Text Mining in Radiology Reports by SVM Classifier

Anuradha K. Bodile
Department of Computer Technology
YCCE
Nagpur, India

Manali Kshirsagar
Department of Computer Technology
YCCE
Nagpur, India

ABSTRACT
Medical text mining has gained increasing popularity in recent years. Now a days, large amount of medical text data are daily generated in health institutions, but never refer again as it is very time consuming task. In Radiology research area, most of the reports are in free text format and usually unprocessed, hence it is difficult to access the valuable information for medical professional unless proper text mining is not applied. There are some systems for radiology report information retrieval like MedLEE, NeuRadIR, CBIR but very few of them make use of text associated with image This paper proposes a text mining system to deals with this problem by using statistical machine translation approach. The radiology report is given to the system as input and system will return the similar report match with the entered report from the database.The SVM classifier is use in SMT approach to find the match report. Precision and Recall accuracy measures are used for evaluation purpose.

Keywords
Text mining, Radiology report, Image feature extractor, report retriever.

1. INTRODUCTION
In medical technology due to wider adoption of electronic medical record systems, many reports and large medical text data are generated in hospitals and other health institutions daily[1]. The medical report include the patient's medical condition in detail like medical history; prescription and results. Although, these text data contain valuable information, not referred to again. The valuable data that are not used to full advantage. A similar situation occurs in the field of radiology. The reports are not in proper format and usually unprocessed, making it difficult for radiology professionals to retrieve and use useful knowledge and information from the reports. Radiologist often face the burden of reviewing prior study reports before reading the patient's current study[10] and it's a very time consuming and difficult task. For making the clinical decision [2],it is advantageous to use the previous report of the same structure, of the same region, and of the same disease. A less experienced radiologists, use a reference text to find images that are similar to the query image for guidance. Hence, medical CBIR systems can aid doctors in analysis by retrieving images with known pathologies that are similar to a patient's image.

A text mining system[1]extracts and uses information from reports. There are so many existing system for radiology report information retrieval like MedLEE**,** NeuRadIR, CBIR but very few of them make use of text and image features togetherly.The proposed system will rerieve the report by using statistical machine translation approach. The system consist of feature extraction module make use of both text and image features providing the more strong probability to get match report. The system consists of main modules are medical term extraction, feature extraction, structured dataset creation, report retriever. The medical finding extraction module extracts medical findings in radiology reports, which describe radiologist's observations of the patient's medical conditions in the associated medical images. The reports are enter in the system require preprocessing to remove irrelevant contents and natural language processing techniques to process the text and to extract the medical findings and associated information. The color and texture feature are extracted in the image extraction module. The structured database are created using extracted text and image feature. The retrieval module takes user's input and returns the reports and images that match the query.

2. RELATED WORK
Tianxia Gong proposed the idea of text mining system which extracts the useful information from reports. They used brain CT radiology reports as domain for testing of system. The System includes feature extraction and retrieval of reports[1]. The Friedman encode radiology reports using semantic approach. Their Medical Language Extraction and Encoding System (MedLEE), uses a grammar and lexicon to determine the structure of the text and transform the text to the target structure and map to vocabulary and determine the phrase using synonyms[6].The Jeffrey Friedlin developed MEDAT system for text mining of medical reports. The system use semantic base index with large number of proposition to perform analysis of medical reports[12]. Taira retrieve medical term or findings in the reports using field theoretical approach focus on building a parser using the "word-word link "concept to output dependency diagram [8]. The parser outputs the dependency diagram using statistical methods.

Dominich proposed a web-based neuroradiological information retrieval system (NeuRadIR). They structured the radiology reports and permit users to retrieve the medical records by using three ways: Boolean, hyperbolic and interaction[5].Krishnapuram proposed a Fuzzy image Retrieval system (FIRST) to represent images using the concept fuzzy attributed relational graph(FARG). The system was based on Content based image retrieval (CBIR) concept[3]. Image feature extraction is the main part of such systems. In the system, region or object is represented by node using attributes like size, shape and relation between them is represented by edge. While many such systems use various image processing techniques to obtain image features, but only few of them make use of associated text to assist the image feature extraction. The Lacoste merge the idea of image associated with text to index the reports and images using the medical concepts from the Unified Medical Language System (UMLS)[2].Two different indexing process are developed for those: Global indexing used for image and Local indexing used for text.

3. PROPOSED SYSTEM

The aim of our text mining system is to extract the medical findings in the text reports, and then use the structured result for radiology report mining applications. The system consist of medical term extractor, image feature extractor , feature storation, report and image retriever.

