Relationship Extraction in Clinical Health Data using NLP

Naveen S Pagad
Assistant Professor
Dept. of ISE, SDM IT Ujire

Gagana H R
4SU14IS009
Dept. of ISE, SDM IT Ujire

Chaithra Kumari
4SU13IS014
Dept. of ISE, SDM IT Ujire

Megha H R
4SU14IS014
Dept. of ISE, SDM IT Ujire

Sharadhi H R
4SU14IS30
Dept. of ISE, SDM IT Ujire

***Abstract*:- Patient data is critical in healthcare domain. Secure, consistent, and coded information increases the efficiency and encourages collaboration within and between organizations. So that Biomedical natural language processing deals with the application of relaxation extraction techniques to clinical documents and to scientific publications in the areas of biology and medicine.**
Processing of such data in e-health or clinical decision support systems is a challenging task. This paper proposes a method that extract semantics from medical discharge summaries using natural language processing approach. And also presents a mapping and transformation system of discharge summaries of hospital written in natural language.

General Terms:- Clinical health data, Relaxation extraction.

Keywords:- Natural language processing(NLP), Conditional Random Fields (CRF), semantics analysis.

1. INTRODUCTION

The constant growth of published biomedical research and of archives of documents generated by clinical practice creates the need for tools that help researchers and practitioners of the biomedical field to cope with large amount of information effectively and efficiently. For the physician, medical records chronicle previous diagnoses and allergic reactions of a patient, providing continuity of information as the patient is transferred from one doctor to another. This work comes under Biomedical NLP (also known as BioNLP) which refers to natural language processing techniques applied to texts and literature of the biomedical and molecular biology domain.

Health data is critical and needs to be stored and retrieved in a structured form. Quality of patient care depends upon the collection and dissemination of information about patients. Health data is critical and needs to be stored and retrieved in a structured form. High level standards are defined which determine the structure of the medical knowledge and provide information model for medical record. They help in information exchange among different healthcare systems.

Moving patient's records from paper or physical fling systems to standardized computer based system creates utility for patients, providers, and decision support systems. It also plays very important role in communicating information within different healthcare providers or stakeholders. Here we tackle the challenge of understanding patient records. They describe an approach to extract salient information from natural language medical text and create appropriate representations of this information.

We focus on medical discharge summaries which are documents that a physician dictates after a patient's hospital stay, highlighting the patient's reasons for hospitalization, test results and findings, diagnoses, and prescribed medications.

2. RELATED WORK

Ayisha Noori V. K and P. C. Reghu Raj [1] proposed a "Extraction of Disease Relationship from Medical Records: Vector Based Approach" .This paper proposes a method that extract semantics from medical discharge summaries using vector based approach. In particular. This work comes under Biomedical NLP (also known as Bio NLP) which refers to natural language processing techniques applied to texts and literature of the biomedical and molecular biology domain. In this paper, they focus on medical discharge summaries which are documents that a physician dictates after a patient's hospital stay, highlighting the patient's reasons for hospitalization, test results and findings, diagnoses, and prescribed medications.

Stefano Ballerio [2] proposed a "Automatic Analysis Of Electronic Discharge Letters As A Means To Evaluate The Continuity Of Information And Of Patient Care". The constant growth of published biomedical research and of archives of documents generated by clinical practice creates the need for tools that help researchers and practitioners of the biomedical field to cope with this large amount of information effectively and efficiently.

Rabia Batool, Asad Masood Khattak and et.al [3] proposed "Automatic Extraction and Mapping of Discharge Summary's Concepts into SNOMED CT". This paper presents a mapping and transformation system of discharge summaries of hospital written in natural language to standardized data, which can be easily used by computer based medical applications.

Sunil Kumar Sahu, Ashish Anand and et.al [4] proposed a "Relation Extraction From Clinical Texts Using Domain In Variant Convolution Neural Network" In this work they focus on extracting relations from clinical discharge summaries. Main objective is to exploit the power of convolution neural network (CNN) to learn features automatically and thus reduce the dependency on manual feature engineering.

Julien Tourille,Olivier Ferret and et.al [5] proposed "Temporal Information Extraction From Clinical Text". Most of the temporal information remains locked within unstructured texts and requires the development of NLP methods in order to be accessed. In this paper, they focus on the extraction of temporal relations between medical events (EVENT),temporal expressions (TIMEX3) and document creation time (DCT).

3. METHODS

To evaluate and predict the health status of a patient ,first need to understand the methodology of diagnosis making and then able to extract from a patient's EHR(electronic health record) the relevant data that can be used in determining the diagnosis.

Electronic Medical Records (EMRs), generated in the process of clinical treatments, refer to systematized collections of patients' clinical information stored in electronic medical records systems . EMRs contain a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. A large amount of medical knowledge, closely related to patients, can be discovered through analyzing these medical records . Generally, EMRs are always described in the form of natural language, from which mining patient-related health-care medical knowledge needs applications of information extraction related technologies.

