer outputs each statement to the Web server and MySQL database (Fig 1).

**Web Server and Database**

The MySQL database structure is as follows: “gender,” “age” at time of study, study “accession number,” and all terms in the “extracted statement” are stored. The birth date, patient name, medical record number, and physician name(s) are not stored. Because the database is separate from the RIS, it requires interval data importing for updates. Furthermore, since the database stores only the accession number, which is used to retrieve the original study in the PACS, there is no need for a separate image repository or exporting step.

**Search Engine**

The RADTF custom NLP performs the following four steps for each query: (a) extraction of RadLex anatomy concept terms; (b) stemming; (c) ranking of results on the basis of detected negation, hedge, or uncertainty expressions; and (d) restricting semantic scope to the sentence level.

**Extraction of RadLex Anatomy Concept Terms.**—RADTF automatically extracts RadLex anatomy concept terms (RadLex MySQL version 2.0) from user queries. This feature enables harmonization with other systems that use RadLex for indexing content. For the query “tear of the ACL,” RADTF recognizes that “ACL” is a RadLex-defined anatomic structure (Fig 3). The extracted RadLex ID can be used to search other RadLex-compatible RIS such as a RadLex-based anatomy atlas (6).

RADTF uses the RadLex lexicon to extract semantic anatomic data from user queries; extracted RadLex IDs can then be used to facilitate data exchange with other standardized, RadLex-aware systems.
Stemming.—To increase the recall for searches, RADTF performs stemming of user queries. Stemming is the process of reducing words to their “stem” or common root. For example, both “pneumonitis” and “pneumonia” share the stem “pneumon.” The stemmed form of the user query is then used for searching the RADTF database. Without stemming, the search phrase “acute appendicitis” will match only results containing the exact terms, such as “consistent with acute appendicitis.”

RADTF stemming can increase the number of potential matches. For example, the user query “acute appendicitis” is conflated or reduced to the roots “acute” and “appendi.” Use of these roots retrieves the following statements: “consistent with acute appendicitis,” “acute inflammation of the appendix,” and “acute appendiceal inflammation.” Without stemming, only the first report is retrieved, since it is the only report that contains the original exact query terms “acute” and “appendicitis.”

RADTF’s stemming algorithm is based on Richard Heyes and Jon Abernathy’s PHP implementation of the Martin Porter stemming algorithm (7–9).

Relevance Ranking.—RADTF ranks results on the basis of detected negative and hedge expressions. The RADTF search engine is called a semantic search engine because it searches for terms that are described in a positive context (not negated and not qualified by hedging). A set of heuristic NLP methods is applied to each statement to determine whether it is a positive statement. Results are stratified into relevancy tiers, from highest to lowest relevance, as follows: (a) reports that do not contain a negative or hedge expression, (b) reports that contain a hedge, and (c) reports that contain a negative expression (Fig 3).

Semantic Scope.—In addition to the processing steps already mentioned, RADTF restricts matches to individual statements (as opposed to the entire report) to improve precision. For example, a search for “thickened gallbladder” will match the statement “the gallbladder is inflamed and thickened.” However, it will not match a report that contains the following two statements: “The appendix is thickened. The gallbladder is normal.” This feature improves the precision of searches that contain multiple terms or concepts and enables more granular searches (eg, “perforated acute appendicitis” instead of “acute appendicitis”). Semantic scope restriction also complements stemming.

Security
RADTF is designed to minimize the amount of patient identifying information stored and shown to users. For clarification, RADTF de-identifies patient identifying information from reports (eg, removal of name and medical record number). However, since the accession number used to link to the original study is stored, RADTF does not anonymize or completely dissociate the patient information from the study.

Additional layers of security are as follows: (a) the generated output database is independent from the PACS and RIS; (b) access is restricted to users within the hospital intranet; (c) all user searches and activity are logged by date, time, and Internet protocol address; (d) the viewing of studies and handling of images require a separate, PACS-handled authentication step in which permission is granted and activity is tracked independently by the IT section of the radiology department; (e) RADTF does not link to the patient electronic medical record; and (f) the number of search results listed is restricted to limit use of the system for research.

