ISSN:0975-9646

Novel Implementation of Text Mining for Reports

K.Suresh,K.Venkatesh Sharma,Y.Srinivas,K.Shalini
Sri indu college of engineering and technology Hyderabad, A.P. India
kanigirisuresh@gmail.com
venkatesh_k123@rediffmail.com
sri_vass999@yahoo.com
shalini8881@gmail.com

***Abstract*—In this paper, we propose a text mining system to extract and use the information in radiology reports. The system consists of three main modules: medical finding extractor, report and image retriever. The medical finding extraction module automatically extracts medical findings and associated modifiers to structure radiology reports. The structuring of the free text reports bridges the gap between users and report database, makes the information contained in the reports readily accessible. It also serves as intermediate result to other components of the system. The retrieval module analyzes user's query and returns the reports and images that match the query. The overall evaluation results are satisfactory, though more thorough testing and evaluation are needed. Our future work includes improving the current system performance and implementing the radiology report generation system using statistical machine translation approach, for which we have designed the general architecture**.

Keywords-component; Text Mining, Radiology reports

I. INTRODUCTION

Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers generally to the process of deriving high-quality information from text. High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering[2], concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). With the advances in medical technology and wider adoption of electronic medical record systems, large amounts of medical text data are produced in hospitals and other health institutions daily. These medical texts include the patient's medical history, medical encounters, orders, progress notes, test results, etc. Although these text data contain valuable information, most are just filed and not referred to again. These are valuable data that are not used to full advantage.[1]

In radiology reports are in free text format and usually unprocessed, there is a great barrier between the radiology reports and the medical professionals (radiologists, physicians, and researchers), making it difficult for them to retrieve and use useful information and knowledge from the reports. As the information is not accessible, it cannot be used for other related applications. Therefore, to provide the needed information to the medical professionals and make use of the information, text mining in the radiology reports provides a solution to the problem.

II. ANALYSIS

A. PURPOSE OF THE SYSTEM

Radiology reports contain rich information describing radiologist's observations on the patient's medical conditions in the associated medical images. However, as most reports are in free text format, the valuable information contained in those reports cannot be easily accessed and used, unless proper text mining has been applied. In this project, we propose a text mining system to extract and use the information in radiology reports.

The system consists of three main modules:

a) Medical finding extractor,

b) Report and image retriever, and

c) text-assisted image feature extractor.

B. Maintaining the Integrity of the Specifications

In Existing System Dominich built a web-based neuroradiological information retrieval system (NeuRadIR). They indexed the radiology reports and allow users to retrieve the medical records by three modes: Boolean, hyperbolic, and interaction. However, as the radiology reports in their databases were originally Hungarian, the English version of which seemed to be simplified. User' queries are also limited as the controlled vocabulary used in indexing does not have enough coverage for the domain.

Content based image retrieval (CBIR) systems have gained popularity in recent years [7]. Image feature extraction is the core part of such systems. While many such systems use various image processing techniques to obtain image features, very few of them make use of associated text to assist the image feature extraction. Sinha et al [10] combined relevant information derived from free-text reports and information derived from the MR images to index the images. Lacoste et al

[8] used medical concepts from the Unified Medical Language System (UMLS) metathesaurus to index the reports and images. However, the indexing process is separate for text and images.

III. PROBLEMS IN EXISTING SYSTEM

In existing system the reports are in free text format and are usually unprocessed, there is a great barrier between the radiology reports and medical professionals making it difficult for them to retrieve and use useful information and knowledge from the reports. As the information is not accessible, it cannot be used for other related applications.

A. PROPOSED SYSTEM

The development of this new system contains the following activities, which try to recover the problems from the previous system: The main radiology report text mining system is to extract the medical findings in the free text reports, and then use the structured result for medical record data mining applications: report and image retrieval and structured text assisted image feature extraction. The existing system takes patients' radiology examination records as input. Each record consists of a report describing radiologist's observation on the examined body part of the patient, and a series of scanned images. The medical finding extraction module focuses on the free text radiology report. It applies natural language processing techniques and uses medical lexicons to extract the medical findings and their attributes from the free text and output them in a structured form. The following figure shows the architecture of our proposed system.(Figure 1)[1]

