Automated Retrieval and Generation of Brain CT Radiology Reports

Gong Tianxia
Supervisor: A/P Tan Chew Lim

October 14, 2007

Abstract

With the advances of medical techniques, large amounts of medical data are produced in hospitals every day. However, most of current works are focusing on mining the medical images, whereas fewer work has been done for the text information associated with the medical images. Radiology reports contain rich information about the corresponding medical images but are often under mined. Therefore, our research topics will be focusing on information extraction from brain CT radiology reports, radiology reports assisted medical image content retrieval, and automatic generation of brain CT reports based on domain knowledge and associated images. Current medical record search systems will benefit from our research so that searching for information is more efficient and convenient. Doctors and radiologists can also be more efficient to conduct their research in the area using the improved system. The automatical generation of reports can give reference to radiologists. Our research will also be helpful to facilitate an education system for junior doctors and researchers in the area.

Acknowledgements

I would like to thank my supervisor, A/Prof. Tan Chew Lim, who has stimulated me to be interested in this research area and given me invaluable advice on my research topic.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advances of medical techniques, large amounts of medical data are produced in hospitals every day. Corresponding computer applications are also developed to analyze the data in various aspects. Among the areas, data mining in medical images is of growing interest. However, as most current works are focusing on mining the medical images (including content based medical image retrieval, medical image indexing, and etc.), fewer work has been done for the text information associated with the medical images.

Radiology reports contain rich information about the corresponding medical images. They usually contain detailed description of normal and abnormal findings of the interested body areas on the images, conclusion from the radiologists, and sometimes some bio data of the patients as well. As the radiology reports are often under mined, the valuable information contained in them are not utilized to build new computer applications or improve the existing applications.

We obtained around 300 patients' data on brain CT (Computer Tomography) scans from National Neuroscience Institute (NNI), Tan Tock Seng Hospital, Singapore. The data for each patient include one series of CT images on brain (18 to 30 slices or images) and one free text radiology report associated with the series of image. Our research topics will be focusing on

information extraction from brain CT radiology reports, radiology reports assisted medical image content retrieval, and automatic generation of brain CT reports based on domain knowledge and associated images.

1.1 Motivation

Most of the current medical image and report system index only on patients' particulars such as name and identity card number. Such system is inefficient and inconvenient as it is impossible for doctors or radiologists to remember each patient's name and id in order to retrieve their medical data. However, if the radiology reports are processed, mined and indexed, the medical record systems can be improved and become more efficient and convenient.

When such improvements are done, the doctors and radiologists can also be more efficient to do their research in the area. Instead of trying hard to gather text or image information based on patients' particulars from memory fragment (which is extremely time and effort consuming and often unsuccessful), they can easily search the improved medical record systems based on medical terms (such as disease, treatment, etc.), gather the information what they wanted much faster, and conduct research and studies on them more conveniently.

Moreover, the automatical generation of reports can give reference to radiologists for them to compare the generated results and their judgement in hospital.

Our research can also help to facilitate an education system for junior doctors and researchers in the area. Online medical record system open for doctors, medical students and researchers can be built based on our proposed research. Such online system can integrate more medical record data from various source and provide a platform for the community to exchange information and knowledge.

1.2 Organization of the paper

The rest of this paper is organized as follows. Chapter 2 includes a literature review covering current approaches on main aspects of the area, including information extraction from radiology reports, automatic generation of medical reports, and question answering systems in medical domain. Chapter 2 will also compare the approaches and discuss their strengths and limitations. Chapter 3 will discuss some possible research topics for the future PhD thesis. We also include some preliminary research and the experimental results in Chapter 4. Chapter 5 will then conclude the paper.

Chapter 2

Literature Review

2.1 Information Extraction from Radiology Reports

2.1.1 MedLEE: Medical Language Extraction and Encoding System

The Columbia University of New York (together with the Columbia Presbyterian Medical Center) has developed an NLP system MedLEE (MEDical Language Extraction and Encoding System) that identifies clinical information in narrative reports and transforms this textual information into a structured and conceptual representation [27, 89]. The main goal is to represent the knowledge of chest X-ray radiology reports, store it in a database and allow physicians to query the knowledge base by means of controlled vocabulary. Another realization is the integration of the NLP module with an automated decision-support system [28]. Although the MedLEE system is primarily semantically driven, the necessity of integrating syntactic knowledge is recognized: the development of a syntactic grammar is foreseen [27]. The semantic grammar consists of 350 directed conceptual graph (DCG) rules specifying

well-defined semantic patterns, their interpretations, and the underlying target structures [26] into which they should be mapped. The grammar rules are directly interpretable by Prolog. MedLEE had been improved and more features had been added over the ten year, and it remains one of the most cited and popular methods to process text for radiology reports.

In order to map the clinical information in the patient documents into a structured form, a formal model was designed to represent the clinically salient information. The fundamental design of this model is based on the information formats developed by the Linguistic String Project [85]. Two of the most relevant components of the representational model are Rad Finding Structure and Modifier, which represent the structures of the findings and the modifiers, respectively.

As shown in the figure 2.1, the text processing of MedLEE consists of three phases: parsing, phrase regularization, and encoding.

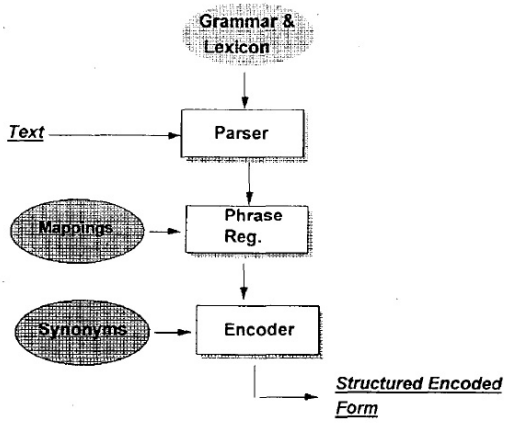

Figure 2.1: The general architecture of MedLEE

In first phase, the parser uses a grammar and lexicon (containing 1720 single-word entries and 1400 multi-word phrases) to determine the structure of the text and generates the preliminary structured output form for the clinical information. As the parser is driven by a semantic grammar that

is highly effective for handling structured text and common patterns, the system is very suitable for the domain of radiology. The grammar consists of rules specifying well-defined semantic patterns, their interpretations, and the underlying target structures into which they should be mapped. In MedLEE, the target structures correspond to the formal model of the domain. The semantic grammar incorporates pattern-matching and semantic techniques into one formalism (the semantic grammar), but is more general.

In second phase, MedLEE uses a mapping knowledge base to regularize the phrases and further reduce stylistic variations that occur in natural language. The mapping knowledge base consists of structured output forms of multi-word phrases that can be decomposed, so that the contiguous and non-contiguous lexical variants can be mapped to standard forms.

In third phase, the system uses a synonym knowledge base to map the standard forms into unique concepts associated with the controlled vocabulary. The synonym knowledge base consists of standard forms and their corresponding concepts in the controlled vocabulary and forms a critical bridge between the language of the text and the unique concepts in the controlled vocabulary.

The Medical Entities Dictionary (MED) developed at Columbia Presbyterian Medical Center (CPMC) was first served as a knowledge base of medical concepts that consist of taxonomic relations in addition to other relevant semantic relations. At later stage, MedLEE also experimented to use UMLS [48] as the knowledge base and had different evaluation results as in [29, 30, 71].

2.1.2 RADA: RADiology Analysis Tool

RADA, the Radiology analysis tool as described in [51] provides a method to index findings and associated information described in free text thoracic radiology reports. The system extracts mass lesion and lymph node findings, and links specific information associated with the findings such as size and

location.

Each glossary entry for RADA is represented by a concept, the smallest fragment of knowledge defined by RADA. A concept encodes both semantic and syntactic knowledge.

RADA's glossaries originate from two main sources, the Unified Medical Language Sources (UMLS) [48] and a specialized thoracic glossary. The specialized glossary augments the data found in the UMLS thus providing additional information necessary for the system. The structure of the glossaries is similar to the structure of Meta-1.4 of the UMLS. It provides a structured method for representing unique concepts by a number of textual representations.

Each concept in RADA system has three attributes as shown in Figure 2.2. The semantic code defines the semantic class to which the concept belongs. Likewise, the syntactic code defines which syntactic class to which the concept belongs. The text string defines the word or phrase to which the concept corresponds. The lexical analyzer uses the text string to match the concept with the text. To provide a human readable form of the concept, the text string is maintained throughout the rest of the system.

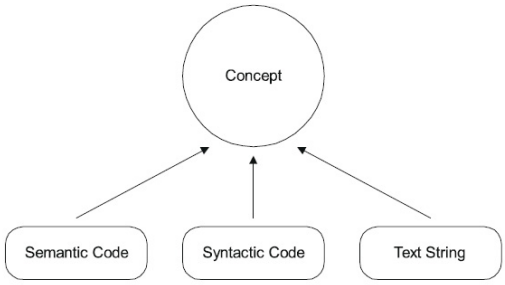

Figure 2.2: The concept representation in RADA

Lexical knowledge is encoded in knowledge hierarchies similar to type abstraction hierarchies [14]. Type abstraction hierarchies are multilevel knowledge structures that emphasize the abstract representation of information.

The meaning of a word or phrase is defined by a hierarchy of related concepts. A concept's semantic code encodes its position in the hierarchy. Different hierarchies exist for different classes of concepts. For example, anatomy concepts form one hierarchy and finding concepts form another. For each concept class they developed a hierarchy of terms and meanings as shown in Figure 2.3. The entries of the glossary are grouped by their meanings.

