Medical Information Extraction Using Natural Language Interpretation

[1]Gunjan Dhole and [2]Dr. Nilesh Uke,

[1]Department of IT, PG Student, Sinhgad College of Engg, Pune, 411041,India
[2]Department of IT, Associate professor, Sinhgad College of Engg, Pune ,411041,India

ABSTRACT

NLP Based Retrieval of Medical Information is the extraction of medical data from narrative clinical documents. In this paper, we review Natural Language Processing (NLP) applications designed to extract medical problems from narrative text clinical documents. This paper also covers the methods used in this field and it also describes the architecture of the proposed system. However extraction of medical information is the difficult task due to complex symptom names and complex disease names. Proposed system is an expert system which will try to understand the input that can be the question about disease or list of symptoms and the system will try to give out the proper answer. Proposed system will consist of modules such as data processing, query processing, data extraction, answer matching, user interface etc. Lots of data are available in the medical field in free text form which is not used by normal rule based systems, so with the help of NLP we can use this free text data such as it will try to give out the answers that we search on our own.

Keywords:

Extraction, medical information, narrative text, NLP

1. Introduction :

The increase in use of electronic health records and the corresponding interest in using these data for quality improvement and research in this field states that the interpretation of free text contained in the records is a difficult step. The biomedical text is another important information source which will benefit from structuring of data in narrative text. Different approaches are implemented for extraction of the medical text. Natural language processing approach uses tools like noun entity recognizers, coreference resolution, part of speech taggers and relationship extractors. However medical text is different from normal text as it contains complex terminologies, so medical information needs advanced versions of these tools.

1.1. Motivation :

Following points are the points express the need for an automated system.

- **Need for Text Processing**: If we consider the data that we want to use as a database it should always be present in the database format. We miss out so much of data just because we don't have it in a proper format. It states that there is a need to develop such a system which can process this data. So that we never should face the problem of lack of data and the data which is already available in free text form can be used for better purpose.

- **Need of Medical Text Processing**: In case of the medical field, there are lots of data available. But medical field does not have that much development with the help of Natural Language Processing. Medical data such as Electronic Medical Records are available. But there are very fewer systems that can take data from these medical documents. So there is a need to process this data which can be achieved with the help of natural language processing.

- **Need for an automated system for diagnosis of disease**: The normal patient doesn't have the facility to predict the disease he is having until he gets to the doctor. Sometimes patients ignore their symptoms in early stages on any disease, which can be harmful. If the patients will be provided with the automated disease diagnosis system then can at least know the severity of the disease

- **Need for Automated system for doctors**: The doctors have to keep track of very big chunk of knowledge. If they can take help of some automated system which can help them to diagnose a disease it will be good. It states that there is a need for automated system for doctors also.

These denote that there is a requirement for developing the system that can process medical documents that can give the diagnosis of the disease and severity of the disease. The aim of this work is to evaluate an automated approach to the risk stratification of general diseases using Natural Language Processing (NLP) on medical documents. The final goal of the research is to increase patient safety by providing him the information about his diseases and severity of his diseases.

1.2. Why To Use Nlp For Information Retrieval :

Most of the systems are using Rule based systems for extraction of information. In rule based systems, the number of rules is limited. Because of the limited number of rules the information extraction gets limited. If some new information is required then the rule based system fails to extract information due to lack of rules. In natural language processing, we can extract data from free text form. Enormous amount of data is available in free text form is available today. But it is not properly utilized. The text such as Electronic Health Records has patient data that can be used for many purposes. The Free text such as disease information, its symptom and causes all can be found in the free text. Natural language processing(NLP) can be used to extract all these and get proper information. It is also found that doctors don't use their whole knowledge for any disease. The automated system for data extraction can help doctors for proper recognition of diseases.

2. Related Work :

Different medical extraction systems like MedlEE, MetaMap, linguistic string project were proposed [9]. MedLEE is developed to extract, structure, and encode clinical information into textual patient reports so that data can be used properly. Carol Friedman developed MedLEE with the Department of Biomedical Informatics at Columbia University, the Radiology Department at Columbia University, and the Department of Computer Science at Queens College of CUNY[9]. Dr. Alan Aronson developed MetaMap that is highly configurable software developed at the National Library of Medicine (NLM) to

convert biomedical text to the UMLS Metathesaurus and also to discover Metathesaurus concepts referred to in text [9]. The Linguistic String Project (LSP) was a developed (1960-2005) in the computer processing of language which is based on the linguistic theory of Zellig Harris: linguistic string theory, transformation analysis, and sublanguage grammar [9].

Different tools that are used for natural language processing are NER, pos-taggers, co-ref resolutions and Relationship extractors. Branimir T. Todorovic and Svetozar R. Rancic proposed a system for Named Entity Recognition and Classification using Context Hidden Markov Model [14]. Mohamed Hashem proposed A Supervised Named-Entity Extraction System for Medical Text. Andreea Bodnari. Louise Deleger proposed system for Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain [14].