The system uses statistical machine translation approach. Basically, there are two main approaches of machine translation, the older one Rules Based Machine Translation (RBMT) and the more recent Statistical Machine Translation (SMT). Basically, the RBMT approach that is usually word based and most modern SMT systems are phrased based and performs translations using Probability function. In the SMT models, the system use SVM (support vector machine) which is helpful in text and hypertext categorization and classification of images and also useful in training part.

SVM classifier basically use for classification and regression purpose. It classifies two classes by using hyperplane. The proposed system use SVM classifiers to find the report match with enter query report. The System use C_SVC type of classifier with RBF kernel type. The extracted feature will be use as trained samples by SVM classifier.

The Medical finding extractor extracts the medical findings or the useful information from the reports. The image feature extractor extracts the features of images like edges and texture. The extracted feature are stored and then giving to SVM classifier for training purpose. Finally, the report and image retriever retrieves the similar report match with the report entered by user.

3.1 Text Extraction

The collected radiology reports [11] contain the patient's medical details and the unnecessary contents like html tags and stopwords. So, by applying preprocessing html tags and stopwords are removed from reports.

The next preprocessing step is removal of stopwords. The stopwords present in the report like at, the, and, or, who, what etc. which are not useful in medical field are removed from report. The stopwords are compared with the list of stopwords that is already created and then removed. After finishing the preprocessing step, applying NLP techniques like stemming, term mapping and semantic rule helps out in finding the medical terms.

The stemming finds the root word of the given word. For e.g. "Cancerous" has a root word "Cancer". In stemming, Porter stemmer function is used for removing all suffixes like "ness", "in", "able". For example, after stemming "radiography" shows its root word "radiograph". Then the term mapping is used to count how many times a particular word is repeated in the document. Finally, the Semantic rule is applied to terms that are obtained after term mapping to find the similar or same meaning of those terms. To search the possible meaning, terms are mapped to Wordnet 2.1 which provides huge collection of synonyms and small definitions of a word.

3.2 Image Extraction

The radiology report contains image along with text report. The System also utilizes the image features for retrieval of report. The image features also supportive for positive result and increasing the possibility of getting more perfect match report. After extracting the text, the image feature extraction is applied to extract the features from images. The Image extraction module extracts the image features like its edges, color feature values and texture values of image. It includes the color feature extraction, edge detection and texture extraction. The color feature extraction extracts the feature vector of the image and the edge detection shows the boundary or the edges of the image. Various type of images contains in the report are brain, abdominal, hand fracture, etc

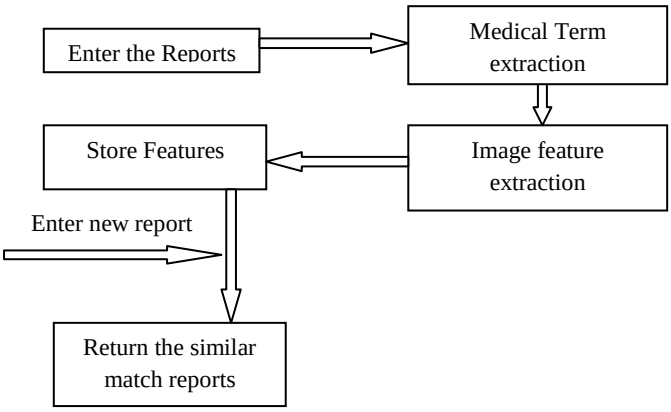


Fig 1. Flow of modules

After extracting the text, the image feature extraction is applied to extract the features from images. It includes the color feature extraction, edge detection and texture extraction. The color feature extraction extracts the feature vector of the image and the edge detection shows the boundary or the edges of the image. Precision and recall measures will be used for evaluating how many reports are correctly retrieved.

4. OBTAINED RESULTS

Report is open in the system that contains the patient's medical description and the unnecessary contents like html tags and stopwords as shown in Fig 1.