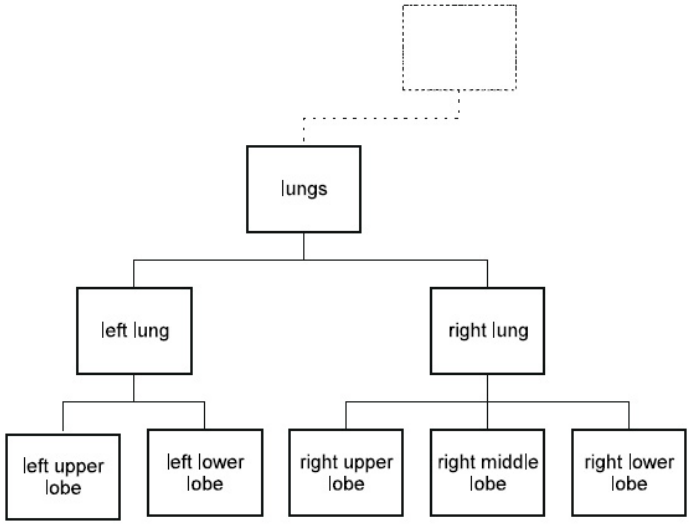

Figure 2.3: The type abstraction hierarchy in RADA

RADA uses entities to structure the details of extracted radiology findings and anatomy. Entities structure knowledge through a well-defined set of attributes. For example, an entity encoding a radiology finding will have attributes describing the size, location and architecture of the finding. Findings found in a report are stored in instances of the finding entity. RADA creates instances of the entities during finding analysis.

As shown in the figure 2.4, RADA consists of four parts: Lexical Analyzer, Finding Analyzer, Joiner and Reference Resolver.

The Lexical Analyzer first decomposes sentences into words and phrases. It matches words and groups of words to a specialized glossary of terms, and

decomposes the sentence into glossary entries.

The Finding Analyzer scans the sentence for articles (the, a, an) and pronouns. When an article is found, a finding entity is created and the sentence is parsed for phrases describing anatomy or findings. Parsing experts process fragments and recognize phrases that can be combined. Dierent parsing experts process the sentence, insuring that the sentence fragment matches one of several known forms. An example is shown in Fig. 4b. Additional parsing experts can easily be added to the system.

The Joiner uses several semantic/syntactic parsers to link concepts into the slots of the finding entity. Each parser is a context free grammar. Joiner iteratively parses the sentence, until no more changes are made to the sentence or any extracted findings. Each grammar parses the sentence in turn, adding concepts to any findings in the sentence and compressing the sentence into a simpler representation. Currently several dierent grammar experts parse sentences combining concepts and findings. This joiner phase also removes unnecessary information from the sentence and insures that negative

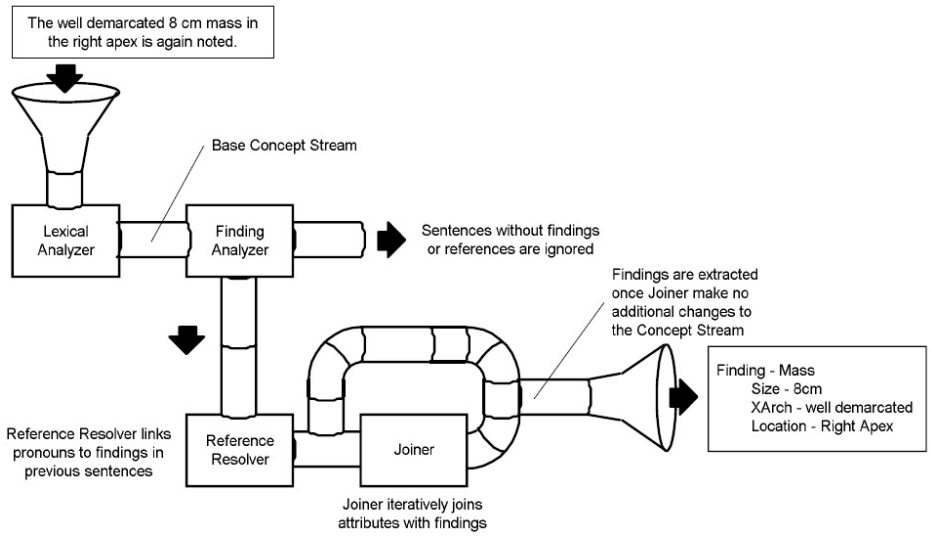

Figure 2.4: The general architecture of RADA

findings are accurately modeled. Concepts that represent a finding are combined into an entity. Within the sentence, the entity replaces the concepts it supersedes. Removing the extraneous concepts simplifies the structure of the sentence. Concept compression is the transformation of a sentence by combining several concepts into one entity. Figure 5 illustrates concept compression. Parsing experts are easier to develop and often more general because of the simpler sentence structure.

RADA resolves pronouns by retrieving the top item from the finding stack. Anaphora, such as "the mass", are resolved dierently. First, the finding analyzer replaces the anaphora with a finding and marks it as a reference. RADA continues to process the sentence, finding attributes are linked to the finding as they are found.

Once finished processing a sentence, RADA passes any findings marked as references to the reference resolver. In fourth phase, the reference resolver compares each reference to the findings stored in the finding stack. Information stored in a reference is added to a finding by matching the reference with one of the findings in the stack. A match is found if both the types and the attributes of the reference and the finding are the same. If a match is not found, the reference is considered a new finding and is added to the finding stack.

2.1.3 Statistical Natural Language Processor for Medical Reports

Taira et al [91, 92] developed statistical natural language processor for radiology reports since most tasks in NLP are classification problems. They focus on the specific sub-problems of sentence parsing and semantic interpretation [80].

The statistical NLP system consists five components: structural analyzer, lexical analyzer, parser, semantic interpreter and frame constructor as shown in Figure

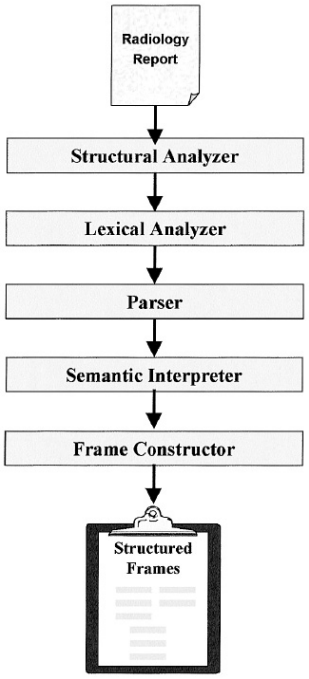

Figure 2.5: The general architecture of Taira et al's statistical medical report NLP.

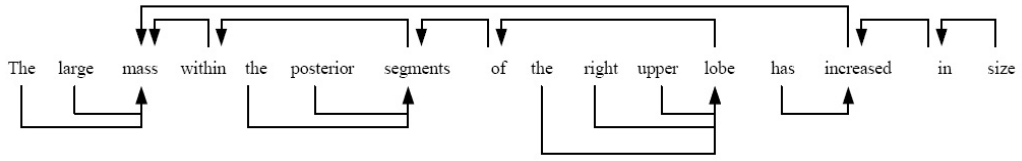

Figure 2.6: A sample sentence from radiology report with arcs showing dependencies between words.

The Structural Analyzer isolates sections of medical reports (e.g., "Procedure Description", "History", "Findings", "Impressions") and individual sentences within sections. It is implemented based on a conversion from a rule-based system to one that uses a maximum entropy classifier.

The Lexical Analyzer looks up semantic and syntactic features of words in a medical lexicon [51] as we discussed in Subsection 2.1.2, normalizes dates and numerical expressions, and tokenizes punctuation.

The Parser creates a dependency diagram between words in an input sentence by adding arcs that indicate a modifier relationship between pairs of words as shown in Figure 2.6. An arc from word A to word B indicates A modifies B. The mechanism of parsing is conceptualized as a dynamics problem similar to how atoms aggregate to form complex molecules. Words initially have no dependencies with other words. They each exist in a free state. As the parsing step proceeds, each word attempts to configure itself into a more favorable steady state of existence. The final state of the parse reflects the configuration of the words that minimizes the overall energy of the system. Words are modeled as active entities characterized by their signal processing behavior. This includes its emission spectrum, its absorption spectrum, and its response function to resonance conditions.

The Semantic Interpreter interprets the links of the parser's dependency diagram and outputs a set of logical relations that form a semantic network for the sentence. The dependency graph that the parser produces has unlabeled arcs between words to show modifier relations. The semantic inter-

Finding	=	"***mass***"
Size	=	"*large*"
Number of	=	1
Location	"*within*"	"*segments*"
Direction	=	"*posterior*"
Part of	=	"*lobe*"
Direction	=	"*right*"
Direction	=	"*upper*"
Trend	=	"*increased*"
Property	=	"*size*"

Figure 2.7: An example of structured representation output from semantic interpreter.

preter applies rules based on semantic features of the words and the direction of the arc between them in the surface structure parse to translate these arcs into the logical relations. Then it bundles logical relations together into output frames that list attributes of a finding, of a therapeutic or diagnostic procedure, or of an anatomic structure. An example of structured representation is shown in Figure

The Discourse Processor determines whether a finding from a sentence is new or a referent to a finding from previous sentences.

2.1.4 Challenges

There are many other systems, such as SAPHIRE [42] and RIME [7, 89] implemented to extract information and structure from semi-structured or free text medical reports or classify the radiology reports [3]. The major challenges for these systems are: negations [27, 12, 83, 3], insufficient understanding of the text [92], ungrammatical writing styles [92], large vocabulary [92] and assumed knowledge between writer and reader [92].

2.2 Automatic Generation of Medical Reports

Currently there is not much research work done in automatically generation of medical reports; however, some works have been done in more general domain of medicine as surveyed by D. Hüske-Kraus in [50].

Current ways of producing text for documents in medicine are less than optimal in several respects. The field of NLG draws on the idea of generating text from a conceptual representation of not only certain facts, but also knowledge about how to express them via (written) language. Unfortunately, NLG does not yet offer ready-to-run solutions for the automatic production of most of the document types in the given typology. It seems, however, highly plausible that the demands of medical informatics for these kinds of systems will be satisfiable as NLG matures.

2.2.1 Structured Data Entry Approach

In contrast to free text documents, Structured Data Entry (SDE) as introduced and described in [16, 33, 58, 65, 73, 96] is to provide the physician with predefined forms, containing the typical graphical user interface (GUI) widgets and to build a structured report from the data entered.

This approach offers guideline for less experienced physicians to complete their observations and facts gives a uniform structure for the medical documents. It also tolerates telegraphic writing style, and is therefore sufficient for a lot of systems [20, 44, 57, 97].