Temporal information extraction from electronic health records driven by the need for medical staff to access medical information from a temporal perspective .Diagnostic and treatment could be indeed enhanced by reviewing patient history synthetically in the order in which medical events occurred.

3.1 Finding The Diagnosis Sections

Most discharge summaries have both a list of diagnoses for the patient as they arrived and as they left. The arrival diagnoses are usually called the "past medical history".The discharge summary is divided into sections usually with a label in upper case and separated with a colon .

Natural Language Processing (NLP) has been widely applied in biomedicine, particularly to improve access to the ever-burgeoning research literature. The input consists of natural language documents containing unstructured text. These documents are fed to the RE system, process of detecting and classifying the semantic relation among entities in a given piece of texts. So the program first tries to recognize how sections are labeled by looking for some standard section names at the beginning of a line and determines how these are formatted.

3.2 Automatic Labeling

Next the program looks for sections labeled as diagnoses or medical history. If there is no labeled admitting diagnosis or medical history section, the program looks for the present illness section and tries to find a paragraph starting with an appropriate phrase. For this we uses a Conditional Random Fields (CRF) [LMP01] because it is the labeling algorithm. Otherwise the program looks for "history of" followed by a disease or procedure in the present illness section.

3.3 Data Processing

This stage also includes the identification of files, that is removal of personal information such as name, address, hospital name, doctor's name, medication details and unwanted sections like observations etc. This step is necessary to maintain the privacy of patients and the hospital authorities demanded this while providing the dataset.

It identifies and characterizes the relations described in the text data. The output of the RE system consists of relation mention triples which include the two entity mentions that take part in the relation and the relation type.

3.4 Data Access

The data can be accessed in multiple ways. The simplest way is by simple SQL queries on the tables of interest. Another access mode is the use of a query language that was developed specifically for this purposes .It is possible to restrict the output to a predefined set of desired patient-IDs which are contained in the output set.

4. FIGURE

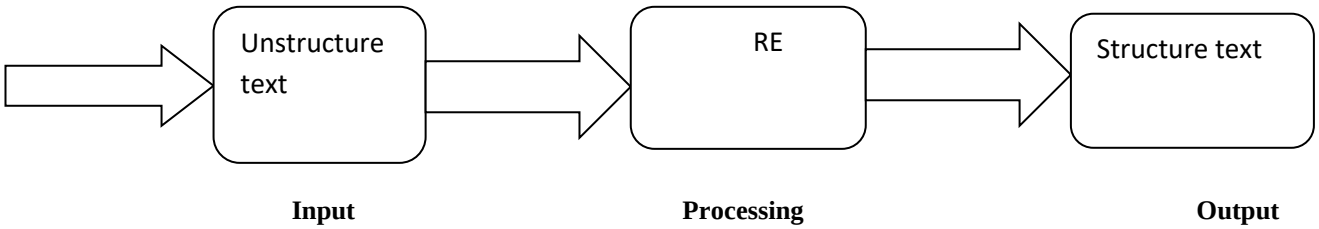

Fig 1: NPL data processing

5. CONCLUSION

This work explores the potential of relation extraction in medical domain. The system provided discharge summary has a unique format. Given this condition, and because of the increasing availability of electronic textual data, NLP represents a way to evaluate the continuity of information and of care. The proposed system helps in processing of natural language to make it usable for healthcare applications like clinical decision support system and making system interoperable. Finally we able to convert the unstructured text to structured format and also gave the relations and relation types using the method of NLP and CRF algorithm.

6. REFERENCES

[1] Ayisha Noori V. K and P. C. Reghu " Raj Extraction of Disease Relationship from Medical Records: Vector Based Approach" International Journal of Latest Trends in Engineering and Technology (IJLTET) Vol. 3 Issue2 November 2013 ISSN: 2278-621X,

[2] Stefano Ballerio "Automatic Analysis Of Electronic Discharge Letters As A Means To Evaluate The Continuity Of Information And Of Patient Care" June 2009 Printed on demand by "Nuova Cultura".

[3] Rabia Batool, Asad Masood Khattak ,Tae-Seong Kim and Sungyoung Lee "Automatic Extraction and Mapping of Discharge Summary's Concepts into SNOMED CT"

[4] Sunil Kumar Sahu, Ashish Anand, Krishnadev Oruganty, Mahanandeeshwar Gattu"Relation extraction from clinical texts using domain in variant convolution neural network" Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 206–215,Berlin, Germany, August 12, 2016. c 2016 Association for Computational Linguistics.

[5] Jianhong Wang, a, Yousong Peng, Bin Liu, Zhiqiang Wu, Lizong Deng, , and Taijiao Jiang "Extracting Clinical entities and their assertions from Chinese Electronic Medical Records Based on Machine Learning" 3rd International Conference on Materials Engineering, Manufacturing Technology and Control (ICMEMTC 2016).