Unique Features
Storing and retrieving radiology cases is an important activity for education and clinical research, but this process can be time-consuming. Perhaps more important, designing systems to integrate with the clinical work flow across different vendors is costly. RADTF represents an NLP-based method to automatically generate a database of teaching material from existing RIS reports while allowing users to view DICOM-quality images without requiring an exporting process or maintaining a separate, costly, data-intensive image database. Using RADTF, we automatically generated a searchable Web-based radi-
ology teaching file containing over 700,000 cases with millions of images for education. Radiologists at our institution have conducted over 4000 searches during the past year. Since all software components are free and open-source (PHP- and MySQL-configured Apache Web server; RADTF code available under the GNU General Public License on request) and hardware requirements are minimal, RADTF can be deployed inexpensively. Furthermore, the Web-based architecture is extensible and facilitates interoperability with different PACS vendors.

Although the RIS-PACS is a rich data mine for teaching and research material, manually assessing the quality of teaching files (eg, determining if a case of appendicitis is “classic”) is still superior. Therefore, RADTF enables an automated, large-scale compilation method but also provides a mechanism for manual authorship. Several excellent examples of manually authored teaching files can be found on the Internet, including the Medical Imaging Resource Center (10–12). Differentiating features of RADTF manual authoring include a real-time auto-suggestion feature that incorporates RadLex disease and anatomy terms to promote standards-based indexing and the ability to view DICOM-quality images without the need for exporting images or maintaining a separate image repository.

Other systems that enable searching of radiology reports have been described, particularly Render, which is an excellent application developed at Massachusetts General Hospital for searching radiology reports (2). There are several unique features of RADTF that differentiate it from previous systems. RADTF (a) does not store full radiology reports but rather de-identified impressions extracted from radiology reports by using simple NLP rules; (b) is a stand-alone database apart from the RIS and PACS; (c) performs query analysis by using RadLex anatomy ID extraction, which enables its resources to be linked to other RadLex-indexed data, and by using stemming to improve recall of user queries; (d) incorporates semantic search to rank negative and hedge findings or diagnoses at the sentence level; and (e) includes education-specific functions such as the quiz and DDX modes. RADTF is open-source, and its extensible Web-based framework facilitates integration with other Web applications or services.

Retrospective reviews based on keyword methods to identify initial cohorts, such as patients with imaging evidence of intussusception, have been accepted in the radiology literature (13–15). The semantic search used in RADTF improves precision by eliminating negative and hedge phrases, thus providing an efficient and potentially useful approach for radiologists to identify initial cohorts for research without relying on radiology IT systems programmers.

Disadvantages of RADTF compared with existing Web search engines such as Google, Yotlook, and GoldMiner (9,10) include lack of query analysis, including synonymy, query expansion, and spell check functionality. Furthermore, integration with other clinical data such as pathology or surgery reports or procedure notes (eg, colonoscopy, laparoscopy, arthroscopy, and even molecular data) are additional potential features that may enhance educational and research value. Teaching systems based on NLP methods can enable radiologists to tap into the educational value of their large repositories of radiology reports; such systems can be augmented with other data such as pathology and surgery reports.

Conclusions

RADTF is an open-source, RadLex-compatible teaching file solution that uses NLP methods to process radiology reports and thus automatically create a searchable teaching resource from the RIS and PACS. Some of the key features are de-identification of data from radiology reports and the use of NLP to extract and de-identify teaching-relevant information and improve the specificity of the search. Using RADTF, we generated a searchable, Web-based radiology teaching file containing over 700,000 cases in less than 1 day. With extensions to our methods, integration with other clinical data such as pathology and even molecular data could enhance the educational and research value of this resource in the future.

Acknowledgment.—We thank Vijay Yeluri for technical assistance with PACS compatibility.
References


The collection of radiology reports and images in existing RIS and PACS can be a useful resource for teaching material.

NLP methods can be applied to unstructured radiology reports to remove extraneous information and to identify key content useful for teaching.

RADTF uses the RadLex lexicon to extract semantic anatomic data from user queries; extracted RadLex IDs can then be used to facilitate data exchange with other standardized, RadLex-aware systems.

Teaching systems based on NLP methods can enable radiologists to tap into the educational value of their large repositories of radiology reports; such systems can be augmented with other data such as pathology and surgery reports.

RADTF is an open-source, RadLex-compatible teaching file solution that uses NLP methods to process radiology reports and thus automatically create a searchable teaching resource from the RIS and PACS.