However, there are also drawbacks of the SDE approach. Stylistic variance is not possible. The rigid one-to-one mapping of a user-entered data item to a phrase in the report not only yields less than natural text, but also leaves no place for the stylistic preferences of different physicians. What may be a desirable quality in one context, ensuring standard terminology, can be a serious obstacle to broad user acceptance under circumstances where physicians do not want their individual formulations to become streamlined by an

information and communications technology (ICT) system. The (temporal) decoupling of data entry and report generation which many of the systems exhibit [58, 97] is a step back from a "what you see is what you get" functionality. Especially in the early phases of system use, it can be irritating for a physician not to have an immediate response regarding what his action in the GUI. will result in in the output text. Naturally enough, the user will have less confidence in the report being a faithful account of his entries the longer it takes until he has the possibility of reviewing it.

2.2.2 Mail Merge Approach

Mail Merge approach is typically found in contexts where a number of data items are stored in a database, such as laboratory systems, patient data management systems (PDMS) in intensive care units, but also in conjunction with SDE forms, the technology of mail merge features can be used to build up a text. With procedural constructs such as "IF ... THEN ... ELSE", "CASE", "LOOP" etc. it is possible to put together quite complicated text structures as in [2]. Nevertheless, an enormous amount of coding has to be carried out just to get the syntactical structure right, at least in languages where there is a high degree of flexion in nouns, pronouns, adjectives and verbs. Even though more would be possible in theory, in reality these systems tend to adopt an only slightly less telegraphic style and are thus subject to the same criticisms as the SDE or "concatenated item labels" approach.

2.2.3 Canned Text Approach

Whereas the SDE and Mail Merge approaches require a system analysis/design phase before actual use and in most cases are configurable only by the ICT vendor, the implementation of canned text phrases is very easy, and they are a commodity item in virtually every text processing environment. The user has total control over the inventory of text phrases. This is, of course, a blessing

in disguise, as the full responsibility for completeness, stylistic adequacy and syntactical congruence of the phrases is imposed on the system user. Thus, in many cases only normal findings are represented as canned text phrases. Wherever complicated sentence structures must be built from several variable text fragments, it is more likely than not that, at least in languages with much inflexion, the number of phrases suffers from combinatorial explosion. This again can be avoided only by adopting an agrammatical "telegraphic" style as in [94]. It is worth noting that, when this is done, the canned text approach has most of the drawbacks of the other two approaches, but none of their advantages, rendering it the least favorable alternative.

Of course a lot of valuable work has been done on the basis of the techniques mentioned above. Recognizing the need for flexible and open, self-documenting and communicable reports, Kahn [53] for instance has dedicated much effort to the task of platform-independent reporting. But irrespective of the accomplishments in medical linguistics, exemplified for instance in the Unified Medical Language System (UMLS) [68, 69], the technology for generating the appropriate text from these reports is still lacking.

2.2.4 Hybrid Approach

Some medical report generation systems not only employ one of the approaches mentioned in previous subsections, but a combination of a few. Linguistic String Parser Project described in [61] or the BAIK system as in [32] used the "canned text mail merge" technique.

2.2.5 NLG Approach

We regard any system which technologically goes beyond the approach of filling strings into a template to generate text from an underlying non-linguistic representation as an NLG system. A hypothetical "universal" NLG system that differs from the systems mentioned in previous subsections would per-

form the following tasks: discourse planning, content determination, document structuring, multilingual microplanning with lexicalization, aggregation, generation of referential expressions, and formatting.

NLG systems in general are more mature than systems employing other earlier approaches and achieve more functionality. Researchers as in [70, 56] addressed the problem of generating medical documents in an ad-hoc, though not necessarily unsuccessful way, whereas others realized the need to draw on an underlying conceptual representation. Bernauer [5, 6] utilized conceptual graphs as a representational formalism from which reports (of bone scans) were generated. At approximately the same time, A. Rector, B. Nowlan et al. addressed the multilinguality issue with PEN&PAD as in [77], and in 1994 Bullock [9] used the semantic net underlying PEN&PAD as input for a text generation component.

It is worth recognizing that whilst the problem of data entry, i.e. achieving a situation in which physicians willingly enter data on their own, was not solved satisfactorily in all of these projects, their failure to come into widespread routine use must not be attributed solely to the drawbacks of the approaches mentioned. The low quality of the resulting texts, especially with respect to their diversity, appropriateness and readability, in conjunction with their limited to non-existent potential to "tune" their generation component to individual needs or the progress of clinical medicine, must be regarded as being of at least equal weight.

The fact that text quality really is an issue may be seen from the fact that although structured data entry (sometimes even based on conceptual representations) is becoming more and more common in medicine [8, 36, 90] there is still no system in routine use which for a nontrivial domain based on NLG principles generates medical reports with fluent, concise and readable text.

2.2.6 Challenges

It is true that NLG does not address all the problems surrounding the intended application of medical document generation, but this is of course also due to the neglect of the medical informatics community. Moreover, NLG still has a long way to go before it can be expected to offer easy and convenient solutions to every generation problem, so that the very idea of employing NLG to generate medical reports may be premature. On the other hand, the body of concepts and techniques already elaborated under the heading of NLG leads the way to far more elaborate systems for generating text in clinical medicine.

Challenges of NLG in general domain also exist in medical domain, more specifically, in our topic of generating medical radiology reports. At micro planning level, problems include aggregation, anaphora, referential expressions, ellipsis and enumerations. At discourse level, problems include discourse planning, content determination, document structuring, morphosyntactical realization, and formatting.

2.3 Free Text Assisted Medical Image Retrieval

Content based image retrieval (CBIR) in medical domain is of growing interest in recent years [74]. As large amounts of medical images are produced in hospitals every day, applications developed upon content based medical image retrieval are very useful. However, many of these medical CBIR systems do not make full use of the information that associated with the images. DICOM headers, which contain information about the bio-data of the patients, specifications of the operation procedures and device settings and etc., are used by some of the systems; whereas the radiology reports associated with the medical are much ignored. Using both image and text information,

the performance of information retrieval has large potential to improve, and systems of more functionalities can be developed.

2.3.1 NeuRadIR: Web-Based Neuroradiological Information Retrieval System

As very few current work has been focusing on CT brain image retrieval with associated radiology reports, our literature review in this specific aspect is harder than others. Nevertheless, we found NeuRadIR as in [23, 24], an neuroradiological information retrieval system with focus on human brain CT based on radiology reports written in Hungarian language.

Most current image indexing and retrieval systems make use of general visual properties such as color, shape and texture to classify or retrieve 2D images. Accordingly, in medical research some content-based information retrieval systems are developed based on image features. However, they chose to adopt a semantic information approach which uses radiology reports for indexing to retrieve brain CT images, because in medical radiology content-based image retrieval cannot be carried out by using completely automated approaches [86], as the clinically useful information in an image typically consists of grey level variations, and the image sets differ only through subtle, domain specific clues [37, 38, 41, 63].

They identified and indexed relevant medical concepts using Johnson et al's approach [51] as we described in Subsection 2.1.2. Three different information retrieval methods which could be used in combination were implemented in the NeuRadIR system to satisfy the needs of different user groups: retrieval of exact matches, retrieval of similar objects (partial match), and retrieval of associated objects. The general architecture of the NeuRadIR system is shown in Figure 2.8.

Retrieval of exact matches is a boolean method which allows to retrieve objects that match exactly the user's query. An exact match is primarily required in undergraduate teaching for demonstration and analysis purposes,

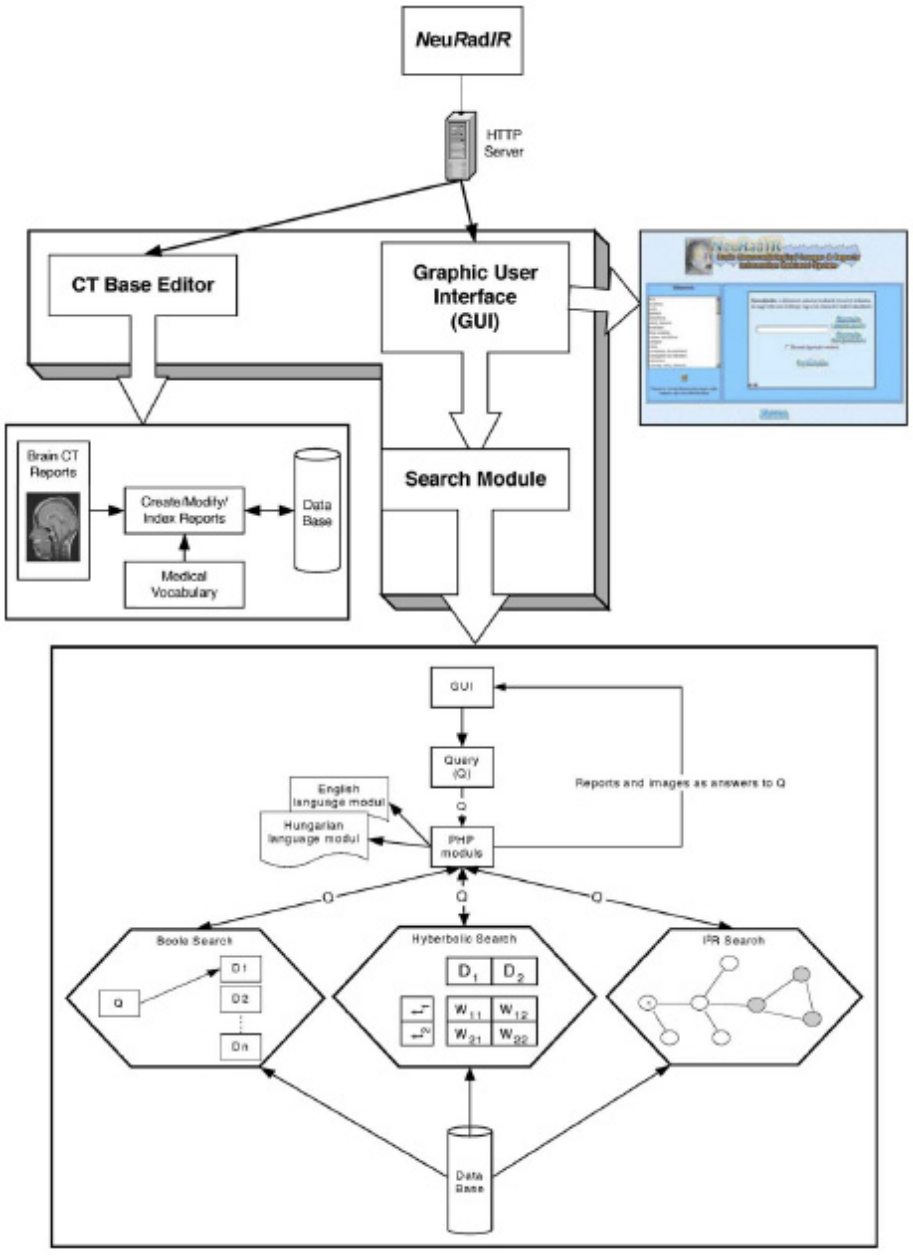

Figure 2.8: The general architecture of the NeuRadIR neuroradiological information retrieval system