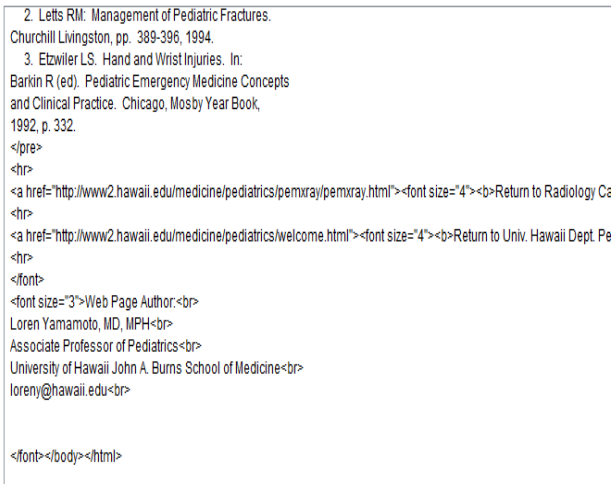
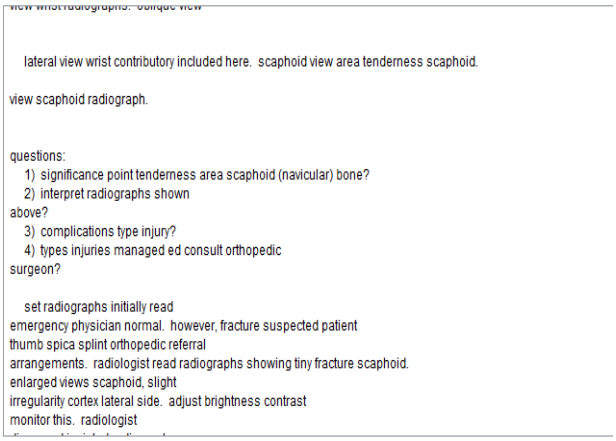


Fig 2. Open the report

The Irrelevant contents from the report are removed by preprocessing in which html tags and stopwords are removed as shown in Fig. 3.



lateral view wrist contributory included here. scaphoid view area tenderness scaphoid.

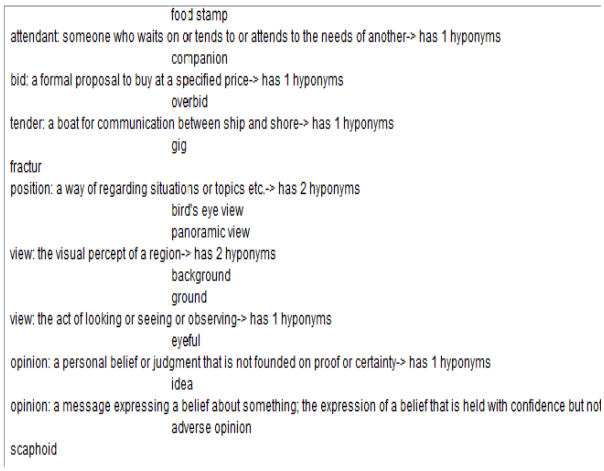
view scaphoid radiograph.

questions:
1) significance point tenderness area scaphoid (navicular) bone?
2) interpret radiographs shown
above?
3) complications type injury?
4) types injuries managed ed consult orthopedic
surgeon?

set radiographs initially read
emergency physician normal. however, fracture suspected patient
thumb spica splint orthopedic referral
arrangements. radiologist read radiographs showing tiny fracture scaphoid.
enlarged views scaphoid, slight
irregularity cortex lateral side. adjust brightness contrast
monitor this. radiologist

Fig 3. Preprocessing Of Report

After preprocessing, medical term is extracted by applying stemming, term mapping and semantic rules from Fig 4. In the text extraction, finally the terms are generated with their hyponyms.



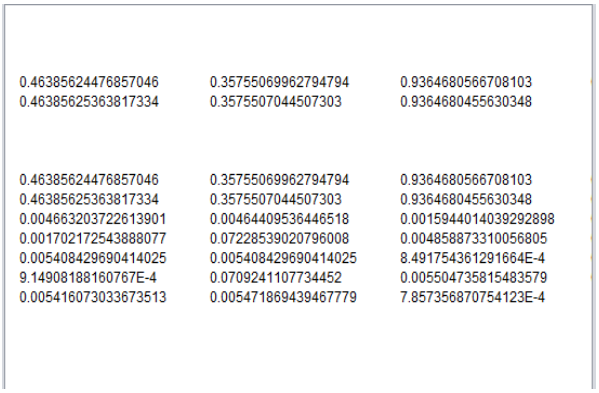
food stamp
attendant: someone who waits on or tends to or attends to the needs of another-> has 1 hyponyms
companion
bid: a formal proposal to buy at a specified price-> has 1 hyponyms
overbid
tender: a boat for communication between ship and shore-> has 1 hyponyms
gig
fractur
position: a way of regarding situations or topics etc.-> has 2 hyponyms
bird's eye view
panoramic view
view: the visual percept of a region-> has 2 hyponyms
background
ground
view: the act of looking or seeing or observing-> has 1 hyponyms
eyeful
opinion: a personal belief or judgment that is not founded on proof or certainty-> has 1 hyponyms
idea
opinion: a message expressing a belief about something; the expression of a belief that is held with confidence but not
adverse opinion
scaphoid

Fig 4. Medical Term Extraction

The image extraction module extracts features like color, texture feature vectors value. Canny edge detector shows the boundary or edges of the image as shown in fig .4.