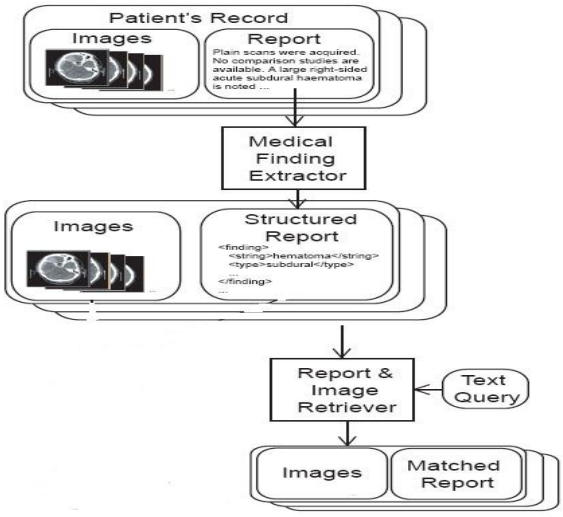

Figure 1. Architecture of proposed system

IV MODULES DESCRIPTION

Study of the System

In the flexibility of uses the interface has been developed a graphics concepts in mind, associated through an AWT & Swing Interface. The GUI's at the top level has been categorized as follows

Number of Modules

The system after careful analysis has been identified to be presented with the following modules:

The Modules involved are

User Interface Module

Medical Finding Extractor

Report and Image Retriever

Text-Assisted Image Feature Extractor

User Interface Module

Rich user interface developed in order to give the number of Nodes. It allows user to select the radiology report and text data. It allows extracting the data from radiology image

Medical Finding Extractor

The goal of medical finding extractor is to extract the medical findings in the radiology reports. It takes the brain CT radiology reports as input, extracts medical findings and their modifiers, and outputs them in a structured form. We used the semantic approach to achieve our text mining task. The system consists of the following components: term mapper, parser, finding extractor, and report constructor.(Figure 2)[1]

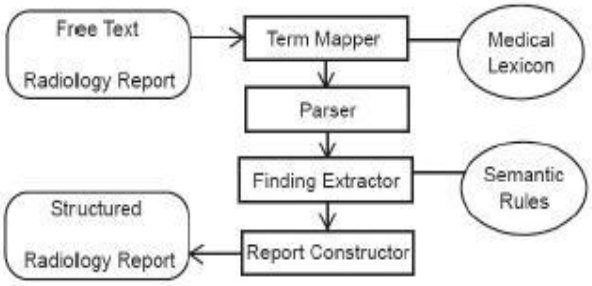

Figure 2. Medical finding extractor

The term mapper maps single-word and multiple-word terms to our medical lexicon and normalizes the terms to standard forms. Medical Subject Headings (MeSH) is a large controlled vocabulary developed by National Library of Medicine used for medical texts indexing [2]. However, as MeSH does not cover the entire set of vocabulary for brain CT radiology reports, it does not reach the degree of specificity required in the reports. Our medical lexicon is constructed from MeSH, other radiology and anatomy thesaurus, and actual brain CT radiology reports our parser is developed based on the Stanford Parser [4] and trained using labeled brain CT radiology reports. The parser parses each sentence and outputs the typed dependency tree, which shows the syntactic relations between the words and phrases in the sentence. For example, the typed dependency graph of the sentence "There is acute subdural

hemorrhage in the left frontal lobe." is shown in Figure. The finding extractor selects the findings and their modifiers according to a set of semantic rules. A medical finding in the brain CT report refers to the abnormality of patient's medical condition.(Figure 3)[1]

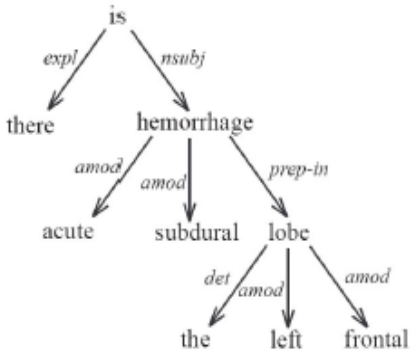

Figure 3. Example parse tree

The query analyzer is essentially the same as the medical finding extractor we described in Subsection. Instead of taking a full radiology report as input, the query analyzer takes the user query as input, which usually consists of a phrase or a few words. When a query from the user is entered to the system, for example, "acute subdural hematoma, no skull vault fracture", the query analyzer extracts the medical finding from the query and structure it as below.

