Processing of Medical Records from Unstructured Medical Transcripts

Vinod Chandra S. S.[1]

ABSTRACT

The work looks at the problem of creating a framework that identifies information in the free text medical records and maps that information into a structured representation containing clinical terms. The content of the medical records are relatively stereotyped sentence types based on its specialized word usage. This regularity makes it possible to determine a set of sublanguage-specific word classes, which correlate with the types of information conveyed in the subfield; these word classes form the bridge between the structure (syntax) of the sublanguage text and their information content (semantics). We can define sublanguages as the part of the English language used in body of the texts dealing with a particular domain. A parser acts between the medical record and the sublanguage, which extracts the medical data from the input document

Keywords : Medical records, Medical data, Medical parser, NLP, Rule base systems

1. INTRODUCTION

Clinical medical record contains a wealth of information, largely in free-text form. Information extraction in structured format from free-text records is an important research endeavor [1]. A medical document in free text format contains information that is useful for various purposes like medical coding. But keeping track of such lengthy documents is a tedious process. This approach is to observe a large number of medical documents and find a common pattern of reporting diagnosis procedures and symptoms. The information contained in the key phrases is represented in a table like structure whose columns corresponds to the major sublanguage word classes. Different column combinations are possible in a basic sub-language sentence type. The idea is to organize the sub-language sentence into compact tabular representation so that the content of the document can be quickly inspected. In order to represent the information uniformly, the syntactically conveyed connections are translated into the occurrences of particular combinations of column entries in the format.

The growing interest in automated and integrated medical records has spurred intense research into indexing, abstraction and understanding clinical text. Several applications are enabled with Natural Language Processing (NLP) technology, but all are essentially text mining operations in terms of source documentation [2] specificity and depth of information required. The required information can range from a desire to determine which course of treatment are effective for particular conditions per patient group to wanting to know where the latest outbreak of community acquired diseases is taking shape so that a sales force can be first to market. Similarly, Demographic information is important first to identify and individual patients across multiple medical encounters.

[1]Department of Computer Science & Engineering, College of Engineering, Thiruvananthapuram, India
e-mail:vinodchandrass@gmail.com

Some of the interesting works related to medical information processing is discussed here. The ARBITER (Arterial Biology for the Investigation of the Treatment Effects of Reducing Cholesterol) is the application developed, and used MEDLINE citations [11] for information extraction. Medical Language Processing (MLP) and tagging of the medical text [4, 5] discusses statistically significant POS n-gram type overlaps of newspaper language and medical sublanguage, which has not been recognized before. A Dialogue-Based System for Identifying Parts for Medical Systems [6] describes a system that provides customer service by allowing users to retrieve identification numbers of parts for medical systems using spoken natural language dialogue. They showed a results of extremely encouraging with the system being able to successfully process approximately 80% of the requests from users with diverse accents. The use of clinical data present in the medical record to determine the relevance of research evidence from literature databases [3, 7] discussed. Here they used conventional information retrieval system for analysis of the patient's data record. Three algorithms are discussed in the above work.

Medical information processing from an unstructured text is highly related to NLP. This work deals with set of sub-languages and a parser. The parser interacts between the sub-languages and input document. A natural language processing tool for analysis the medical text and the information is extracted in a specified format using a parser.

2. Parser Design

Document Analysis

Medical document analysis that occurs in clinical documents and their associated lexical attributes are shown in Figure 1. We can observe that the lexical attributes appearing in the medical documents are the lexicon or semantic classes which we have to develop in order to arrive at the detailed analysis of the clinical documents. The statement of a medical fact is composed of a subject and a predicate (SUBJECT and PREDICATE) each of which has associated an atomic attributes or lexicons. The SUBJECT may be physically absent in the statement being modeled, but if so, it is implicit. The Medical Fact can be divided into subtypes like Clinical Fact, Treatment Fact and Response Fact [8]. Clinical Fact subtypes are distinguished by the paragraph they occur in: Examination, Diagnosis, Lab Test, History etc. The Treatment Fact subtype is subdivided into general medical management (GEN), Surgery (SURG), medications (MEDS) and all other therapies (Comp). The idea of classifying the medical document into various fact classes is the use of different lexicons in each class. The Laboratory Fact will be characterized by the description of a test (Text box) with its associated attributes shown in the box attached to the test box. These attributes are the set of lexical or semantic classes which distinguishes the Laboratory statements from the others. The Treatment Fact is likewise distinguished from other statement types by the lexical classes shown in the box attached to the Treatment Fact box, and similarly for the Response Fact. An instance of treatment fact is often coupled to a Response fact via a response relation.