Fig 5. Image Extraction

In Feature extraction, extracting the color and texture feature values of image. The feature vector values of Color and texture of image are shown in fig 4.



0.46385624476857046	0.35755069962794794	0.9364680566708103
0.46385625363817334	0.3575507044507303	0.9364680455630348
0.46385624476857046	0.35755069962794794	0.9364680566708103
0.46385625363817334	0.3575507044507303	0.9364680455630348
0.004663203722613901	0.00464409536446518	0.00159440140392928
0.001702172543888077	0.07228539020796008	0.004858873310056805
0.005408429690414025	0.005408429690414025	8.491754361291664E-4
9.14908188160767E-4	0.0709241107734452	0.005504735815483579
0.005416073033673513	0.005471869439467779	7.857356870754123E-4

Fig 5.Feature Extraction

SVM Traing Started
SVM Parameters :
SVM Type : C_SVC
Kernel Type : RBF
gamma : 0.5
nu : 0.5
cache size : 20000
C : 1
eps : 0.001
p : 0.1

Fig 6. Train SVM

SVM is trained for retrieval of report as shown in Fig.6.

5. CONCLUSION

This paper proposes a text mining system which will retrieve the similar radiology reports from structured database that match with entered report or required by user. The text and image extraction is performed to find the medical term and image feature resp.In the text extraction, extract medical findings and canny edge detection is done in image extractions.For training of reports,trained SVM classifier as it gives better performance in text and image categorization.In future,system will retrieve the similar report using SVM classifier.OurThe system saves the burden of reviewing prior study reports, saves the valuable time of medical professionals (radiologists, physicians, and researchers) and greatly helpful for less experienced radiology practitioner for guidance purpose.

6. REFERENCES

[1] Tianxia Gong,Chew Lim Tan, Tze Yun Leong, Cheng Kiang Lee, Boon Chuan Pang, C. C. Tchoyoson Lim, Qi Tian, Suisheng Tang, Zhuo Zhan, "Text mining in Radiology Reports", 8th IEEE International Conference on Data Mining, 2008, pp.1550-4786.

[2] C. Lacoste, J. H. Lim, J. P. Chevallet, D. T. H. Le, "Medical-image retrieval based on knowledge-assisted text and image indexing", IEEE Transactions on Circuits and Systems for Video Technology, 2007, pp. 889–900.

[3] 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,2004, pp.1185–1199.

[4] Jeffrey Friedlin Malika Mahoui, Josette Jones, Patrick Jamieson, "Knowledge Discovery and Data Mining of Free Text Radiology Reports", 1st IEEE International

Conference on Healthcare Informatics, Imaging and Systems Biology, 2011, pp. 4407-9780.

[5] S. Dominich, J. Goth, T. Kiezer, "Web-based neuro radiological information retrieval system using three methods to satisfy different user aspects", Journal of Computerized Medical Imaging and Graphics, 2006, pp. 263-272.

[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, 1994, pp. 161-174.

[7] Issam El-Naqa, YonGyi Yang, Nikolas Galatsons, "A Similarity Learning Approach to Content Based Image Retrieval: Application to Digital Mammography", IEEE Transaction on Medical imaging, 2004, pp. 245-263.

[8] R. K. Taira, V. Bashyam and H. Kangarloo, "A field theoretical approach to medical natural language processing", IEEE Transactions on Information Technology in Biomedicine, 2007, pp. 364-373.

[9] R. K. Taira, S. G. Soderland, and R. M. Jakobovits. "Automatic structuring of radiology free-text reports." Radiographics, 2001, pp.237–245.

[10] Aisan Maghsoodi, Merlijn Sevenster, Johannes Scholtes, Georgi Nalbanto Philips Research, "Sentence-based Classification of Free-text Cancer Radiology Reports", 25th International Symposium on Computer-Based Medical Systems , 2012, pp.978-4678-2051.

[11] www.hawaii.edu/medicine/pediatrics/pemxray.