in the activity of a general practitioner when monitoring the patient's medical history, or by the neuroradiologist in support of the diagnoses. The boolean model (BM) of information retrieval is, historically, the first retrieval method developed for the purpose to retrieve information from large computer databases. The BM is used by virtually all commercial IR systems today. The BM is based on Boolean logic and classical set theory. The objects (e.g., documents, images) to be searched are conceived as sets of terms, while the users query as a Boolean expression of terms. Retrieval is based on whether an object contains a query term; more exactly, those objects are retrieved that satisfy, in terms of Boolean logic, the query. The BM can be used when one wants to retrieve objects that match exactly the condition expressed in the query.

Retrieval of similar objects (partial match), a hyperbolic method [35], makes it possible to retrieve similar objects in response to a query, ranked by their degree of similarity. A partial match is primarily required by the neuroradiologist to help make a clearer decision, and in postgraduate courses (or undergraduate teaching) for analysis purposes. In the hyperbolic information retrieval (HIR) model, both the objects and the query are mapped to points of a feature space based on frequencies of terms (or keywords) that characterize them. In the HIR, the feature space is mathematically modeled by a non-Euclidean Geometry called a CayleyCKlein Hyperbolic Space (CCKHS) in which the query is at the origin of the space. Retrieval is based on how "similar" the query and the objects are to each other. The degree of similarity is evaluated based on the geometric distance (hyperbolic distance) between points. Those objects are retrieved which are similar enough, i.e., which are closer to one another than a predefined distance. Notice that, in HIR, the query does not have the form of a Boolean expression (like in the BM). The HIR model can be used when one wants to retrieve similar objects in response to a query, ranked by their degree of similarity.

Retrieval of associated objects is an interaction method [22]. This method

allows for retrieving objects that are associated with the query, even if they do not contain any of the query terms. Associated match is primarily required in medical research, or by the neuroradiologist to help discover relationships that may support decision making. While in both the BM and the HIR methods the objects to be searched are considered to be entities that are isolated from each other, in the interaction information retrieval (I2R) method the objects form an interconnected network. Before being answered, the users query is interconnected with the objects, thus it restructures, partly, the initial network of objects. The retrieval process means a spreading of activation started at the query, and those objects are being returned that form a reverberative circle ("recalled memory") which emerges during this process. The I2R method is used when one wants to retrieve objects that are associated with the query, even if they do not contain any of the query terms.

According to their system and real user evaluation, the relevance effectiveness (recall) of the NeuRadIR system was found to be such that it fulfils its aim set by their project.

2.3.2 Information Retrieval on MR Brain Images and Radiology Reports

Compared to the work for brain CT, more research has been done for MR brain image content based retrieval in association with free text radiology reports as in [87]. They developed a method that combines relevant structured information derived from free-text radiology reports by a natural language processor (NLP) with an automated registration algorithm that maps patient images to a labeled brain atlas.

The method consists of four phases: automated image volume registration to a labeled digital atlas, identification and refinement of organ-specific contours, computation of quantitative imaging features such as texture, heterogeneity, shape, and size, and automatic extraction and structuring of dictated findings from the corresponding radiology report. The general architecture is

shown in Figure 2.9. The particulars of registration, segmentation, and feature extraction algorithms are not specified because the general architecture does not depend on these details. The choice of a specific algorithm depends on various factors including imaging modality, anatomy, and patient demographics. Algorithms (and associated parameters) can either be selected by the user or set by the system (based on predefined expert rules). It is also possible to incorporate machine learning techniques for automated selection of different image-processing algorithms.

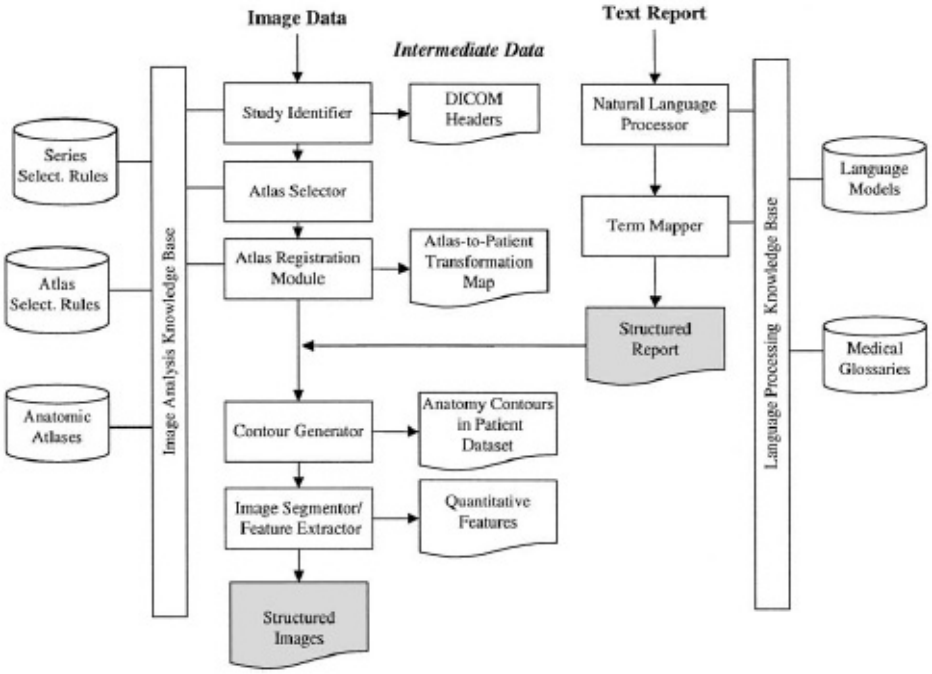

Figure 2.9: The general architecture of the MR brain image content extraction module

The Study Identifier extracts information from the DICOM headers in the studies, including patient demographics, imaging modality, anatomy, reason for study, image geometry, and acquisition parameters. The Atlas Selector selects the most closely matched anatomic atlas based on patient age, disease condition, and imaging modality. Registration Selector chooses proper regis-

tration algorithm from a number of 2D and 3D registration algorithms that require little or no manual intervention. The Contour Generator performs registration using the algorithm (and parameters) selected by the Registration Selector and also identifies the images associated with the location descriptions of findings extracted from the radiology report. The Natural Language Processor (NLP) takes radiology reports as inputs, and then outputs a set of structured frames which contain the important attribute descriptions reported for a particular finding. The NLP they used is based on Taira et al's approach [91, 92] as we discussed in Subsection 2.1.3. The Term Mapper takes either the set of location-specific logical relations from the NLP or anatomic terms entered from a structured data entry module as input. Then the Term Mapper accesses a dictionary that defines a list of synonyms for each atlas term. This synonym dictionary represents an attempt to standardize the anatomic descriptions found in free-text reports. Alternatively, all structures listed in the atlas could be coded to a standardized terminology such as SNOMED [88] or UMLS [49]. The Image Segmenter then further segments the image generated from Contour Generator. The Feature Extractor calculates quantitative features for each finding segmented by the Image Segmenter. These features include descriptors of shape, size, border characteristics, and homogeneity, which can then serve as more precise image content indexes complementing the traditional qualitative descriptions found in radiology reports.

2.3.3 Challenges

The major challenge for this area of research is the complexity of the system. Such systems contain many functional components that needs specialities in various research area. Therefore, it is difficult for any individual researcher to build such system or for a group of researchers to build such system in a short period of time. For each component, the choice of approach or algorithm is also important, as different systems cater for different purposes and thus

should adapt the approach that best serves its own purpose.

Chapter 3

Possible Research Topics

Based on the literature review in Chapter 2, our possible research topics will be in the corresponding aspects in a more specific domain of brain CT: information extraction from brain CT free text radiology reports, automatic generation of brain CT reports based on domain knowledge and associated images, and radiology reports assisted brain CT images retrieval. We will also investigate other suitable and related research areas if possible.

3.1 Information Extraction from Brain CT Radiology Reports

Our major task in this research aspect is to extract structured medical findings from the free text radiology reports. As shown in Figure 3.1, our system will have four components: structure analyzer, lexical analyzer, parser, and composer.

Most of our radiology reports obtained from NNI consist of three parts: reasons for examination, detailed description of observations and findings, and comments or conclusion. The structure analyzer will decompose the radiology report into these three sections. As the writing style in our radiology

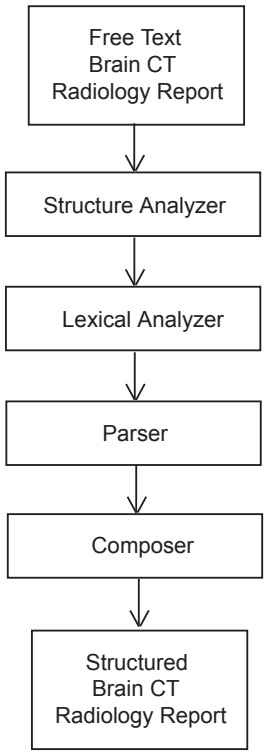

Figure 3.1: The general architecture of our brain CT radiology report information extraction system

reports are consistent, we will first implement the structure analyzer based on rules. We will use other approaches (e.g. statistical approach) at later stage in order to adapt writing style variations.

