

# Text Mining in Radiology Reports

Tianxia Gong, Chew Lim Tan, Tze Yun Leong  
Department of Computer Science, School of Computing, National University of Singapore  
{gong\_tianxia, tancl, leongty}@comp.nus.edu.sg

Cheng Kiang Lee, Boon Chuan Pang, C. C. Tchoyoson Lim  
National Neuroscience Institute, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433  
{cheng\_kiang\_lee, boon\_chuan\_pang, tchoyoson\_lim}@nni.com.sg

Qi Tian, Suisheng Tang, Zhuo Zhang  
Insitute of Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632  
{tian, suisheng, zzhang}@i2r.a-star.edu.sg

## Abstract

*Medical text mining has gained increasing interest in recent years. 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 paper, 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, a report and image retriever, and a text-assisted image feature extractor. In evaluation, the overall precision and recall for medical finding extraction are 95.5% and 87.9% respectively, and for all modifiers of the medical findings 88.2% and 82.8% respectively. The overall result of report and image retrieval module and text-assisted image feature extraction module is satisfactory to radiologists.*

## 1. Introduction

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.

A similar situation occurs in the field of radiology. As

the 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.

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, and text-assisted image feature extractor. The medical finding extraction module automatically extract medical findings in the brain CT radiology reports, which describe radiologist's observations of the patient's medical conditions in the associated medical images. The system takes a set of unstructured reports as input, applies natural language processing techniques to process the text and the rules in frequent medical finding pattern database to extract the medical findings and related information. The retrieval module takes user's free text query as input and returns the reports and images that match the query. The image feature extractor takes radiology images as input, makes use of the information extracted from associated report, and extract features from the image.

## 2. Related Work

Text mining applications in medical domain began to attract attention. These applications include medical text classification and categorization [2], medical knowledge dis-

covery [11], clinical events detection and surveillance [10] etc.

One of the prerequisites of these applications is to extract the information in the medical text and represent it in a structured form. Friedman et al [5] used a semantic approach to encode radiology reports. Their Medical Language Extraction and Encoding System (MedLEE), employs a grammar and lexicon to determine the structure of the text and transform the text to the target structure. Then it uses a mapping knowledge base to reduce stylistic variations and encode the regulated phrase into unique concepts using a synonym knowledge base. They tested on 230 chest x-ray reports with four types of diseases; the recall and precision of the system were 70% (85% if training on query) and 87% respectively. Taira et al [14], on the other hand, used a statistical approach to extract medical findings in the reports, with focus on building a statistical parser to output dependency diagram using the concept “word affinity” [13]. The parser uses statistical methods to outputs the dependency diagram that maximizes the “affinity” and “valence” probabilities. Their system evaluation was focused on word-word link classification (statistical parser) and link interpretation.

Dominich et al [4] 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 [12] 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) meta-thesaurus to index the reports and images. However, the indexing process is separate for text and images.

### 3. The text mining system in radiology

#### 3.1. Background

A CT head examination usually consists of a radiology report and a set of brain images. The text below is the medical finding description part of a brain CT radiology report.

A large right-sided acute subdural haematoma is noted with a maximal depth of 20 mm. Basal cistern effacement is noted. A small left occipital scalp haematoma is seen. Likely left frontal lobe contusion. No evidence of skull vault fracture.

Medical findings in brain CT examination refer to disorders and diseases, i.e. the abnormality of the brain. For example, “haematoma” and “midline shift” are medical findings in the example report. Apart from the findings, radiologists also note down more specific details of the findings in the reports. They can be considered as attributes or modifiers of the findings, which include type, duration, location, amount, size, direction, probability, and seriousness.

#### 3.2. General system architecture

The goal of our 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.

As shown in Figure 1, the 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.

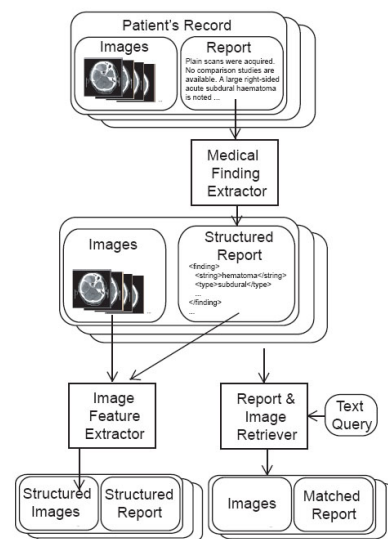


Figure 1. General architecture of the system

The patient record database is then indexed using the extracted medical findings. The report and image retrieval

module takes user’s free text query as input and outputs the reports with associated images that are matched to the query. There are two modes of matching, exact match and partial match, to cater different needs of the users.