<finding>

<string>hematoma</string>

<type>subdural</type>

<duration>acute</duration>

</finding>

<negative finding>

<string>fracture</string>

<location>skull vault</location>

</negative finding>

The retriever then searches the structured reports and images in the database and returns the ones that match the structured query. There are two modes of retrieval: exact match and partial match. Exact match returns results that are mostly needed by the user, whereas partial match returns results similar to what the user queries and facilitates the user to compare similar cases. Under the exact match mode, only the reports containing exactly the same findings and modifiers as in the query are returned. Take the same query "acute subdural hematoma, no skull vault fracture" for example, in the mode of exact match, reports with "acute subdural hematoma with fracture in skull vault" is not returned, as the "skull vault fracture [5]" finding in the query is negative, whereas in the report it is positive. The explicit labeling of positive and negative findings in medical finding extractor described in Subsection is also for more accurate retrieval here. Reports with "chronic hematoma" or "longitudinal fracture through right temporal bone" are also not returned in the retrieval results, as their modifiers (duration, type, location) of the findings do not match with the queries. On the other hand, if the user chooses to use partial match, then reports with findings and modifiers that applications such as content based information retrieval (CBIR), image classification etc. These features are usually extracted based on the image's information by image processing only. It is an advantage if associating text can assist in the process, as more information is utilized. With the assistance of structured reports associated to the images, features are more accurately extracted from images

One of the goals of radiology image mining is to detect any abnormality of the body part examined. For brain CT images, hematoma and hemorrhage detection is one of the major tasks [9]. We use structured report of brain CT scan of severe head injury in our project to help to extract features of any hematoma or hemorrhage in the brain. If there is such abnormality in the brain, the detailed information about it is depicted in the structured report in terms of medical finding and its modifiers. In the example shown, "type", "duration", and "location" are modifiers for medical finding "hemorrhage". Hematoma and hemorrhage types include "subdural", "epidural", "intracerebral", "intraventricular", "subarachnoid" etc. When they appear in the structured report as modifier values [6], they entail the shape, the location and sometimes the size information of the hematoma or hemorrhage as well. Other modifiers like "size" also entail image feature related information and are helpful for the image feature extraction module.

In the radiology image feature extraction module as shown in Figure, after we register the brain CT image to brain atlas, along with the image features we extracted from the image itself, we use the location information from the "type" and "location" modifiers in structured report to select the area of interest in the image. When the candidates of abnormality regions are produced from image mining procedures, we use the shape, size and intensity information from structured report to resize the area of interest, draw contours, and segments the abnormality region in the brain. The image feature extraction module then uses other image processing techniques to select other features. The extracted features form a structured image, which can be used for further image mining tasks, such as medical image classification

USE CASE DIAGRAM

Use case diagram shows a set of use cases and actors special kind of class) and their relationships. Use case diagrams address the static use case view of a system. These diagrams are especially important in organizing and modeling the behaviors of a system.

In the use case diagram the actor is Radiologist and the use cases are record, image, report, MFE, image feature extraction, structure for image and structure for report, query.(Figure 4)

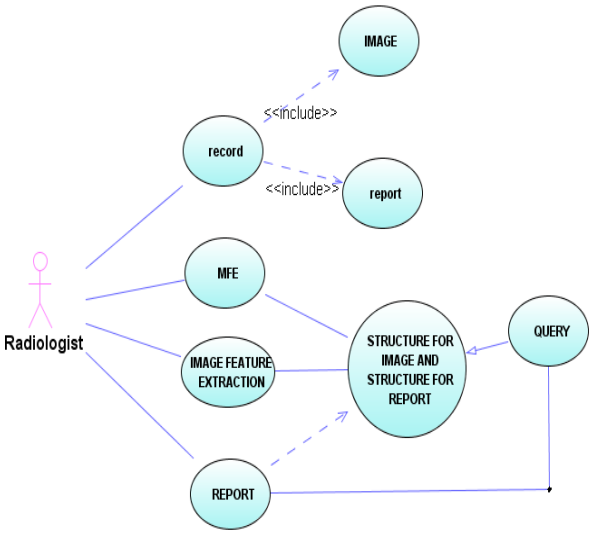

Figure 4. Use case diagram

SEQUENCE DIAGRAM

Sequence diagram is a kind of an interaction diagram. Interaction diagrams address the dynamic view of a system. A sequence diagram is an interaction diagram that emphasizes the time-ordering of messages. Here in this sequence diagram the objects are record, MFE, IFE, SI&R, Query, and Report.(Figure 5)

