High Accuracy Information Retrieval and Information Extraction System for Electronic Clinical Notes

Jon Patrick and Min Li
Health Information Technology Research Laboratory, School of IT
The University of Sydney
Sydney, NSW 2006, Australia

jonpat@it.usyd.edu.au mili9528@uni.sydney.edu.au

Abstract

There is a great demand for highly accurate and timely Information Retrieval and Information Extraction in medicine and health care. To meet this need, we have developed a novel system, Intelligent Clinical Notes System (ICNS) to assist doctors in retrieving clinical notes based on concept searching. This has required dealing with the both the software engineering and natural language processing aspects of the task. This system has been installed and integrated into the existing clinical information system in the Intensive Care Unit, Royal Prince Alfred Hospital, Sydney.

Keywords: Information Extraction, SNOMED CT (SCT), Text to Snomed, Intelligent Clinical Notes System (ICNS).

1 Introduction

Clinical notes is a new domain for information extraction, where health professionals have a demand for high accuracy access to information, while the demand for intelligent techniques in their information seeking tasks has become much stronger. However, a good information extraction system in health care should not only retrieve information accurately but obtain it quickly in time critical situations.

Although the advance of clinical information systems is self-evident, serious limitations in semantic retrieval in these systems are still present. The principal reason is that clinical retrieval, in which correct semantics are crucial, are quite different from the traditional keyword based search, as performed by Google. In particular, a clinical concept may be known by a large variety of different names, consequently the traditional keyword can’t retrieve all instances of the same concepts.

In this paper, we present an Intelligent Clinical Notes System (ICNS), built to Intensivists requirements that can retrieve patient notes and extract useful information from them. The main objective for this system is to use natural language processing (NLP) to serve clinicians and improve their productivity and efficiency thus contributing to patient quality and safety. For example, the concept based search engine can automatically identify synonyms for use in a search request. In addition, automatic spelling correction, as well as, most of the abbreviations and acronyms are identified and expanded, which makes the clinical notes more readable for the non-author.

During the development of this system, several external resources are used. They are a medical ontology SNOMED CT, a process for converting text into SNOMED CT terminology, gazetteers and dictionaries.

1.1 SNOMED CT

SNOMED CT (SCT) is a comprehensive medical ontology constituting a reference terminology in a relationship hierarchy with approximately 350,000 concepts and 1.4 million relationships (Wua et al., 2004). The computable concept definitions and relationships, which it provides can help determine the semantic categories in collected data. With the help of its reference terminology, clinical notes can be codified automatically, and data extraction and analysis relating to the causes of disease, the treatment of patients, and the outcomes of the overall health care process can be much more easily researched (Spackman, et al., 1997).

1.2 Text to SNOMED CT Converter (TTSCT)

In order to index all the medical terminology in the patient notes, an existing algorithm [Text To SNOMED CT (TTSCT) (Patrick, et al., 2007)] for mapping text to SCT is used in this system. The implementation of TTSCT is based on the strategy that the input is collected into “chunks” based on their meaning by utilizing natural language processing then the text tokens are matched to SCT concept tokens. Finally, a matching algorithm uses a pre-computed matching matrix to rank SCT concept descriptions against the phraseology of the chunks to identify the highest ranking match for the longest phrase. Subsequently negations and qualifications are separately recognised.

TTSCT is utilised in the ICNS by scanning through the whole patient note, identifying all the medical concepts within the free text and mapping them into concept IDs of SCT. This function then enables searches to be made on a broad range of phrases equivalent to SCT concepts rather than literal strings of descriptions.

1.3 Gazetteers

A gazetteer is a word list of one class of content, such as a list of staff names, new medical terminology, etc which is used to improve the quality of information extraction. In ICNS, we use words extracted from clinical notes consisting of terminology not in SCT to build gazetteers (such as the abbreviations gazetteer, acronym gazetteer, clinical staff gazetteer, etc), and then recognize them in the patient notes.

Copyright © 2010, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 108. Anthony Maeder and David Hansen, Eds. Reproduction for academic, not-for-profit purposes permitted provided this text is included.
Gazetteers serve a different purpose to SCT terminology. The accurate encoding of all content in the patient notes is not possible due to limitations of natural language processing functions and the failure of SCT to provide 100% coverage of clinically relevant content. The gazetteers have been trained, from a 60 million token ICU corpus, with the content that is unknown to SCT or expressed in forms that are not readily identifiable in SCT. One important advantage of the gazetteers is that they support clinicians’ shorthand expressions and idiosyncratic forms of expression not otherwise recognisable, as well as non-medical categories of information, e.g. occupations, named institutions, etc.