Jiaping Zheng proposed a system for coreference resolution for the clinical narrative [15]. Wafaa Tawfik, Abdel-moneim proposed a system for Clinical Relationships extraction techniques from patient narratives [7]. The Ontology Development corpus and Information Extraction corpus annotated for co reference relations consists of 7214 coreferential pairs, forming 5992 pairs and 1304 chains. Classifiers can be trained with semantic, syntactic, and surface features pruned by feature selection. For the three system components for the resolution of relative pronouns, personal pronouns, and noun phrases. Support vector machines with linear and radial basis function kernels, decision trees, and perceptrons can be used for machine learning [10].

Ddan shen PROPSED A MedPost: a part-of-speech tagger for biomedical text [13]. This tagger was developed to meet the need for the high accuracy part-of-speech tagger trained from the MEDLINE corpus. This program currently accepts text for the purpose of tagging in either native MEDLINE format or XML. MEDLINE is a database of publications in health sciences, biology and related fields. It currently contains over 12 million records and nearly 7 million include an abstract [13].

Semantic relations can be extracted with the help of annotation approach which relies on linguistic patterns and domain knowledge which consists of two steps [8]:

(i) Recognition of medical entities
(ii) Identification of the correct semantic relation between each pair of entities.

The first step is obtained by enhanced use of metamap. The second step depends on linguistic patterns that are built semi-automatically from a corpus selected. According to semantic criteria, evaluation of the treatment relations between a treatment and a disease can be extracted.

3. Information Retrieval :

The research in "Natural Language Processing" (NLP) is going on from many years which were formed in 1960 as a sub-field of Artificial Intelligence and Linguistics, with the aim of studying problems in the automatic generation and understanding of natural language. A general text retrieval system consists of three entities

1. Records or full text documents
2. Indexer
3. Information retrieval tools

With the help of indexer, records are indexed and retrieved with the help of extraction tools. Natural Language Processing can be added at any or all of these stages. NLP interprets and stores meaning at for both the query and the document. If we add NLP information retrieval following stages should be followed.

Step 1: Document Processing
Step 2: Query Processing
Step 3: Query Matching
Step 4: Ranking & Sorting

4. Proposed Work :

The proposed system consists of two major modules: Document Processing with Natural Language processing and query processing with natural language processing. Both are very important phases of the project. The proposed system consists of 5 modules as shown in the architecture. Also, it contains the knowledge base. Knowledge base actually stores all the medical documents:

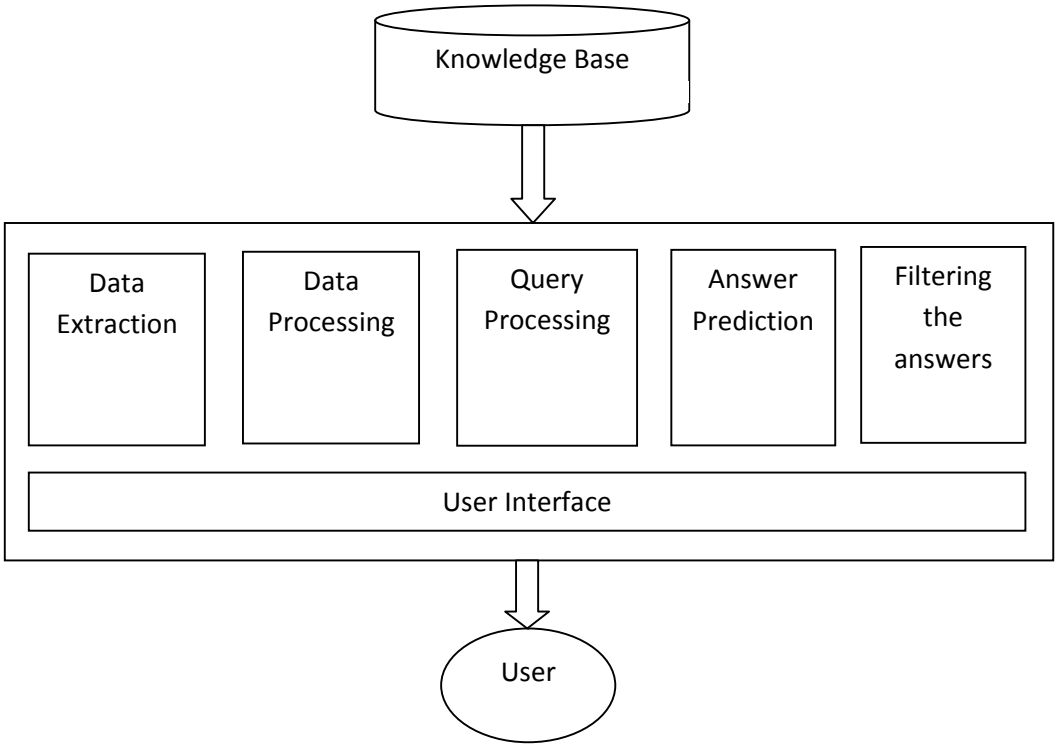

Fig. 5.1 the Architecture of Proposed System

4.1. Flow Of The Proposed System:

Query will be given as input to the query processing module. Important entities will be extracted from the query. According to these entities, documents will be processed for extraction proper answers to the queries.