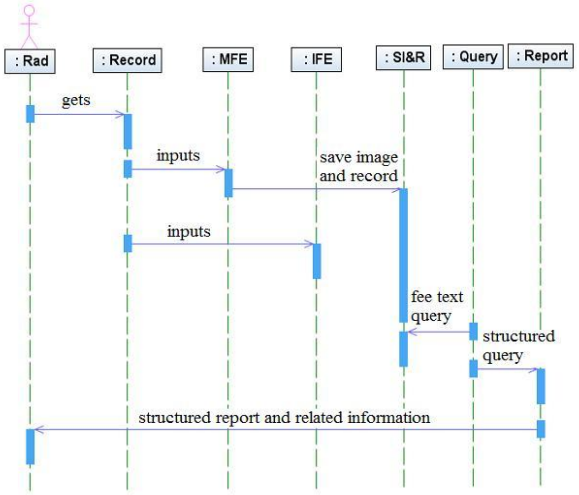

Figure 5. Sequence diagram

Image is browsed and loaded similar to how we load the text report ……(Figure 6)

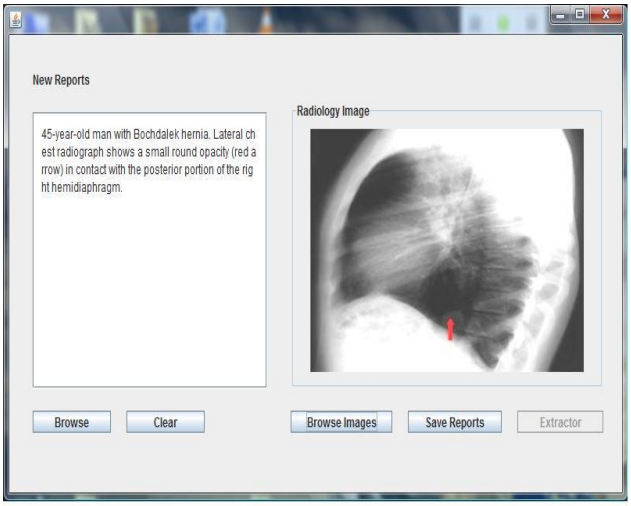

Figure 6 Loading of text report

The loaded image and text are to be saved in the database with a respective name.

The window consists of a text box where a text query can be entered related to the diagnosis required. This gives two kinds of matches-partial and exact matches… (Figure 7)

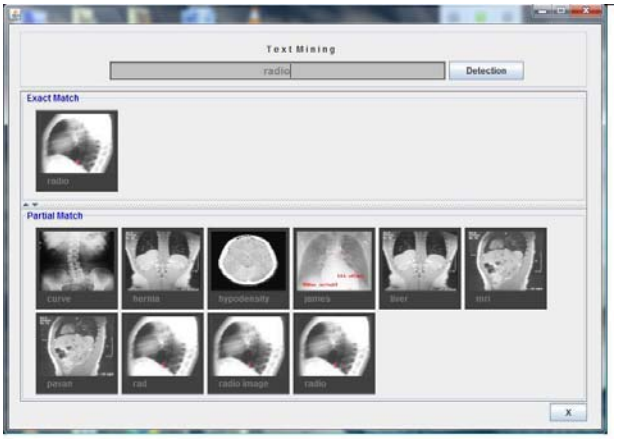

Figure 7. Matching results

Clicking on any image from the partial or exact matches pops a window that shows the radiology image and the corresponding text description of the image…(Figure 8)