The lexical analyzer takes a single sentence as the input, and outputs the semantic and syntactic features of each word or phrase in the sentence. In this labeling process, punctuation is identified; dates, numeric measurements, special symbols, and proper nouns are recognized; words and phrases are looked up in medical lexicons; and prefixes and suffixes are analyzed. Any words that remain unknown after this process are inserted into a database of unrecognized words together with their frequency of occurrence. We will use both UMLS and published radiology sources (e.g. radiology textbooks, radiology review manuals, radiology word compilations, and published radiology glossaries) as our medical lexicon.

The parser uses a grammar and lexical definitions to identify and interpret the structure of the sentence, and to generate an intermediate structure based on grammar specifications. The grammar is a set of rules based on semantic and syntactic co-occurrence patterns. The output of the parsing phase generates a list structure, where the output consists of primary findings and associated modifiers. The detailed implementation of parser is yet to be decided in the future studies.

The composer then reduces redundant structures obtained from parser, combines the structures and outputs the final overall structure of the radiology report. The detailed implementation of composer is also to be decided in the future studies.

3.2 Automatic Generation of Brain CT Radiology Reports

We will make use of current NLG techniques and previously analyzed and indexed brain CT radiology reports and images to automatically generate

radiology reports for new brain CT scans. This part of research will be done in collaboration with the other fellow graduate student Liu Ruizhe, who will be in charge of the image part.

Shown in Figure 3.2, our general architecture will follow the NLG architecture described in [52, 84]. Our approach will be corpus based, which typically uses a collection of example inputs and associated output texts. The example inputs in our case are the extracted features from the analyzed brain CT images and radiology reports in the database; while the output are the associated human-authored free text radiology reports.

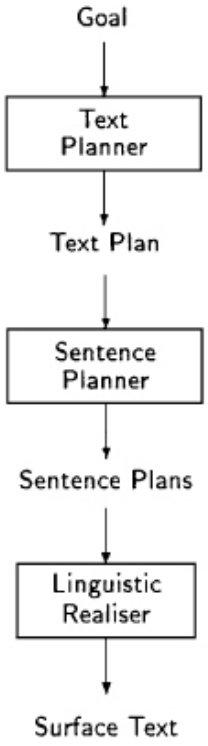

Figure 3.2: The general architecture of a natural language generation system.

Our system will contain the basic parts of a typical NLG system: content determination, discourse planning, sentence aggregation, lexicalization, referring expression generation, and linguistic realization.

Content determination is the process of deciding what information should be communicated in the text. This process is to create a set of messages from the system's inputs or underlying data sources, namely the extracted features from the new brain CT images in our case. These messages are the data objects then used by the subsequent language generation processes.

Discourse planning is the process of imposing ordering and structure over the set of messages to be conveyed, as good structuring can make a text much easier to read. In the simplest possible terms, this is akin to a story having a beginning, a middle and an end; but most documents have much more discernible structure than this. In our case of brain CT radiology report, according to the usual writing style of most radiologists, the beginning is often the reason for CT scanning (which can be extracted from the DICOM header associated with the brain CT image), the middle part is the detailed description of medical observations and findings, and the end is the conclusion or summary of the findings. We will use Rhetorical Structure Theory (RST) [67] to organize the text based on relationships that hold between parts of the text. An example of a simple radiology report in rhetorical structure is shown in Figure 3.3.

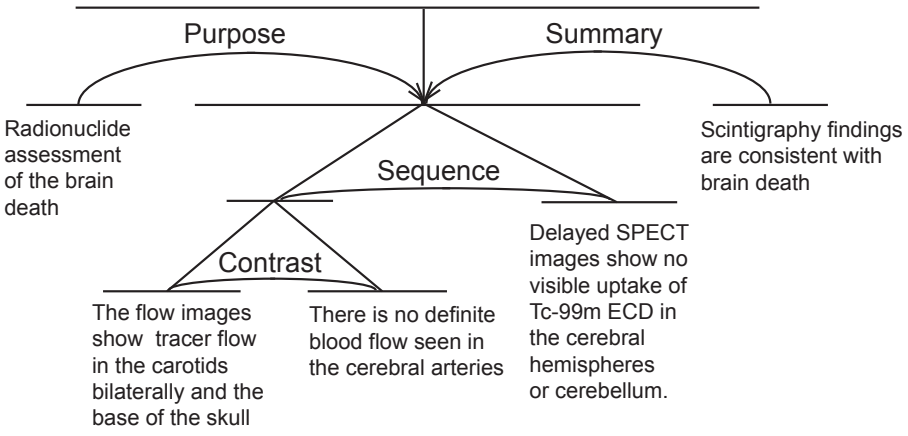

Figure 3.3: The rhetorical structure of a simple radiology report of brain CT scan.

Sentence aggregation is the process of grouping messages together into sentences. In principle, aggregation techniques can be used to form paragraphs and other higher-order structures as well as sentences; however, this is a less well understood process, and is one of the most challenging part.

Lexicalization is the process of deciding which specific words and phrases should be chosen to express the domain concepts and relations which appear in the messages. In many cases, lexicalization can be done trivially by hard-coding a specific word or phrase for each domain concept or relation; however, in some other cases, fluency can be improved by allowing the NLG system to vary the words used to express a concept or relation, either to achieve variety or to accommodate subtle pragmatic distinctions. In our case of brain CT radiology report, we will first employ the former approach which hard-codes specific word and phrases to standardize the output language radiology reporting [78, 81]. At later stages, we will employ the latter approach to generate radiology reports of various writing styles to cater different user groups.

Referring expression generation is the task of selecting words or phrases to identify domain entities. Referring expression generation is closely related to lexicalization, since it is also concerned with producing surface linguistic forms which identify domain elements. However, unlike lexicalization, referring expression generation is usually formalized as a discrimination task, where the system needs to communicate sufficient information to distinguish one domain entity from other domain entities. This generally requires taking account of contextual factors, including in particular the content of previous communications with the user (generally referred to as the discourse history). Referring expression generation also remains as one of the most challenging subtask in NLG. We will study this area in more depth and search for a practical solution for our case in the future.

Linguistic realization is the process of applying the rules of grammar to produce a text which is syntactically, morphologically, and orthographically

correct. We will also study this part in more detail in the future.

Besides the challenges for general medical report generation as we discussed in Subsection 2.2.6, for our project, the detail level of the features extracted from brain CT images poses another challenge[1] as our automatic generation of radiology report will be based on these extracted features as well.

3.3 Radiology Reports Assisted Brain CT Images Retrieval

NNI (National Neuroscience Institute, where our research data are from) currently use GE PACS system to store the patient data, which consist of one or more series of scans and radiology reports corresponding to each series of scan. The patient data are indexed using only name and id number, which is very inconvenient as we discussed in Chapter 1. Our system will use indexed information from radiology reports and images other than patient name and id for users to search.

As shown in Figure 3.4, we will extract information from both images[2] and reports as discussed in Subsection 3.1. For each series of examination or scanning, we will have a set of features or findings extracted from both the brain CT images and the associated radiology report.

The general architecture of our brain CT image and radiology report system is shown in Figure 3.5. The system can take three types of queries: a short text string (just like text search in most cases, e.g. "extradural hemorrhage, no fracture"), a sample brain CT radiology report, and a sample brain CT scan image. The system can take any one of the three types or all types as input query.

[1]This challenge was reminded by Robert Dale during a consultation when he made a academic visit to National University of Singapore in August 2007.

[2]This part will be done by Liu Ruizhe.

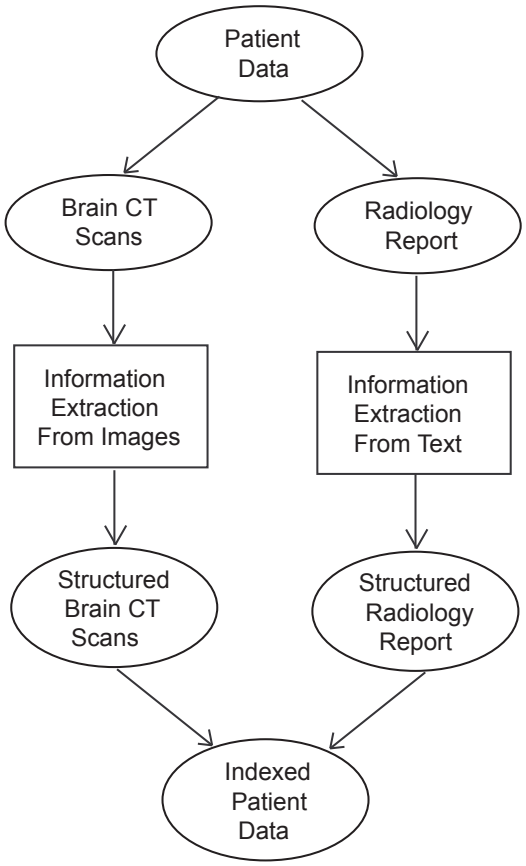

Figure 3.4: Our information extraction system (from both brain CT images and radiology reports).

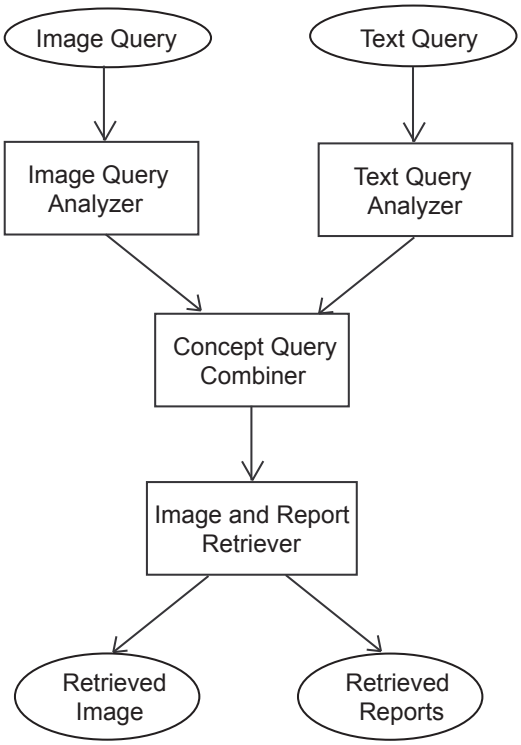

Figure 3.5: The general architecture of our brain CT image and radiology report system.