The structured medical findings and their attributes (like size and location) are helpful to extract image features as well. Image feature extraction usually makes use of the information from the images only; however, with the help of structured text that is associated with the images, the feature extraction from the images can be more accurate.

### 3.3. Medical finding extraction from radiology reports

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. The system architecture is shown in Figure 2.

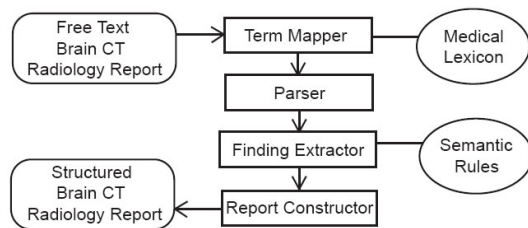


Figure 2. Medical finding extraction module

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 [1]. 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 [3] 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 3.

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 the pa-

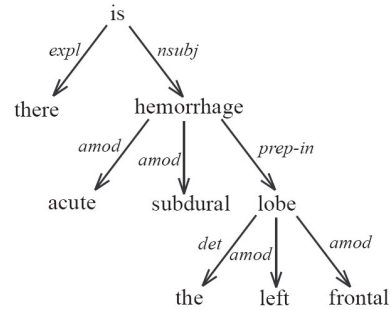


Figure 3. The typed dependency graph of example sentence

tient’s medical condition. For example, “hematoma”, “fracture”, “midline shift” are common findings in brain CT radiology reports. Each finding may have several modifiers that describe the properties of the finding, such as “type”, “location”, “duration”, “probability” etc. The finding extractor makes use of the intermediate result from term mapper and parser to locate medical findings and the modifiers, and uses a set of semantic rules to translate the syntactic relations in the result from the parser to logical relations between findings and their modifiers. The semantic rules were manually constructed and cover frequent patterns how the findings appear in the sentences. Negations are also detected by the finding recognizer. Findings associated with negative expressions are outputted as negative findings. It is necessary to extract negative findings and include them in the structured report explicitly as both negative and positive findings are significant to the medical personnel accessing the report.

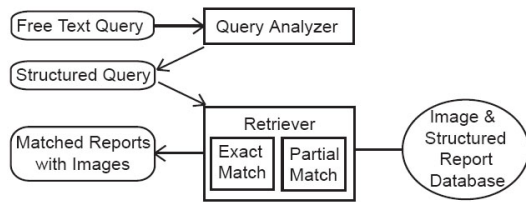
The report constructor then output the result from finding extractor in XML format. The output result of findings and modifiers of the example sentence in Figure 3 is shown in below. //

```
<finding>
  <string>hemorrhage</string>
  <type>subdural</type>
  <duration>acute</duration>
  <location>in left frontal lobe</location>
</finding>
```

### 3.4. Radiology report and image retrieval

The retrieval module builds a connection between the medical professionals and the report database, makes the content of the reports directly and conveniently searchable and retrievable. After the reports are structured in XML format, they can now be retrieved conveniently using queries. As the images are usually associated with reports to consist a complete patient record of examination, the images are retrieved as well along with the reports. As shown in

Figure 4, our report and image retrieval system consists of query analyzer and medical record retriever.



**Figure 4. Report and image retrieval module**

The query analyzer is essentially the same as the medical finding extractor we described in Subsection 3.3. 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 return 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” are not returned, as the “skull vault fracture” 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 3.3 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 query’s. On the other hand, if the user chooses to use partial match, then reports with findings and modifiers that

match part of the query are returned as well as the exactly matched ones. For example, the reports with “acute subdural hematoma with fracture in skull vault”, “chronic hematoma”, “longitudinal fracture through right temporal bone”, which are rejected in exact match mode, will be returned under the partial match mode.

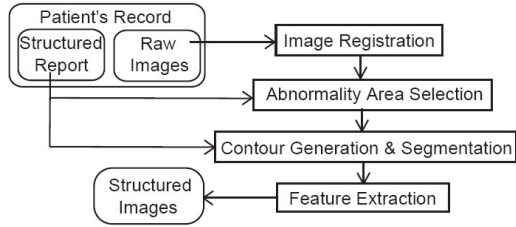
### 3.5. Text assisted medical image feature extraction

Like extracting information from free text, feature extraction from images are necessary for many image mining 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 in Section 3.3, “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, 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 5, 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 [6].



**Figure 5. Report and image feature extraction module**

In the process of contour generation and segmentation, the parameters the image processing procedures are adjusted under the effect of structured information from report to achieve better result. We consider it as a text-assisted training process for the image feature extractor too. When new images are obtained without accompanying report, the image feature extractor can use the set of parameters trained by the assistance of reports to output more accurate image features.