Answer matching will try to find the best possible answer from set of answers
Proposed system contains following modules:

1) Data Extraction
2) Data Processing
3) Query Processing
4) Answer Matching
5) Filtering the Answers

4.2. Data Extraction:

This module will manage knowledge Base. It will try to extract some data from Internet. Knowledge base will also consist of different resources that contain biomedical texts regarding different systems. It can have any kind of free text that involves disease's information.

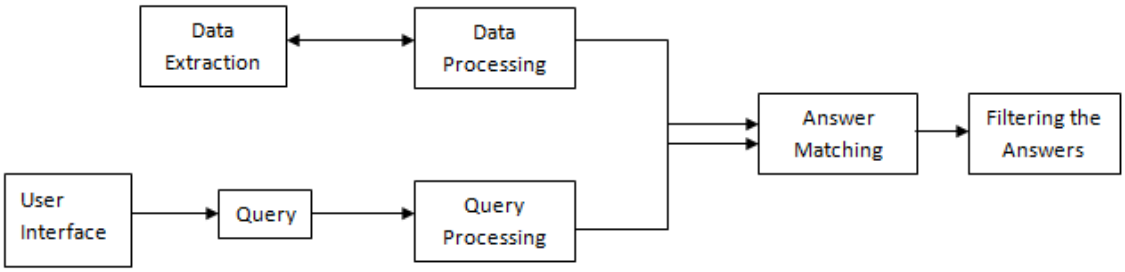

Fig. 5.2. The flow of the project

.

4.3. Document Processing:

Document processing involves processing of documents with the help of Natural language processing. It will involve parts such as section splitting, tokenization, relationship extraction, etc.

4.4. Query Processing:

Query processing will involve the processing of a query with the help of natural language processing. It will extract all the important relationships and keywords from the query.

4.5. Answer Prediction:

Answer prediction will involve the prediction of answer from the given set of relations and keywords.

5. Outcome Of The Project:

The aim of the project is to get the answers of the disease related queries with the best precision and recall.

If the user enters set of symptoms then system should be able to answer with the probable disease name.