1.4 Dictionaries
Besides SNOMED CT and the gazetteers, several other external resources have also been adopted to identify medical terms and common words. They are:
1. Unified Medical Language System (UMLS) dictionary (Zieman and Bleich 1997).
2. Common words (Moby) dictionary.
3. Besides the above dictionaries, some other custom dictionaries have also been adopted, which were generated from the clinical notes mentioned previously. One of the largest is the abbreviation dictionary. It contains nearly 1000 abbreviations with their matching full expressions. Moreover, the misspellings dictionary (a list of misspelt words and their corrected spellings) and the unknown-words dictionary (a collection of unknown words) also have been derived from the clinical notes. The unknown words lexicon is important, as the recognition of unverifiable words is a progressive activity, which is never completed. An external process for verifying unknown words is in place but it always lags behind the generation of new unknown words, which need to be recognised in the presentation layer.

2 System Architecture and Implementation
The ICNS is based on a Client/Server/Data Warehouse configuration. All the client requests are processed at the server side. In other words, the client computer has only one main task, which is to submit the user request, such as the retrieval request, annotation request etc. Meanwhile, all the substantive tasks, such as indexing, retrieval, proof reading and so on, are processed on the server side.

Furthermore, the server is also responsible for controlling the data stores. Figure 1 shows the system configuration.

The system architecture is divided into three main components. They are the data warehousing component, server system component, and client system component.

2.1 Data Warehousing
Figure 2 shows the workflow of the data warehouse processing. This component is responsible for the data transfer, index and annotation.

The basic process is: firstly, the original notes undergo the proof reading process. Proof reading provides correction for lexical formation, spelling corrections and canonicalization of non-words such as measurements. It also adds attributes to each token such as its semantic category.

![Figure 2. Data Warehouse Processing](image)

Figure 3 presents the preprocessing required for the proof reading. The main purpose of this preprocessing is to extract four custom dictionaries from the manually appraised list of unknown words, which come from the ICU clinical corpus. The generated dictionaries will be used in the proof reading process. There are three main steps in the pre-processing:

1. The manual evaluation of unknown words is done in spreadsheets compiled to have a certain number of context examples for each unknown word. Manually, in one column the correct word is inserted and in another the gazetteer name is placed.
2. Subsequently, the columns are filtered for reuse where for example, duplicated rows and empty rows are deleted.
3. Finally, the reorganized rows are used to generate three dictionaries and approximately 50 gazetteers, which are used in the next stage.
The flow chart for the next stage of the proof reading is presented in Figure 4. Besides the unknowns lexicon, abbreviations lexicon, misspellings lexicon and gazetteers, three external resources also are used, they are the SCT (SNOMED CT) ontology, UMLS lexicon and Moby lexicon. The basic workflow for this process is:

1. The plain token is checked as to whether it belongs to the abbreviations lexicon. If not an abbreviation it will pass to the next step otherwise, this abbreviation will be expanded to the full name. Meanwhile, each word in the expanded name will go to the next step.

2. Next, if the token belongs to a gazetteer, the gazetteer class name will be added to that word’s attribute list as an output for the annotated content. Otherwise, it will be passed to the next step.

3. Next, if it belongs to the SCT ontology, an SCT tag, Concept Id and description is attached as a tag attribute. This task is achieved by a query to the SCT terminology server.

4. In the fifth and sixth steps, the word will be checked with the UMLS dictionary and Moby dictionary respectively. If it matches either it will be tagged and exported.

5. In this step, the Misspellings lexicon is used to verify if it is misspelt in which case the corrected spelling will be returned. Subsequently, every word in this corrected form will go back to the second step. This is a recursion point in this proof reading processing.

6. Finally, if the word does not belong to any of the resources mentioned above, it is checked with the Unknown Dictionary. If it exists, an unknown tag will be added to it, otherwise, it will be stored in the Unknown dictionary and given an unknown tag. The strategy of keeping an explicit lexicon of unknown words is aimed at progressively identifying and reducing the list of unknowns though manual intervention.
2.2 Server System and Client System

The other two system components are the server system and the client system. The main task for the client system is to submit client requests. These include the retrieval requests and annotation requests. In contrast, the role of the server system is to process those requests and return the results from the data warehouse. As well, the user accounts management function and auto-completion of search expressions function are also fulfilled by the server system. Figure 6 is a diagram for the Client and Server system.