For image query, we will first extract information from the input image[3]; and for both types of text query, we will use our system discussed in Subsection 3.1 to extract concepts from the queries.

The concept query combiner will reduce the redundant concepts and resolve conflicting concepts generated from both analyzers, and then combines the concepts into one query. The detailed implementation is yet to be decided in future studies.

Then the image and report retriever will use the concepts generated from concept query combiner to retrieve images and reports accordingly from the database. The detailed implementation is also to be decided in future studies.

3.4 Others

We may also investigate some other interesting aspects of research for our proposed work, including question answering system for brain CT image and report database. Our possible research topic may also include these aspects as well in the future.

[3]This part will be done by Liu Ruizhe.

Chapter 4

Preliminary Work

In this chapter, we describe some preliminary work that has been done in collaboration with the other fellow graduate student Liu Ruizhe on automatic classification of trauma types based on brain CT scans. The detailed method and experimental results are published in 2nd IAPR International Workshop on Pattern Recognition in Bioinformatics (PRIB 2007) [34]; nevertheless, we will have a brief description in this paper.

4.1 Problem Definition

The input of the problem is a set of brain CT scan images in JPEG format[1] The system is to classify these images into three major head trauma categories or normal category [60]: EDH[2], SDH_Acute[3], ICH[4] or normal.

[1]The most common format for medical images is DICOM [54, 75] in recent years; however, we need the images in JPEG format in order to process them.

[2]Extradural hemorrhage, aka. Epidural hemorrhage.

[3]Acute subdural hemorrhage.

[4]Intracerebral hemorrhage.

4.2 Our Approach

In our proposed approach, the system consists of three phases: preprocess[5], feature extraction and classification. In the preprocessing phase, we segment the hemorrhage regions from the CT brain image using ellipse fitting [25], background removal and wavelet decomposition technique [19, 72]. The segmented result is a binary image with potential hemorrhage regions in white and the others in black. Then in phase two, for each of the potential hemorrhage regions, we extract information about size, shape and position, and create a feature vector accordingly. Lastly in phase three, we use a machine learning algorithm to classify the potential hemorrhage regions into different hemorrhage types or normal regions according to the extracted features. The CT brain images are then classified according to the classification of its potential hemorrhage regions.

4.2.1 Preprocessing

The preprocessing algorithm consists of four steps. Step one takes figure 4.1 as input, and removes the skull and fits an ellipse to the skull to construct an "interior region", the brain inside the skull. Step two removes the gray areas in the "interior region", which consists of mainly grey matter and white matter. Step three then uses a wavelet decomposition to reduce noise and set a threshold automatically to identify the hemorrhage regions. The last step generates a binary image containing the hemorrhage regions in white whereas the others in black as shown in Figure 4.2.

4.2.2 Feature Extraction

The brain CT image features that the human doctors use for classification include the size, the shape and the relative position of the potential hem-

[5]The preprocessing phase was done by Liu Ruizhe.

orrhage region. Hence, we also need the similar features for our automatic classification. In our approach, we find out and quantify the such features using Matlab Image Processing Toolbox [1]. For each potential hemorrhage region, we use the Matlab embedded function *regionprops* to extract the area, major and minor axis lengths, eccentricity, solidity and extent of the potential hemorrhage regions. Features for the skull and the background are extracted from the labeled skull and background regions as well [18]. As these features are similar to the features doctors manually use in such a way that they also describe the size, shape and position of the potential hemorrhage region; therefore, we consider them useful for classification. The class of each feature vector is one of the following values: EDH, SDH_Acute, ICH and normal. The details and implicit meaning of each feature are shown in Table 4.1.

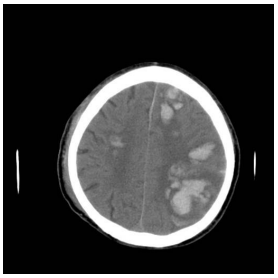

Figure 4.1: A sample ICH image taken as input.

Figure 4.2: The output image of the ICH sample after preprocessing.

	Name	Description [1]
1	Area	The actual number of pixels in the region.
2	Major axis length	The length (in pixels) of the major axis of the ellipse that has the same second-moments as the region.
3	Minor axis length	The length (in pixels) of the minor axis of the ellipse that has the same second-moments as the region.
4	Eccentricity	The eccentricity of the ellipse that has the same second- moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length.
5	Solidity	The proportion of the pixels in the convex hull that are also in the region.
6	Extent	The proportion of the pixels in the bounding box that are also in the region. Computed as the area divided by area of the bounding box.
7	Skull	Whether the region is adjacent to skull or not.
8	Background	Whether the region is adjacent to background or not.

Table 4.1: Features extracted from each possible hemorrhage region for classification.

4.2.3 Classification

According to the features extracted in section 4.2.2, we classify the regions into five categories: EDH, SDH_Acute, ICH, other and normal, where the first three classes refer to the three types of hemorrhages we focus on, and normal means that the region is not a hemorrhage. For example, the potential hemorrhage regions of Figure 5 classified as ICH are shown in Figure 6.

As there may be more than one type of hemorrhage present in a brain CT image, the class for each image cannot have only one of the class values as the regions have. Instead, the class for each image is a boolean vector <EDH, SDH_Acute, ICH, normal>, where each boolean value indicates the presence of certain type of hemorrhage. The class of the image is classified according the classifications of its regions. If the regions are classified as some type(s) of hemorrhage (EDH, SDH_Acute, or ICH), the image is also classified as the same type(s) of the hemorrhage(s). Otherwise if all regions are classified as normal, the image itself is also classified as normal.

4.3 Experiment Results

We obtained 35 CT brain images (15 EDH, 9 SDH_Acute, 6 ICH and 5 normal) belonging to 12 patients from the National Neuroscience Institute, Tan Tock Seng Hospital, Singapore. After preprocessing, we obtained 818 potential hemorrhage regions (15 EDH, 19 SDH_Acute, 47 ICH and 737 normal).

	EDH	SDH_Acute	ICH	normal
Precision	60.0%	53.8%	60.0%	95.9%
Recall	60.0%	36.8%	44.7%	98.2%

Table 4.2: Detailed testing results for each class.

We used J48 classifier, a decision tree classifier based on C4.5 [79], from WEKA [45] to train and test the region features. 10-fold cross validation

was used. The average accuracy (correctly classified regions / all regions) is 93.0%. As there are many more normal class cases than the other classes, the data is highly imbalanced, which causes high accuracy for normal class and relatively lower accuracy for other classes. The detailed testing results for each class are reported as shown in Table 4.2.

The decision tree obtained from J48 is shown in Figure 4.3. The knowledge represented by the decision tree is actually very close to the doctors knowledge in classifying potential hemorrhage regions. For example, if the regions area is less than or equal to 2891 pixels (6.89cm2) and greater than 91 pixels (0.22cm2), and the eccentricity is less than or equal to 0.9426 (the greater the eccentricity is, the elongated is the region), and the region is not adjacent to skull, then the region is ICH. This is also a typical rule for doctors to recognize ICH manually.

The classification of the image is considered as: 1. correct, if the predicted class(es) and the actual class(es) are exactly the same; 2. partially correct, if the actual class(es) is/are included in the prediction, but other class(es) is/are also predicted; 3. incorrect, if the predicted class(es) is different from the actual class. Among the 35 images, 18 are classified correctly, 6 are classified partially correctly, and 11 are classified incorrectly.

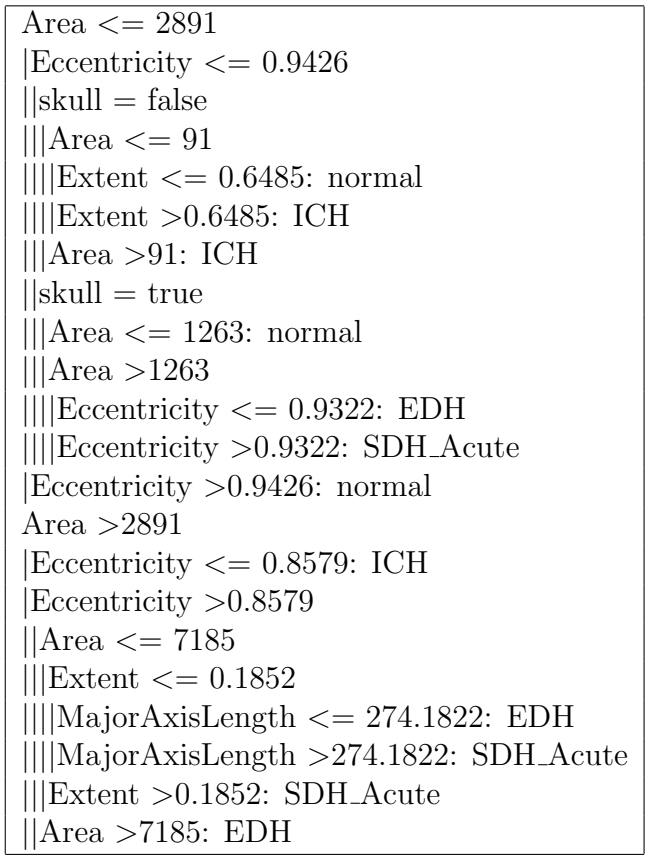

Figure 4.3: The decision tree obtained from the training data using the J48 classifier.

Chapter 5

Conclusion

With the advances of medical techniques, large amounts of medical data are produced in hospitals every day. Corresponding computer applications are also developed to analyze the data in various aspects. Among the areas, data mining in medical images is of growing interest. However, as most current works are focusing on mining the medical images (including content based medical image retrieval, medical image indexing, and etc.), fewer work has been done for the text information associated with the medical images.