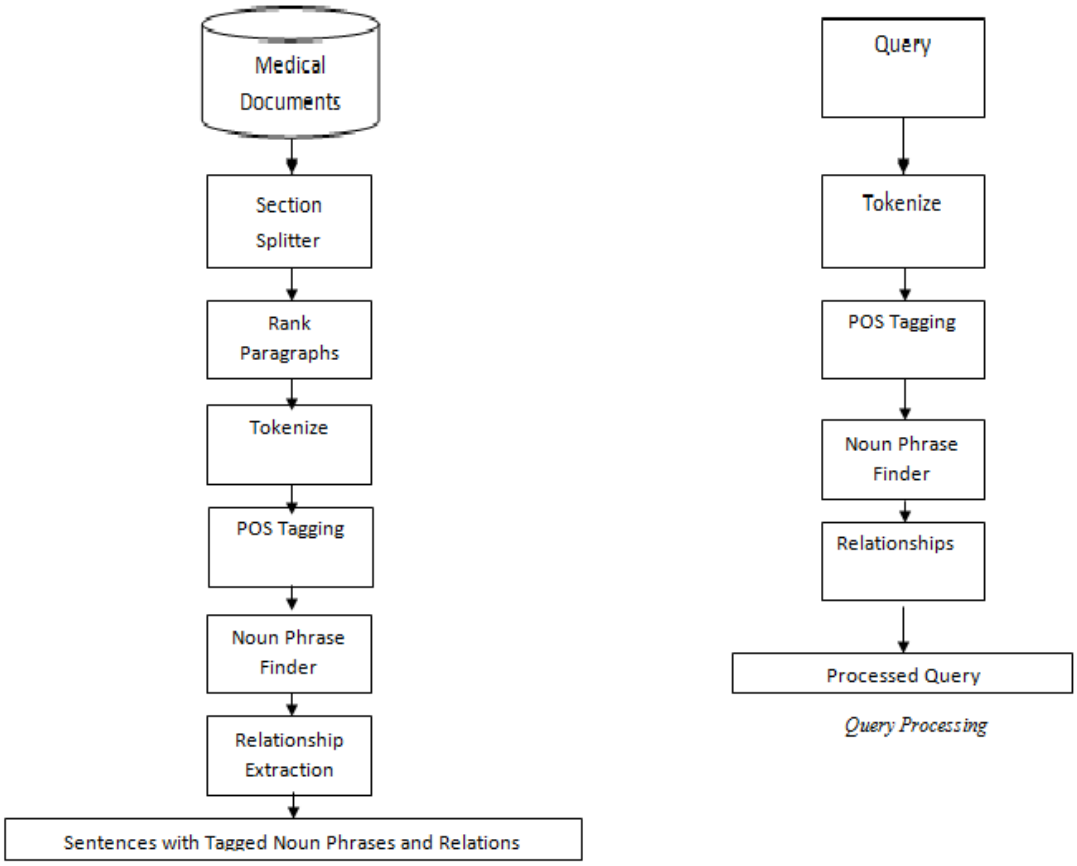

Fig.4.3.1. Document Processing

6. Conclusion:

Lots of researches are going on in the field of extraction of medical text with the help of NLP. As medical text is different from normal text, it needs advanced tools as compared to the normal NLP tools. Also, some of the systems are proposed for text extraction. Still there is a need for better medical text extraction systems. The proposed system extracts disease information according to the query currently.

References

[1] Andreea Bodnari, Louise Deleger, Thomas Lavergne, "A Supervised Named-Entity Extraction System for Medical Text"
[2] Ngô Thanh Nhàn "linguistic string project - medical language processor" at http://www.cs.nyu.edu/cs/projects/lsp/
[3] The Brandeis University, "MedLEE" at <http://www.medlingmap.org/taxonomy/term/80>
[4] James Freeman-Hargis "Introduction to Rule-Based Systems" at <http://ai-depot.com/Tutorial/RuleBased.html>

[5] Asma Ben Abacha, Pierre Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach" From Fourth International Symposium on Semantic Mining in Biomedicine (SMBM)

[6] D. Nagarani, Avadhanula Karthik, G. Ravi, "A Machine Learning Approach for Classifying Medical Sentences into Different Classes", IOSR Journal of Computer Engineering (IOSRJCE) Volume 7, Issue 5 (Nov-Dec. 2012), PP 19-24

[7] Dan Shen Jie Zhang Guodong Zhou, "Effective Adaptation of a Hidden Markov Model-based Named Entity

[8] Faguo ZHOU Enshen WU, "The Design of Computer Aided Medical Diagnosis System Based on Maximum Entropy" 978-1-61284-729-0111 2011 IEEE

[9] Hinxton, UK. 25-26 October 2010

[10] Jiaping Zheng,1 Wendy W Chapman,2 Timothy A Miller,1 Chen Lin, "A system for coreference resolution for the clinical narrative", J Am Med Inform Assoc (2012). doi:10.1136/amiajnl-2011-000599

[11] Khan Razik, Dhande Mayur , "To Identify Disease Treatment Relationship in Short Text Using Machine Learning & Natural Language Processing", Journal of Engineering, Computers & Applied Sciences (JEC&AS), Volume 2, No.4, April 2013

[12] Kyle D. Richardson1, Daniel G. Bobrow1, Cleo Condoravdi1, Richard Waldinger2, Amar Das3, "English Access to Structured Data", 2011 Fifth IEEE International Conference on Semantic Computing

[13] L. Smith1, T. Rindflesch2 and W. J. Wilbur, "MedPost: a part-of-speech tagger for bioMedical text", Vol. 20 no. 14 2004, pages 2320–2321, bioinformatics/bth227

[14] Lucila Ohno-Machado, Editor-in-chief, Prakash Nadkarni, Kevin Johnson "Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature", amiajnl-2013-002214

[15] Romer Rosales, Faisal Farooq, Balaji Krishnapuram, Shipeng Yu, Glenn Fung, "Automated Identification of Medical Concepts and Assertions in Medical Text Knowledge Solutions" , AMIA i2b2/VA text mining challenge

[16] Stéphane M. Meystre, MD, MS, Peter J. Haug ,"Comparing Natural Language Processing Tools to Extract Medical Problems from Narrative Text", MD AMIA 2005 Symposium Proceedings

[17] téphane Meystre, Peter J Haug, R. Engelbrecht et al., "Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx) Connecting Medical Informatics and Bio-Informatics", ENMI, 2005

[18]Stphane Meystre, Peter J. Haug, "Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation", Journal of Biomedical Informatics 39 (2006) 589–599

[19]Wafaa Tawfik Abdel-moneim1, Mohamed Hashem , "Clinical Relationships Extraction Techniques from Patient Narratives", JCSI International Journal of Computer Science Issues, Vol.10, Issue 1, January 2013