The extracted information is grouped under several medical classes like treatment fact, medical fact, test fact etc. Some information may have associated contents like affected body parts, test/procedure performed, extent of damage etc. The output must clearly state the patient state, diagnosis and procedures. The input document statements are expected to follow simple grammar rules

so that the correct sentence structure can be defined and matching patterns can be found out. Also transcript templates may affect accuracy. So a common template is expected. Lexicons that cover the required words need to be present. The sizes of word storage files cause an increase in loading time but decrease errors. Still steps may be taken to avoid extra loading time and increased response time. The design overview is shown in Figure 2. A brief explanation of each section is given.

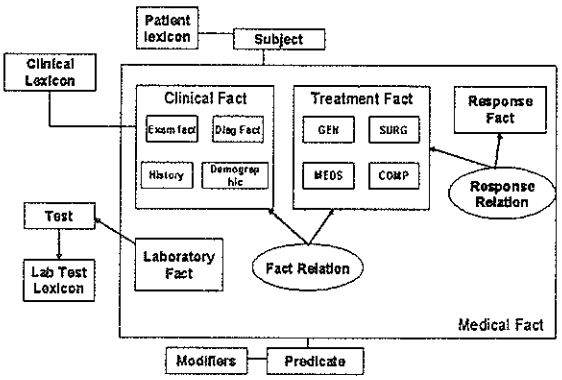

Figure 1 : Analysis of Medical Documents

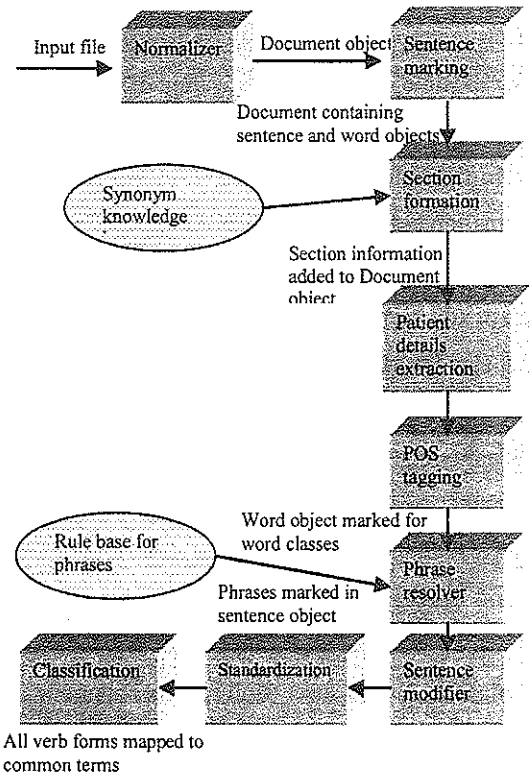

Figure 2: Parser Design Overview

Normalization

This step performs the initial preparation of text document. The document is read line by line. From every line each character is examined. ':' is replaced by new line. Remove '.' other than delimiters. For example the line is searched for patterns like 'Mr.', 'Mrs.' etc and '.' are removed. Multiple spaces are grouped into single space. Finally '.' is replaced by new line. This step effectively helps in correctly identifying sentence boundaries.