Radiology reports contain rich information about the corresponding medical images. They usually contain detailed description of normal and abnormal findings of the interested body areas on the images, conclusion from the radiologists, and sometimes some bio data of the patients as well. As the radiology reports are often under mined, the valuable information contained in them are not utilized to build new computer applications or improve the existing applications.

We obtained around 300 patients' data on brain CT scans from National Neuroscience Institute (NNI), Tan Tock Seng Hospital, Singapore. The number will grow to around 2000 in the near future as the hospital has already contained this amount of data on patients but not yet been extracted for our research use. The data for each patient include one series of CT images

on the brain (18 to 30 slices or images) and one free text radiology report associated with the series of image.

Our research topics will be focusing on information extraction from brain CT radiology reports, radiology reports assisted medical image content retrieval, and automatic generation of brain CT reports based on domain knowledge and associated images. We have surveyed related works in these research areas, and have design general architectures of our systems for our possible research topics.

Most of the current medical image and report system index only on patients' particulars such as name and identity card number. Such system is inefficient and inconvenient as it is impossible for doctors or radiologists to remember each patient's name and id in order to retrieve their medical data. However, if the radiology reports are processed, mined and indexed, the medical record systems can be improved and become more efficient and convenient.

When such improvements are done, the doctors and radiologists can also be more efficient to do their research in the area. Instead of trying hard to gather text or image information based on patients' particulars from memory fragment (which is extremely time and effort consuming and often unsuccessful), they can easily search the improved medical record systems based on medical terms (such as disease, treatment, etc.), gather the information what they wanted much faster, and conduct research and studies on them more conveniently.

Moreover, the automatical generation of reports can give reference to radiologists for them to compare the generated results and their judgement in hospital.

Our research can also help to facilitate an education system for junior doctors and researchers in the area. Online medical record system open for doctors, medical students and researchers can be built based on our proposed research. Such online system can integrate more medical record data

from various source and provide a platform for the community to exchange information and knowledge.

Bibliography

[1] *. *Image processing toolbox user's guide version 3*. The MathWorks. 2002.

[2] Parametrix Solutions AG. *Phoenix*. 2001. www.phoenix.ch

[3] David B. Aronow, Fangfang Feng, and W. Bruce Croft. *Ad hoc classification of radiology reports*. J Am Med Inform Assoc. 1999;6(5):393-411.

[4] Douglas S. Bell, Edward Pattison-Gordon, and Rober A. Greenes. *Experiments in concept modeling for radiographic image reports*. J Am Med Inform Assoc. 1994;1:249-262.

[5] Bernauer J, Gumrich K, Kutz S, Linder P, Pretschner P. *An interactive report generator for bone scan studies*. Ann Symposium Computers Applied Medical Care (SCAMC 92) Proc. Clayton PD, ed. McGraw-Hill (IEEE Computer Society Press). 1992;858-860.

[6] Bernauer J. *Conceptual graphs as an operational model for descriptive findings*. SCAMC 92 Proc. Hanley & Belfus; 1992;214-217.

[7] Catherine Berrut. *Indexing medical reports: The RIME approach*. Inf Process Manage. 1990;26(1):93-109.

[8] Birkmann C, Diedrich T, Ingenerf J, Rogers J, Moser W, Engelbrecht R. *A formal model of diabetological terminology and its application for*

data entry. Med Inform Europe 97 (MIE 97) Proc. Pappas C, Maglaveras N, Scherrer J-R, eds. Amsterdam: IOS Press; 1997;426-430.

[9] Bullock JC. *Text generation from semantic network based medical records (Masters thesis)*. Manchester: University of Manchester; 1994.

[10] Keith E. Campbell, Diane E. Oliver, and Edward H. Shortliffe. *The Unified Medical Language System: toward a collaborative approach for solving terminologic problems*. J Am Med Inform Assoc 1998; 5(1):12-6.

[11] Pengyu Cao, Masao Hashiba, Kouhei Akazawa, Tomoko Yamakawa and Takayuki Matsuto. *An integrated medical image database and retrieval system using a web application server*. Int J Med Inform. 2003;71(1):51-55.

[12] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. *A simple algorithm for identifying negated findings and diseases in discharge summaries*. J Biomed Inform. 2001;34(5):301-310.

[13] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh. *Knowledge management, data mining, and txt mining in medical informatics*. In Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh. Medical informatics: knowledge management and data mining in biomedicine. Springer NY. 2005;3-33.

[14] Wesley W. Chu and Qiming Chen. *A structured approach for cooperative query answering*. IEEE Trans. on Knowledge and Data Engineering. 1994;6(5):738-749.

[15] Aaron M. Cohen, William R. Hersh *A survey of current work in biomedical text mining*. Brief Bioinform, 2005;6(1): 57-71.

[16] Colburn C. *Structured text-documentation meets technology*. J AHIMA 1997;2.

[17] Gregory F. Cooper and Randolph A. Miller. *An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text*. J Am Med Inform Assoc 1998;5(1):62-75.

[18] Dubravko Ćosić and Sven Lonŏarić. *Rule-based labeling of CT head image*. AIME Proc. 1997;1211:453-456.

[19] Ingrid Daubechies. *Ten Lectures on wavelets*. SIAM: Society for Industrial and Applied Mathematics. 1992.

[20] DeFriece RJ. *MedIO: A program for intelligent clinical data entry*. 19th Annual Symposium on Computer Applications in Medical Care (SCAMC 95) Proc. Gardner RM, ed. New Orleans, Louisiana, USA: Hanley & Belfus. 1995;1995:1005.

[21] Thierry Delbecquea, Pierre Jacquemarta and Pierre Zweigenbauma. *Indexing UMLS semantic types for medical question-answering*. Stud Health Technol Inform. 2005;116:805-10.

[22] Sándor Dominich. *Mathematical foundations of information retrieval*. Dordrecht, Boston, New York: Kluwer Academic Publishers; 2001.

[23] Sándor Dominich and Júlia Góth. *Retrieval of brain CT reports and images using interaction information retrieval*. Stud Health Technol Inform, 2002; 90: 325-9.

[24] Sándor Dominich, Júlia Góth and Tamás Kiezer. *NeuRadIR: Web-based neuroradiological information retrieval system using three methods to satisfy different user aspects*. Comput Med Imaging Graph, 2006; 30(4): 263-272.

[25] Andrew Fitzgibbon, Maurizio Pilu, and Robert B. Fisher. *Direct least square fitting of ellipses*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1999;21(5):476–480.

[26] Carol Friedman, James J. Cimino, and Stephen B. Johnson. *A conceptual model for clinical radiology reports*. Proc Annu Symp Comput Appl Med Care. 1993;:829-33.

[27] Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino and Stephen B. Johnson. *A general natural language text processor for clinical radiology*. J Am Med Inform Assoc. 1994;1(2):161-174.

[28] Carol Friedman, Stephen B. Johnson, B. Forman, and J. Starren. *Architectural requirements for a multipurpose natural language processor in the clinical environment*. Proc Annu Symp Comput Appl Med Care. 1995;:347-51.

[29] Carol Friedman, George Hripcsak.*Evaluating natural language processors in clinical domain*. Proc Conference on Natural Language and Medical Concept Representation. 1997:41–52.

[30] Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. *Automated encoding of clinical documents based on natural language processing*. J Am Med Inform Assoc. 2004;11:392–402.

[31] Carol Friedman and Stephen B. Johnson. *Natural language and text processing in biomedicine*. In Edward H. Shortliffe and James J. Cimino, eds. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, NY, 2006;312-343.

[32] Giere W. *BAIK - Befunddokumentation und Arztbriefschreibung im Krankenhaus*. Taunusstein: Media; 1986.

[33] Astrid M. van Ginneken. *Structured data entry in ORCA: the strengths of the two models combined*. 20th Symposium on Computer Applications Med Care (SCAMC 96) Proc. 1996: Hanley & Belfus, Philadelphia. 1996;1996:767-801.

[34] Tianxia Gong, Ruizhe Liu, Chew Lim Tan, Neda Farzad, Cheng Kiang Lee, Boon Chuan Pang, Qi Tian, Suisheng Tang, and Zhuo Zhang. *Classification of CT brain images of head trauma*. PRIB Proc. 2007;2007:401-408.

[35] Júlia Góth. *Hyperbolic information retrieval*. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Proc. Tampere, Finlad, August 11C15. 2002;2002:61-77.

[36] Gregory J, Mattison JE, Linde C. *Naming notes: transitions form free text to structured entry*. Methods Inf Med 1995;34(1-2):57-67.

[37] Guy C, Fftyche D. *An introduction to the principles of medical imaging*. Imperial College Press; 2000.

[38] Hajnal J, Hawkes DJ, Hill D. *Medical image registration*. CRC Press; 2001.

[39] Peter J. Haug, David L. Ranum, and Philip R. Frederick. *Computerized extraction of coded findings from free-text radiologic reports*. Radiology. 1990; 174:543-548.

[40] Heathfield HA, Kirby J, Hardiker N. *Data entry in computer-based care planning*. Medical Informatics in Europe 94 (MIE 94) Proc. 1994;1994:186-189.

[41] William Hersh. *Information retrieval: a health care perspective*. Springer Verlag; 1996.

[42] William Hersh, Mark Mailhot, Catherine Arnott-Smith, and Henry Lowe. *Selective automated indexing of findings and diagonoses in radiology reports*. J Biomed Inform. 2001;34:262-273.

[43] Lynette Hirschman and Rob Gaizauskas. *Natural language question answering: the view from here*. J Nat Lang Eng. 2001;7(4):275-300.

[44] Hohnloser JH, Engelmeier T. *Dateneingabe durch Ärzte in elektronische Patientenakten: Vergleich zweier Methoden im 5-järigen klinischen Routinebetrieb*. GMDS 96 proc. 1996.

[45] Geoffrey Holmes, Andrew Donkin, and Ian H. Witten. *WEKA: a machine learning workbench*. 2nd Australia and New Zealand Conf. on Intelligent Information Systems Proc., Brisbane, Australia. 1994:357-361.

