International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 2 228 – 231

Text Mining in Radiology Reports

Ms.Anuradha k. Bodile
Computer Science and Engineering
Yeshwantrao Chavan College Of Engineering
Nagpur, India
anu28bodile@gmail.com

Dr.Manali Kshirsagar
Computer Science and Engineering
Yeshwantrao Chavan College Of Engineering
Nagpur,India
manali_kshirsagar@yahoo.com

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 not in proper format. 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. This paper proposes a text mining system to deal with this problem by using statistical machine translation approach. 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 features. 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 system consist of medical term extractor ,image feature extractor, structured report creation, report and image retriever. Precision and Recall accuracy measures are used for evaluation purpose.**

Keywords- *Text mining, Radiology report, Image feature extractor*.

*****

I. INTRODUCTION

In medical technology [1] 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. The medical texts 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. These 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 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 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.

II. LITERATURE REVIEW

The Friedman [6] 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.By using radiology as a analysis domain,they analyzed four type of disease on 230 chest x-ray reports. Taira [9] 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 [5] 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. Krishnapuram [3] 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. 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 [2] 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).Two different indexing process are developed for those : Global indexing used for image and Local indexing used for text.

III. TEXT MINING 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 ,report and 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 perform 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 that will be useful in training part.

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 database are created by using extracted features with the help of SVM classifier. Finally, the Report and image retriever retrieves the similar report match with the report entered by user.

The collected radiology reports[11] contains the patient's mledical details and the unnecessary contents like html tags and stopwords which are not essential. So, by applying preprocessing html tags and stopwords are removed from reports.

For removing html tags, the replaceall () function is used which replaces first argument with the second argument. The html tags present in the report like \href, \br, are replaced with empty spaces" ".

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. It is a part of text feature extraction.

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. Finally, all possible synonyms of terms are generated.

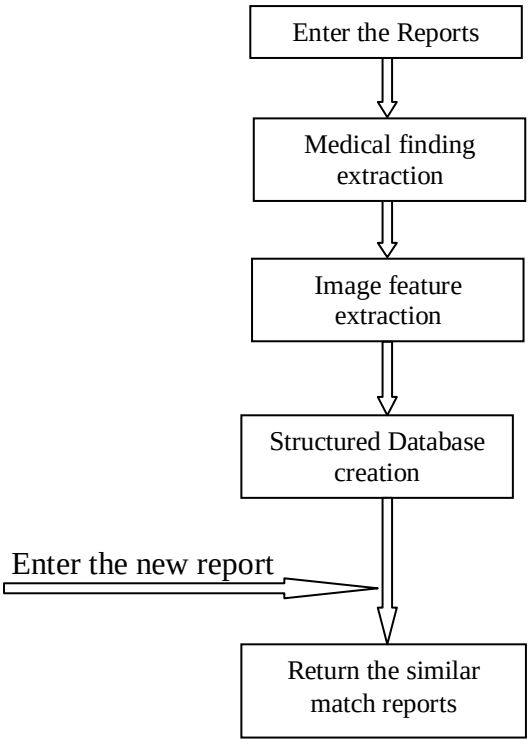

Fig 1.Flow of modules

 ISSN: 2321-8169
228 – 231

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.

IV. OBTAINED RESULTS

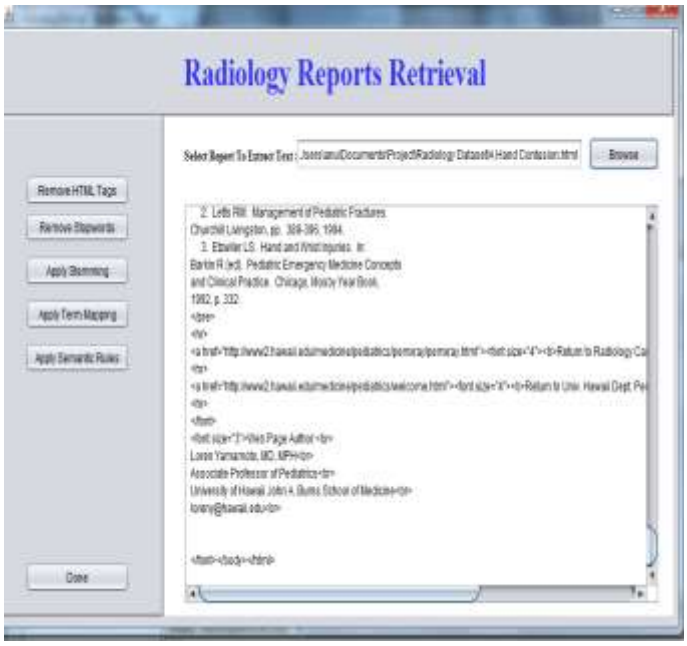

Fig 1.Open the Report

Report is open in the retrieval system that contains the patient's medical description and the unnecessary contents like html tags and stopwords as shown in Fig 1. The contents are removed from report by preprocessing shown in fig 2.