Sentence Marking

Document is split into several sentences based on delimiter (new line) and stored in sentence objects. Each sentence consists of several word objects. Words are marked using spaces as delimiters. The success of this step depends on how effectively all symbols other than proper delimiters are removed in the previous step

```
Procedure noun_finding ( )
Begin
  Read input line;
  Search a word in lexicon;
    If found
      Mark (noun) & Exit ( );
      If word ends with's' or 'es'
        Search a word in lexicon;
          If found
            Mark (noun, plural) & Exit ( );
      If word ends with 'ies'
        Replace 'ies' by 'y'
      Search a word in lexicon;
        If found
          Mark (noun, plural );
End;
```

Figure 3. Noun Finding Algorithm

Section Formation

A list of candidate section headings is created by examining each line of text and comparing the scanned text against regular expressions similar to the stored patterns. Each section heading may appear in several forms. These are mapped to common terminology. For example: Complaint: chief complaint, complaint

Allergy

allergy, allergies, allergy medication, allergy to medication

Physical exam

physical examination, physical exam, physical findings.

A section list is maintained, which stores the section headings in the order they appear in the document along with the offsets, with respect to the original note, of the first and last sentence in the document. The module also identifies a single section heading split across two lines.

Patient Details Extraction

Due to the special structure of medical document, certain sections formed will contain patient specific information like name, age, temperature etc. Several separate parsers are used to extract each type of information. Identify section heading 'Name': the content will be patient name. Then find sentences with words 'pulse' and find following number to get pulse. Next find sentence containing patterns like nn/nn and record as blood pressure. Then find word temperature or locate patterns nn0C or nn0 degree. The number will be patient temperature. Finally remove all sentences which are processed above.

Parts of Speech Tagger (POS)

POS tagger [9] aims at marking all the word classes like noun, verb or prepositions. The tagger is divided into two main phases of operation - lexical analysis, and confextual analysis. The lexicon contains all the possible parts of speech, such as noun, verb, or adjective, appropriate to each word contained therein. Each word from the text document is first marked with all the parts of speech listed for that particular word in the lexicon. If a word does not appear in the lexicon, the tagger will default to mark it as an unknown noun. The algorithm for finding a noun is given Figure 3.

Similar steps are undertaken for verbs which appear in forms like verb, verb + 3rd person, verb + ing, verb in past tense, irregular verbs. Verb in past tense form produce an ambiguity since it is past participle. So the word is marked as ambiguous. Each word object contains a status word where each bit corresponds to particular word class. Using several contextual rules, the contextual analysis phase processes the text further to ensure that the part-of-speech tags are disambiguated. First rule is used for this - if precedes with have/has/had mark it as participle. With this information, the tagger is able to determine the final part-of-speech tag for each word.

Phrase Processing

This processes the extracted phrases and transforms complex phrases into simple canonical syntactic structures of the medical sublanguage stored into the knowledge base [10]. The main functions includes the use rules to extract phrases from sentences, unknown words are grouped to phrases (this may be a medical term) and mark as unknown others. Noun phrases are extracted using a finite set of rules, composed of different sequences of part-of-speech tags [11]. The limit for the longest recognizable noun phrase pattern was set to seven words in length, with the shortest

pattern being obviously a noun phrase of length one, the single noun. The seven-word limit can lead to some error, as the tagger is likely to misidentify noun phrases longer than seven words as two completely separate noun phrases, which themselves may or may not be valid terms. The rules were applied to the tagged words from the text, using a sliding "window" of seven words. As the window slides over the words of the text, the noun phrase patterns are applied to the window contents. When encountered, the sentence delimiters will truncate the window. Since some of the rules are subset of other rules, the longest matching rule is used to determine the "best" noun phrase. Once a noun phrase is located, the window will slide to the next word following the phrase and commence reading the contents of a new seven-word window. Similar steps are undertaken for identifying verb phrases and preposition phrases.

Standardization

Standardization is performed so that there is uniformity through out in representing related terms. This helps in reducing the number of required sentence forms to be identified and aids in better understanding of the sentence. There are three forms of standardizations. Normalizing the phrase is used in canonical names wherever possible, to make text syntactically uniform. Which replace the modifiers by its root word through lookup in the synonym table. The Decomposition simplifies the complex phrases by replacing them by two simple phrases (Example. "Suffering from severely increasing cough and fever" to "suffering from severely increasing cough" & "severely increasing fever"). Finally, flatten the document, changes a sentence structure (Example, "no evidence of pneumonia" to "no pneumonia".