There are three main boxes in the diagram, which represent the three system components. The data warehouse component has already been introduced in the previous section. In this section, the server system component and client system component are discussed.

2.2.1 Client System

Generally speaking, the client system is a user interface to submit client requests and receive the reorganized results. The client requests include the retrieval requests, annotation requests, user registration requests and login/logout requests. Five retrieval keywords are available in the retrieval request, namely patients’ MRN (medical reference number), SCT Concept, String, Care Provider Name and Storage Date. Alternatively, five types of annotation are included in the annotation request, namely, SCT Concepts, Abbreviation, Acronym, Gazetteer, and Unknown words annotation.

2.2.2 Server System

The box in the middle of Figure 6 is the server system component, which is mainly responsible for processing the client requests, managing the user accounts and organizing the results for presentation. Meanwhile, the server system is also a bridge to connect the data warehouse component and client system component.

The basic workflow for the server system is that after it receives the registration and login requests from the client system, the retrieval requests will be processed by the search expression auto-completion process.

After input of the request data, the retrieval request will be divided into three categories, such as SCT Concept, String, and Other keywords (MRN, Care Provider Name and Store Time). The reason for this classification is that the SCT Concept retrieval and String retrieval cannot be processed by an SQL query directly. For example, the SCT Concept needs to go to the TTSCT first. When the concept ID is returned from the TTSCT, an SQL query can be used to read the SCT index and return the list of notes ID which contain this SNOMED concept. The String retrieval is similar, which is processed by the string mapping and index lookup initially. After the indices are retrieved an SQL query will be used to retrieve the relevant notes from the data warehouse.

1. Once the original results are returned from the data warehouse, the user can then request annotations to the text. These requests will be processed in the Results Reorganizer. Finally, the reorganized results will be returned to the client system for screen presentation to the user.

2. The server also has the process to record every user retrieval request and annotation request in a log file. This file is stored in the data warehouse.

3 The Interface and Functionality

Four web pages are designed as the user interface in the ICNS, namely the Main Page, User Registration Page, User Login Page and Retrieval Page. In this section, the Retrieval Page will be introduced in detail, as the main functions are accessed in this page. Figure 7 is a screen shot of the retrieval page.
In Figure 7, the retrieval page is divided into three main parts. The top of the image is the retrieval specification area. This area is used to submit the client retrieval request. The next area is the annotation area, which is located in the centre of the interface (beneath the purple line with “Corrected Spelling”). It is used to submit annotation requests. The remaining part is the results area. The notes index, retrieval results and annotation results are displayed in this area.

In the retrieval area, user friendly functions are available which are used to assist clients to submit their requests. The most important is the auto-completion for search expressions. With the help of this function, users don’t need to know the exact spelling of a word (like SCT Concept) or entire number (like patient MRN). This function can predict these when users only input a part of the full spelling or digits. This function can be used in the MRN, SCT Concept, String, and Care Provider Name field.

Another user friendly function is implemented in the Storage Date field. Where users don’t need to type the storage time themselves. Rather, a calendar will pop up for users to choose the desired date. There is auto-completion for search expressions in the SCT Concept field and the pop-up calendar in the Storage Date field.

Besides these functions, the retrieval function is one of the most significant functions in the retrieval area. There are two types of retrieval methods.

The first method is searching patient notes focused on one patient. In this retrieval method, a patient MRN is input, and hence, all the notes which are retrieved only belong to that particular patient. On input a list of encounters with their time periods for the patient is populated in the encounter panel just below the MRN window. Meanwhile, an index of notes for the selected encounter period, with their storage time, is displayed in the notes index window on the left side of the retrieval panel. In the encounter panel, multiple encounter periods can be chosen according to the user’s retrieval requirements. When these operations are finished, the user can go to “step2” of the retrieval area and fill in some input fields (See Figure 7 Retrieval notes for one patient). When the above operations are completed, the search button is clicked and the results will be displayed in the results panel.

The second retrieval type is the global and it searches the complete collection of notes in the database, so the patient MRN is not required. Clients can go to “step2” of the retrieval area directly.

Another function is the notes annotation, which is used to help clinicians extract useful information from the contents in the results area. This function is used in the annotation area. In the first instance, the original content of the notes will be displayed in the results area. However, when the corrected spelling checkbox is checked, the original notes are replaced by the spelling corrected notes. See Figure 8, where all misspellings are corrected.

After spelling correction, the notes can be annotated for the different categories, such as Abbreviations, Acronyms, SCT Concepts, Gazetteers and Unknown-words. This operation is done by clicking on the checkboxes that appear under the corrected spelling checkbox, thus providing multiple annotations in the annotation area.