#### 4. Evaluation

We obtained 467 brain CT radiology reports from National Neuroscience Institute of Tan Tock Seng Hospital, Singapore. 367 of the reports were used for training, and 100 were used for evaluation of our system. The average length of the reports is 12 sentences, or 157 words. 753 positive findings, 167 negative findings, and 1520 modifiers were labeled in the testing reports. There are 26 different types of medical findings (positive and negative) in the testing reports.

The overall weighted average precision and recall for findings are 95.5% and 87.9% respectively. The detailed evaluation result for findings is shown in Table 1. Lower percentage of negative findings were correctly extracted compared to positive findings due to the additional task to recognize patterns indicating negation in the sentence. Nevertheless, we listed negative findings as a separate category opposed to positive findings, for radiologists and physicians often want to find past reports with explicit presence or absence of certain disorder/disease.

**Table 1. Evaluation result for medical findings extraction**

|                   | Precision | Recall |
|-------------------|-----------|--------|
| positive findings | 96.0%     | 89.5%  |
| negative findings | 93.1%     | 81.0%  |

For those correctly extracted findings, we calculated the

precision and recall for modifiers. The weighted averaged precision and recall for modifiers are 88.2% and 82.8% respectively. The detailed evaluation result for each type of modifier is shown in Table 2.

**Table 2. Evaluation result for modifiers extraction**

|             | Precision | Recall |
|-------------|-----------|--------|
| type        | 93.3%     | 85.1%  |
| duration    | 96.2%     | 93.8%  |
| location    | 86.1%     | 81.2%  |
| amount      | 81.6%     | 77.5%  |
| size        | 85.0%     | 81.9%  |
| direction   | 94.7%     | 90.0%  |
| probability | 91.7%     | 82.5%  |
| seriousness | 83.3%     | 75.0%  |

The system performed well on medical finding types and modifier types that are more frequent. For example, the extraction of “hematoma” and “midline shift” have higher accuracy compared to “inflammation”. Abbreviations, misspellings and short-hand writing affect the recall too, as they are more difficult to map to medical lexicon. The presence of ambiguous sentence structure often confuses the parser and may cause the parser to build the wrong dependency tree and thus creating wrong association between medical findings and modifiers or negation indicators.

As we index the reports with detailed modifiers of each medical finding, users are able to search reports or images with more specific request, instead of a more general query such as “CT head”) compared to the system we surveyed. Nevertheless, Web interface with public access to our system is yet to be implemented, in order for the end users to access and evaluate. For text assisted image feature extraction, the images features extracted are more accurate compared to those without the help of text. However, a detailed comparative study is needed to evaluate such improvement quantitatively.

The overall result is satisfactory to the radiologists in the hospital where the reports were obtained. However, more medical professionals from other medical institutions are yet to form a panel to evaluate the system independently.

#### 5. Future Work

To improve the performance of the medical finding extraction module, we will look into problems like abbreviation mapping, term normalization (including misspellings), and coreference 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 includ-

ing 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.

## 6. Conclusion

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, and text-assisted image feature extractor. The medical finding extraction module automatically extract medical findings and associated modifiers to structure brain CT 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 image feature extractor uses the extracted medical findings and their modifiers like location, type, and size etc. as additional information to that it obtains from pure image processing modules, to select area of interest (abnormality area), draw contours, segment the abnormality region, and extract features of the image. Although the training and testing of our system are in the domain of brain CT, the system can be extended and applied to other domains.

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.

## 7. Acknowledgement

This research is supported by NUS FRC grant R252-000-290-112 and MOE ARC grant R-252-000-349-112.

## References

- [1] Fact sheet: Medical subject headings. <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] C. Lacoste, J.-H. Lim, J.-P. Chevillet, 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.
- [9] 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.
- [10] E. A. Mendonca, J. Haas, L. Shagina, E. Larson, and C. Friedman. Extracting information on pneumonia in infants using natural language processing of radiology reports. *Journal of Biomedical Informatics*, 38:314–321, 2005.
- [11] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond. Medical data mining: Knowledge discovery in a clinical data warehouse. In *Proc. American Medical Informatics Association Annual Fall Symposium*, pages 101–105, 1997.
- [12] U. Sinha, A. Ton, A. Yaghmai, R. K. Taira, and H. Kangaroo. Image content extraction: Application to mr images of the brain. *Radiographics*, 21(2):535–547, March April 2001.
- [13] R. K. Taira, V. Bashyam, and H. Kangaroo. A field theoretical approach to medical natural language processing. *IEEE Transactions on Information Technology in Biomedicine*, 11(4):364–373, July 2007.
- [14] R. K. Taira, S. G. Soderland, and R. M. Jakobovits. Automatic structuring of radiology free-text reports. *Radiographics*, 21(1):237–245, 2001.