Classification

The basic sub language sentence types are identified from the given set of sentences. The characteristic combinations of sublanguage word classes in the SVO (subject-verb-object) relation. For example, in the sublanguage of clinical reporting, a frequent SVO sequence consists of a subject from the PATIENT class (usually patient or a pronoun) followed by a VERB in the V-PT class (a verb whose subject is characteristically a PATIENT noun in the sublanguage, Example, have, develop), followed by a word in the SIGN-SYMPTOM class (Example, cough, dyspnea). This basic sequence type (SVO = PATIENT + V-PT + SIGN-SYMPTOM) is seen in such text occurrences as Patient has had cough for past 4-5 months (SVO = patient + have + cough) and she had an episode of severe dyspnea while in Bangalore (SVO = she + have + dyspnea).

Another similar and rather simpler rule is BODY-PART + TEST + V-SHOW + SIGN-SYMPTOM which is illustrated in the sentence Chest x-ray shows aggressive bilateral pneumonia. Due to the complexities of the natural language, a pattern larger than the subject + verb + object pattern is needed to capture information patterns of the sublanguage. For the sentences like (chest x-ray shows density) there is a larger pattern like BODY-PART + TEST + V-SHOW + RESULT. It is possible that in some cases the BODY-PART may not be explicitly present but is implicit in the TEST word (urinalysis

implies urine). These medical classes and their associated information are stored in an object for medical summary.

Synonym Knowledge Base

The synonym base is used for mapping the regularized structured form to controlled vocabulary concept which is the final stage of Phrase Processing module described in later part of the report. The synonym knowledge base consists of associations between standard output forms and controlled vocabulary concepts. Some of the examples are shown below. The first argument of the synonym specification is the target or standard form of the textual phrase, the second is the controlled vocabulary concept, and the third is the semantic category of the synonym.

Synonym (show, appear, moderate certainty, certainty)
Synonym (enlarged heart, cardiomegaly, central finding)
Synonym (nodular density, nodular opacity, partial finding)
Synonym (severe, high degree, degree)
Synonym (smaller body_location, decrease in body_location size)
Synonym (without, with no, certainty)

Rule Base

Rule base is required to connect individual words to noun or verb phrases. The rules for noun phrases may be represented in Extensible Markup Language (XML) file as follows

<NounPhrase><Val>N</Val> </NounPhrase>
<NounPhrase><Val>PN</Val> </NounPhrase>

<NounPhrase><Val>PN</Val> </NounPhrase>

<NounPhrase><Val>DT</Val> <Val>N</Val> </NounPhrase>

The rule base is also required to map the words of the key phrases in the composite tabular structure. The simplest rule which was also earlier described in the synopsis containing a SUBJECT + VERB + OBJECT is PATIENT + V-PT + SIGN-SYMPTOM. XML files store these rules in the following format

<Rule><term>PT</term> <term>V-PT</term> <term>SYM</term></Rule>

3. Results And Discussion

This paper described a system for processing the patient discharge summaries and mapping the information into a database. The word classes and the controlled vocabulary of a sublanguage grammar have been shown effective for mapping textual information into a semantically structured database. Medical text mining is characterized by market requirement for very precise information at moderately deep level. An important part of the implementation process would be to maintain the precision and accuracy of the information extraction. The volume of data is growing every year and large portion of the data may be idiosyncratic or specific to the hospital or organization. The proposed framework with its inclusion of controlled and custom vocabulary takes care of the region or hospital specific jargon. Once we have the above knowledge base with us what remains is the implementation part. In order to ensure that the proposed technique when applied to medical documents produces reliable results a procedure should be developed for the quality control.

Case Study

The clinical information from a medical text after parsing is shown below. Here the input file is an unstructured text document with many pages.