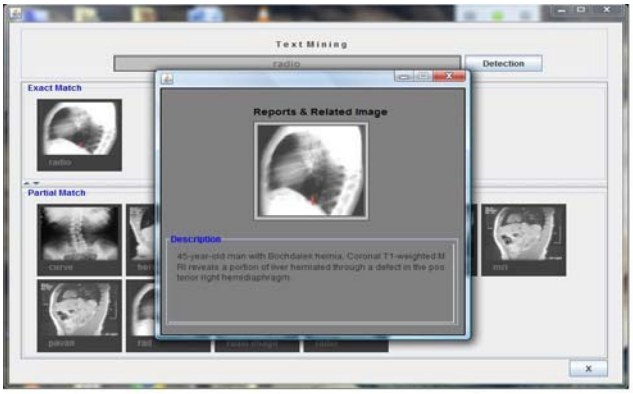

Figure 8. Radiology report and its description

On entering an unrelated query a window containing the message 'there is no record' is popped.

For details of implementation of this paper, please refer to the website http://sites.google.com/site/kpresearchgroup

CONCLUSION & FUTURE ENHANCEMENTS

To improve the performance of the medical finding extraction module, we will look into problems like obbreviation mapping, term normalization (including misspellings), and co reference resolution. For the report and image retrieval module, currently available query method is by free text. In the future, we will add more query methods including image query, so that the user can submit an image (instead of text) and search for similar images in the database. We will improve partial match mode of report and image retrieval by using report ranking, similar to page ranking in webpage searching. More comprehensive testing and evaluation of the system is also yet to be carried out in the future. As many records of radiology examination in the hospital database or public educational database and records of newly scanned images without radiologist's reports consist of only images, we will develop a system to automatically generate radiology reports using statistical machine translation (SMT). In our project, radiology image is considered as a special language, and our goal is to translate the image to free text report. This system is yet to be implemented in the future.

ACKNOWLEDGMENT

The authors wish to thank P.Satya Shekhar Varma, V.Pranav Kumar and R.Pavan Kumar for implementing these concepts.

REFERENCES

[1] Tiamxia Gong, Chew Lim Tan, Tze Yun Leong, "Text Mining in Radiology reports", IEEE, 2008

[2]Factsheet : Medical subject headings.

http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[3] D. B. Aronow, F. Feng, and W. B. Croft. Ad hoc classification of radiology reports. Journal of the American Medical Informatics Association, 6(5):393–411, September October 1999.

[4] M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In Proc. The fifth international conference on Language Resources and Evaluation (LREC2006), Genoa, Italy, 2006.

[5] S. Dominich, J. Goth, and T. Kiezer. Neuradir: Web-based neuroradiological information retrieval system using three methods to satisfy different user aspects. Computerized Medical Imaging and Graphics, 30:263–272, 2006.

[6] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson. A general natural language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1(2):161–174, March April 1994.

[7] T. Gong, R. Liu, C. L. Tan, N. Farzad, C. K. Lee, B. C. Pang, Q. Tian, S. Tang, and Z. Zhang. Classification of ct brain images of head trauma. In Proc. The second IAPR International Workshop on Pattern Recognition in Bioinformatics (PRIB2007), pages 401–408, 2007.

[8] R. Krishnapuram, S. Medasani, S.-H. Jung, Y.-S. Choi, and R. Balasubramaniam. Content-based image retrieval based on a fuzzy approach. IEEE Transactions on Knowledge and Data Engineering, 16(10):1185–1199, October 2004.

[9] C. Lacoste, J.-H. Lim, J.-P. Chevallet, and D. T. H. Le. Medical-image retrieval based on knowledge-assisted text and image indexing. IEEE Transactions on Circuits and Systems for Video Technology, 17(7):889–900, July 2007.

[10] R. Liu, C. L. Tan, L. T. Yun, C. K. Lee, B. C. Pang, C. C. T. Lim, Q. Tian, S. Tang, and Z. Zhang. Hemorrhage slices detection in brain ct images. In Proc. The nineteenth conference of the International Association for Pattern Recognition (IAPR2008), 2008. Accepted.

[11] U. Sinha, A. Ton, A. Yaghmai, R. K. Taira, and H. Kangarloo. Image content extraction: Application to mr images of the brain. Radiographics, 21(2):535–547, March April 2001.