[46] George Hripcsak, Carol Friedman, Philip O. Alderson, William DuMouchel, Stephen B. Johnson and Paul D. Clayton. *Unlocking clinical data from narrative reports: a study of natural language processing*. Ann Intern Med. 1995;122(9):681-688.

[47] George Hripcsak, Gilad J. Kuperman, Carol Friedman, and Daniel F. Heitjan. *A reliability study for evaluating information extraction from radiology reports*. J Am Med Inform Assoc. 1999; 6(2):143-150.

[48] Betsy L. Humphreys and Donald A. B. Lindberg. *The UMLS project: making the conceptual connection between users and the information they need*. Bull Med Libr Assoc. 1993 April; 81(2): 170-177.

[49] Betsy L. Humphreys, Donald A. B. Lindberg, Harold M. Schoolman, and G. Octo Barnett. *The Unified Medical Language System: an informatics research collaboration*. J Am Med Inform Assoc 1998; 5(1):1-11.

[50] Dirk Hüske-Kraus. *Text generation in Clinical Medicine-a review*. Meth Inf Med. 2003;42(1):51-60.

[51] David B. Johnson, Ricky K. Taira, Alfonso F. Cardenas and Denise R. Aberle. *Extracting information from free text radiology reports*. Int J Digit Libr 1997;1:297-308.

[52] Daniel Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall, 2000.

[53] Charles E. Kahn Jr. *A generalized language for platform independent structured reporting*. Methods Inf Med 1997;36:163-71.

[54] Charles E. Kahn Jr, John A. Carrino, Michael J. Flynn, Donald J. Peck, and Steven C. Horii. *DICOM and radiology: past, present, and future*. Journal of the American College of Radiology. 2007;4(9):652-657.

[55] Debra S. Ketchell, Leilani St. Anna, David Kauff, Barak Gaster, and Diane Timberlake. *PrimeAnswers: a practical interface for answering primary care questions*. J Am Med Inform Assoc. 2005;12(5):537-45.

[56] Kraus D, Miche E. *ArztbriefCgenerating medical reports in a multimedia environment*. Medical Informatics Europe (MIE 91) proc. Adlassnig KP, Grabner G, Bengtsson S, Hansen R, eds. 1991;Wien: Springer. 1991;900-904.

[57] Kraus D. *Suregen2Ca model based generator for surgical reports*. MIE 2000. IOS Press.

[58] Kuhn K, Zemmler T, Reichert M, Heinlein C, Roesner D. *Structured data collection and knowledge-based user guidance for abdominal ultrasound reporting*. Ann Symposium on Comput Applied Med Care Proc. 1993;1993:311-315.

[59] Caroline Lacoste, Joo-Hwee Lim, Jean-Pierre Chevallet, and Diem Thi Hoang Le. *Medical-image retrieval based on knowledge-assisted text and image indexing*. IEEE Trans. on Cir. and Sys. for Video Tech. 2007;17(7):889-900.

[60] S. Howard Lee, Krishna C.V.G. Rao, and Robert A. Zimmerman. *Cranial MRI and CT (4th Edition)*. New York: McGraw-Hill, Health Professions Division. 1999.

[61] Li P-Y, Evans M, Hier D. *Generating medical case reports with the linguistic string parser*. 5th National Conference on Artificial Intelligence (AAAI 86); August 11-15, 1986; Philadelphia, PA: Morgan Kaufmann. 1986;1986:1069-1073.

[62] Donald A. B. Lindberg, Betsy L. Humphreys, Alexa T. McCray. *The Unified Medical Language System*. Meth Inf Med. 1993;32:281-291.

[63] Yanxi Liu; Rothfus, W.E.; Kanade, T. *Content-based 3D Neurological Image Indexing and Retrieval: Preliminary Results*. IEEE International Workshop on Content-Based Access of Image and Video Database Proc. 1998;1998:91-100.

[64] Henry J. Lowe, Antipov I, William R. Hersh, Catherine Arnott-Smith, and Mark Mailhot. *Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record*. Methods Inf Med. 1999 Dec;38(4-5):303-7.

[65] Lussier YA, Maksud M, Desruisseaux B, Yale P-P, St-Arneault R. Pure MD. *A computerized patient record software for direct data entry by physicians using a keyboard-free pen-based portable computer*. Frisse ME, ed. SCAMC 92; 1992; Baltimore: McGraw-Hill; 1992:261-263.

[66] Burke W. Mamlin, Daniel T. Heinze, and Clement J. McDonald. *Automated extraction and normalization of findings from cancer-related free-text radiology reports*. AMIA Annu Symp Proc. 2003;2003:420-424.

[67] W. C. Mann and S. A. Thompson. *Rhetorical structure theory: A theory of text organization*. Technical report RS-87-190, Information Sciences Institute. 1987.

[68] McCray AT, Hole WT. *The scope and structure of the first version of the UMLS semantic net-work*. 14th Annual Symposium on Computer Applications in MedicalCare (SCAMC-90) Proc. Miller RA, ed.Washington DC: IEEE Computer Society Press. 1990;1990:126-133.

[69] McCray AT, Nelson SJ. *The representation of meaning in the UMLS*. Methods Inf Med 1995;34(1-2):193-201.

[70] McDonald CJ,Tierney WM,Blevins L. *The benefits of automated medical record systems for ambulatory care*. Implementing Health Care Inform Systems. Orthner HF, Blum BI, eds. New York: Springer. 1988.

[71] Eneida A. Mendonça, Janet Haas, Lyudmila Shagina, Elaine Larson, and Carol Friedman. *Extracting information on pneumonia in infants using natural lanuage processing of radiology reports*. J. of Biomed. Inform. 2005;38:314-321.

[72] Michel Misiti, Yves Misiti, Georges Oppenheim and Jean-Michel Poggi. *Wavelet toolbox user's guide version 4*. The MathWorks. 2007.

[73] Peter W. Moorman, Astrid M. van Ginneken, Johan van der Lei, and Jan H. van Bemmel. *A model for structured data entry based on explicit descriptional knowledge*. Methods Inf Med. 1994;33:454-463.

[74] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. *A review of content-based image retrieval systems in medical applicationsClinical benefits and future directions*. Int J Med Informat. 2004;73:1-23.

[75] Neumann M. *DICOM-current status and future developments for radiotherapy*. Zeitschrift für medizinische Physik. 2002;12(3):171-176.

[76] Torbjørn Nordgård, Martin Thorsen Ranang, and Jostein Ven. *An Approach to Automatic Text Production in Electronic Medical Record Systems*. KES Proc. 2005;2005:1187-1194.

[77] Nowlan W, Kay S, Rector AS,Horan B,Wilson A. *PEN&PAD: A multilingual patient care workstation based on a unified representation of the medical record and medical terminology.* MIE 91 Proc; 1991;Vienna,Austria: Springer. 1991;1043-8.

[78] Rita Noumeir. *Radiology interpretation process modeling.* Journal of Biomedical Informatics. 2006;39:103-114.

[79] J. Ross Quilan. *C4.5: Programs for machine learning.* Morgan Kaufmann, San Francisco, CA, 1993.

[80] A. Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Ph.D. dissertation, Dept. of Computer and Information Science, University of Pennsylvania, 1998.

[81] Bruce I. Reiner, Nancy Knight, and Eliot L. Siegel. *Radiology reporting, past, present, and future: The radiologist's perspective.* J Am Coll Radiol. 2007;4:313-319.

[82] Ehud Reiter and Robert Dale. *Building applied natural language generation sytems.* Natural Language Engineering. 1997;3(1):57-87.

[83] Roni Romano, Lior Rokach, Oded Maimon. *Automatic discovery of regular expression patterns representing negated findings in medicla narrative reprots.* 6th International Workshop on Next Generation Information Technologies and Systems (NGITS) Proc. 2006;300-311.

[84] Ehud Reiter and Robert Dale. *Building applied natural language generation systems.* J Nat Lang Eng. 1997;3(1):57-87.

[85] Sager N, Friedman C and Lyman MS. *Medical language processing: computer management of narrative data.* Reading, MA: Addison-Wesley, 1987.

[86] Shyu C-R, Brodley CE, Kak AC, Kosaka A, Aisen AM, Broderick LS. *ASSERT: a physician-in-the loop content-based retrieval system for HRCT image databases*. Comput Vis Image Understanding 1999;75(1C2):111C32.

[87] Usha Sinha, Anthony Ton, Amy Yaghmai, Ricky K. Taira and Hooshang Kangarloo. *Image Content extraction: application to MR images of the brain*. Radiographics. 2001;21(2):535-547.

[88] Spackman KA, Campbell KE, Cote RA. *SNOMED RT: a reference terminology for health care*. AMIA Fall Symp Proc 1997;1997:640C644.

[89] Peter Spyns. *Natural language processing in medicine: an overview*. Methods Inf Med. 1996;35(4-5):285–301.

[90] Tange HJ, Hasman A, Robb PFdV, Schouten HC. *Medical narratives in electronic medical records*. Int J Med Informatics 1997;46:7-29.

[91] Ricky K. Taira and Stephen G. Soderland. *A statistical natural language processor for medical reports*. AMIA Fall Symp Proc 1999;1999: 970-974.

[92] Ricky K. Taira, Stephen G. Soderland, and Rex M. Jakobovits. *Automatic structuring of radiology free-text reports*. Radiographics. 2001;21:237-245.

[93] Rafael M. Terol, Patricio Martnez-Barco, and Manuel Palomar. *A knowledge based method for the medical question answering problem*. Comput Biol Med. 2007;37(10):1511-1521.

[94] Veigel B. *OrthoStar*. 2001. www.medistar.de/sites/fachaerzte/ orthostar.html

[95] Adam B. Wilcox and George Hripcsak. *The role of domain knowledge in automating medical text report classification.* J Am Med Inform Assoc. 2003;10(4):330338.

[96] Yamazaki S, Satomura Y. *Standard method for describing an electronic patient record template: Application of XML to share domain knowledge.* Methods Inf Med. 2000;39:50-55.

[97] Zacher M. *Computergestüzte Befunddokumentation in der Arthrosonographie (M. D.).* Giessen: Justus-Liebig-Universitä Giessen; 1994.