Patient Name : PEREZ, LUIS
Sex : Male
Age : 35 years

History Of Present Illness

Luis Perez was seen at the request of Dr. Webb for evaluation of severe back and right leg pain. He is a pleasant 36 year old male who states symptoms began in August and it persisted. He states he can only walk with use of a walker. He denies any previous symptoms or problems. Currently he denies any weakness, loss of bowel or bladder function. He complains of severe pain frequently in his right leg all the way to his toes.

Physical Examination

His examination demonstrates young male who using a walker. He has some notable antalgic gait. He has tenderness with spasm at the paralumbar region. Range of motion is very limited. He has notable nerve tension sign with positive straight leg raise, positive Lasegue, positive flip _ test of the right lower extremity. Long track signs are same. He has weakness with both toe walk and with his anterior tib on the right compared to the left.

Medications

He is on Lorcet, Soma, naproxen, Xanax.

Allergies : Codeine, Penicillin

Lab Results

X-rays were obtained on this visit. These were an AP and lateral of his lumbar spine. These show the old fusion of L4-5.

REFERENCES

[1] Xiaohua Zhou, Hyoil Han, Isaac Chankai, Ann Prestrud, Ari Brooks, *"Computer Applications in Health Care (CACH): Approaches to Text Mining for Clinical Medical Records"*, Proc. of the ACM-SAC '06, April 2006.

[2] United States Patent 6915254,*"Automatically assigning medical codes using natural language processing Document:"* ,1999.

[3] Susan L. Price, Lois M. Delcambre, Marianne Lykke Nielsen, *"Medical document indexing and retrieval: Using semantic components to express clinical questions against document collections"*, Proceedings of the international workshop on HIKM '06, ACM Press, November 2006

[4] Thomas C. Rindflesch, Jayant V. Rajan, Lawrence Hunter, *"Extracting Molecular Binding Relationships from Biomedical Text"*, Proceedings of the sixth conference on Applied Natural Language Processing, Morgan Kaufmann Publishers Inc., PP. 188-195, April 200

[5] Udo Hahn, Joachim Wermter, *"High-Performance Tagging on Medical Texts"*, Proceedings of the 20th International Conference on Computational Linguistics COLING '04, Association for Computational Linguistics, August 2004

[6] Amit Bagga, Tomek Strzalkowski, and G. Bowden Wise , *"PartsID: A Dialogue-Based System for Identifying Parts for Medical Systems"*, Proceedings of the Sixth Conference on Applied Natural Language Processing, M K Publishers Inc., PP 29-36, April 2000

[7] Eneida A. Mendonca, Stephen B. Johnson, Yoon-Ho Seo and James J. Cimino, *"Analyzing the Semantics of Patient Data to Rank Records of Literature Retrieval"*, Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain,

Association for Computational Linguistics, PP 69-76, July 2002

[8] Sandra L. Johnson, *"Understanding Medical Coding: A Comprehensive Guide"*. Delmer Publishers, 2005.

[9] Hideki Hirakawa, Kenji Ono, Yumiko Yoshimura, *"Automatic refinement of a POS tagger using a reliable parser and plain text corpora"*, Proceedings of the 18th Conference on Computational Linguistics - Volume 1, Association for Computational Linguistics, July 2000.

[10] Prashant Baronia, *"Intelligent Information Retrieval from Medical Records"*, Computer Science Department, IIT Bombay, 1997.

[11] Bennett NA, He Q, Powell K, Schatz BR. *"Extracting noun phrases for all of MEDLINE"*, Proc AMIA Symp. PP 671-5, 1999.

Author's Biography

Vinod Chandra S S presently worked as a faculty member in Department of Computer Science & Engineering, College of Engineering Thiruvananthapuram, Kerala. He had his BTech (Computer Science & Engineering) from University of Calicut, Kerala and MTech (Software Engineering) from Cochin University of Science and Engineering, Kerala. Presently he is doing PhD in Computational Biology in University of Kerala. Since 1999, he has taught in various Engineering Colleges.He has a modest number of research publications in National and International levels including journals such as ACM, IEEE Conference Proceedings IJCAI Proceedings and Journal of Computer Society of India. He is a member of IEEE, Computer Society of India, Indian Society for Technical Education, and Institution of Engineers.