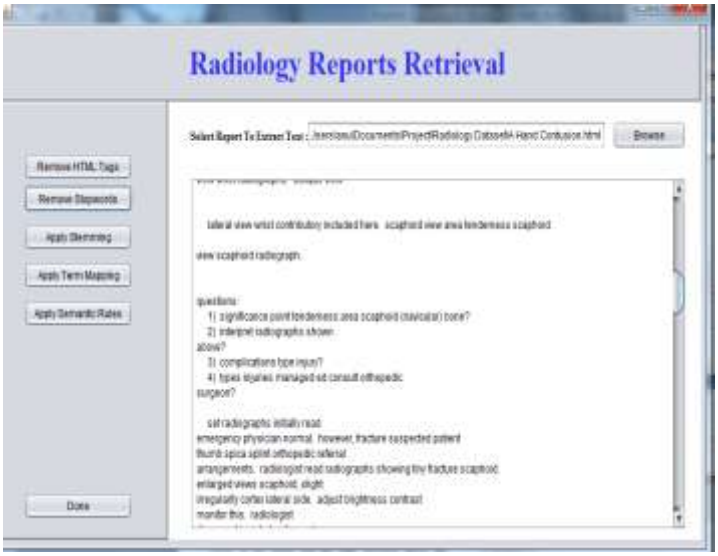

Fig 2. Preprocessing Of Report

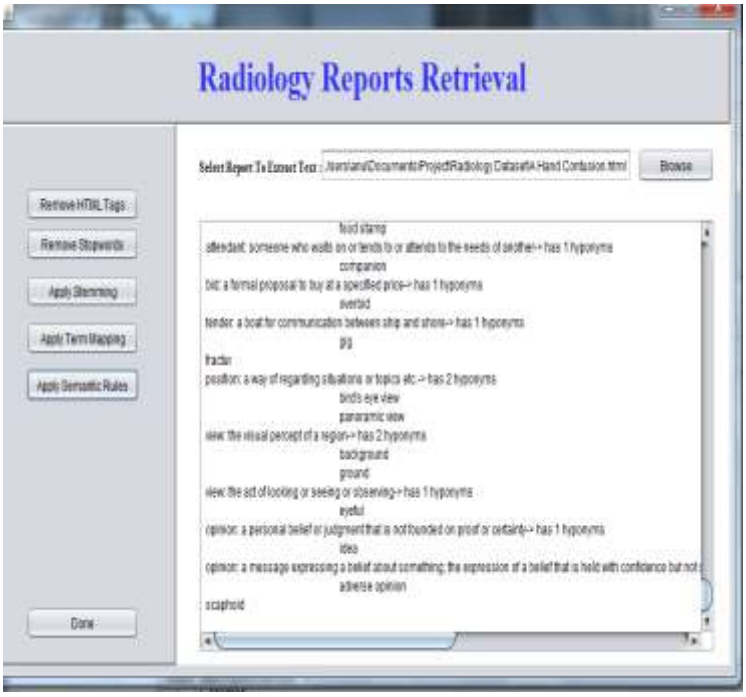

Fig 3. Medical term Extraction

After preprocessing, medical term is extracted by applying stemming, term mapping and semantic rules from fig 3. In the text extraction, finally the terms are generated with their synonyms.

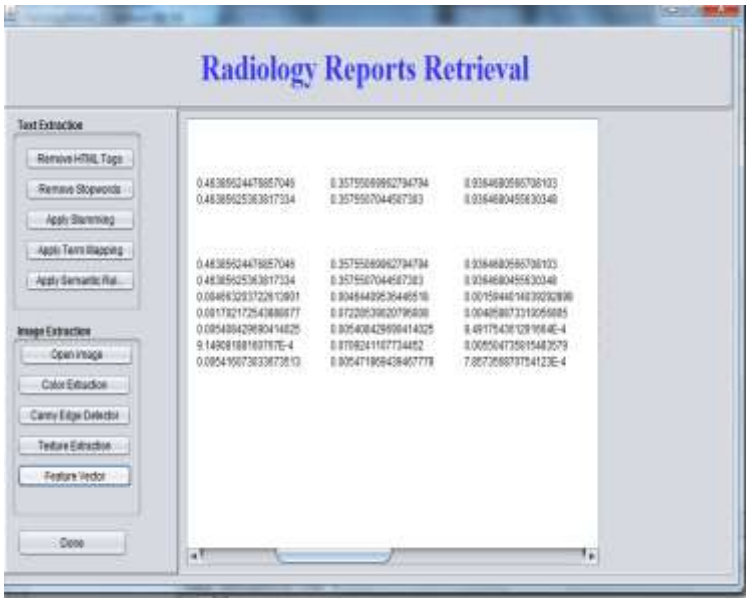

Fig 4. Image Extraction

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

V. CONCLUSION

This paper proposes a text mining system which returns the similar radiology reports from structured database that match with entered report. It provides the library system to refer the previous reports using text and image features. The 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.

VI. 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 aJamieson, "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 Mamography", 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