Figure 8 is the screen shots for Abbreviation annotation, Acronym annotation and Unknown-words annotation.

The final function for this interface is the log function. The purpose of this function is to record every user request, which includes the retrieval requests and annotation requests. This information is stored into the database and can be exported as a log file. With the help of this function, we will learn which functions are used mostly and receive feedback from the users on the functions that are and are not successful by their assessment.

4 Evaluation

Currently, comprehensive evaluation for the ICNS has not been completed. However, a small user interview was held for an initial evaluation. During the evaluation, 2 doctors, one junior and one senior, were interviewed while they were operating the ICNS. Both of them believe the ICNS brings worthwhile convenience to their daily work, since if they want to search for a concept or keyword in the patients’ notes without using the ICNS they need to read all notes for a given patient until they find the desired content. A sample demonstration given by the senior doctor was a question he had about the justification for a patient being on a certain antibiotic. He needed to find the laboratory results from previous “blood cultures” without knowing the time range of the search but excluding all notes not about the topic. The system recognised 5 notes out of 100+ notes with reference to blood cultures and automatically found the same content referenced by the abbreviation “bc” and orthographic variant “BC”.

Also, many good suggestions for improvements were given by the two doctors. The junior doctor mentioned some ideas for improving the user interface, for instance:

1. A search field for the patient name, since sometimes the doctor can only remember the patient’s name.
2. Combination of the store time in the ‘Note Index’ with other information, such as the patient name and his/her care provider for each note. This would be much more user friendly giving better aid for identifying the desired note.

On the other hand, the feedback from the senior doctor was focused on the results from the retrieval functions:

1. When we search multiple words in the string search field or SCT search field, the issue of whether the multiple words should be treated as a single phrase or separate words needs to be considered.
2. Ambiguity of abbreviations/acronyms needs to be resolved, e.g. FROM (Free Range of Movement) is also a standard preposition.

Since these interviews a special window for collecting feedback has been built into the User Interface to gather the users’ written feedback. Based on this information, we can do further investigations and research. We realize that only when the user interface is friendly enough, will the users be willing to use the new technology actively and persistently. In this way, it’s very important to know the functions that the users like and dislike. We believe with the continuous help of this feedback, a sophisticated and comprehensive system can be persistently improved to serve the clinicians well.
5 Conclusion
In this paper, a universal system for clinical information extraction is presented. The Intelligent Clinical Notes System can be easily adapted for different clinical departments. The only change for the system is to connect to the data source. Furthermore, this system is easy to change due to its three independent components. For example, besides the introduced annotation categories, many other categories can be easily added into the interface (like UMLS, Digit, etc.) without changing the Data Warehouse and Server System components, because all these categories have already been annotated in the Data Warehouse processing. Meanwhile, the criterion for annotation can be easily modified by changing the external resources (such as the Abbreviation dictionary, Gazetteers, SNOMED Dictionary, etc.) without modifying the whole system. Currently, the Intelligent Clinical Notes System is installed in the Intensive Care Unit of Royal Prince Alfred Hospital for testing purposes.

6 Future Work
Although this system has many advantages, there are still many improvements that are needed in the future. Currently, we are designing the function to allow users to revise the annotated data, since sometimes there are some mistakes in the external resources (like abbreviations, gazetteers) which will lead to inaccurate annotation in the patient notes. This function will enable processing the revised data automatically. In other words, once the corrected data is submitted by the users, the system can modify the annotation resources which it is using immediately. Finally, this system is a part of a Health Information Technologies Research Laboratory (HITRL) large project. Some of the external resources, such as the custom dictionaries and gazetteers come from HITRL’s other projects. We plan to change the system architecture so we can fetch the newest data from these projects and pass them to this system’s data resources (such as the log file and unknown words collection) to improve the services automatically.

7 Acknowledgement
We would like to thank Drs. Robert Herkes, Michael O’Lerie, and Angela Ryan and other staff in the Intensive Care Unit of Royal Prince Alfred Hospital for providing the development and evaluation environment.

8 References


Zieman, Y. L., & Bleich, H. L. (1997): Conceptual mapping of user’s queries to medical subject headings Proc AMIA
Figure 7. Retrieved notes for one patient. “SNOMED Concept” retrieves any text related to the specified concept, whereas “String” retrieves only literal strings.

Figure 8. User Interface selections for Abbreviations (pink), acronym (purple) and unknown-word (blue) annotations, auto-corrected misspellings (in grey), and dropdown list of Gazetteers.