Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Disease Treatment Analysis for Knowledge Based System Using Natural Language Processing

Poonam Kadam*, Priyanka Kakade, Meenal Omase, Sanchita Sarkar, Amol Bavaskar
Computer Department & Pune University, Maharashtra,
India

Abstract*— In this era of fast world the standard of healthcare relies on the delivery and disposal of healthcare systems due to constraints like time and money. The delivery of modern health care depends on groups of trained professionals and paraprofessionals coming together as interdisciplinary teams. Patient medical record is a vital constraint in healthcare domain and it should be secure, consistent and perfectly coded for the secure transfer from one potential user to another. In this paper, we tackle the problem of ``understanding'' patient related data for a machine learning approach using Natural Language Processing (NLP). An imperative area of Natural Language Processing is semantic analysis, the study of the meaning of lingual utterances. Machine learning is an experimental domain automatic learning which is used in various fields such as medical decision support systems, medical image mapping, protein to protein interaction, extraction of medical facts or medically related data, and for patient's complete management care. We target on three semantic relations: Cure, Prevent, and Side Effect. This is the first step towards achieving reliable/accurate disease and treatment related information which is useful for better understanding.

The prime intention of this work is to show what Machine Learning (ML) and Natural Language Processing (NLP) techniques used for illustration of data and what classification algorithms are convenient for determining and classifying relevant medical information in short texts. In this paper, we focus on diseases and symptoms and the output of these subjective domains would be presented in the form of cure, prevent, side-effects. The first work is to identify the sentences which are published in Pub MED websites. The second task has a deeper semantic dimensional relation and it focus on identifying semantic relations exists between disease-treatment.

Two main challenges in the area of machine learning are the selection of a good learning algorithm and feature representation technique. Therefore the proposed technique integrates with any medical management system to make better medical decisions and inpatient management system by automatically mining the biomedical information from digital repositories.

Keywords*— "Data Mining", "Electronic Healthcare Record System (HER)", "Healthcare", " Machine Learning", " Natural Language Processing", " Pub MED", " Sentence extraction".

I. INTRODUCTION

The steady expansion of medical knowledge has made it more difficult for the physician to remain abreast of medical innovations outside a narrow domain. Consultation with a doctor is a solution when the clinical problem lies beyond the physician's ability, but frequently specialist or expert opinion is not always available as it depends on the circumstances. Attempts have been made to develop computer programs that can serve as virtual doctors.

The healthcare environment is generally rich for information but relatively poor for knowledge. There is a wealth of data available within the healthcare systems but they lack effective analysis tools to discover hidden relationships and trends in data. A major challenge posed to the healthcare decision makers is to provide quality services. Quality service implies diagnosing patients correctly and administering treatments that are effective. The proposed system aims at simplifying the task of doctors. When the doctor fires a query regarding symptoms or disease then the system provides the information regarding preventive measures, cure and side effects of the inferred disease. Basically this project aims at reaping the benefits of the two fast developing research areas i.e. Data Mining and Machine Learning techniques by discovering a framework that integrates both the research areas. The tools that are able to identify reliable information in the medical domain stand as construction blocks for this healthcare system. In this system, we focus on diseases and treatment, prevention and side-effects related information, and the relation that exists between these many entities. The approach used to solve the two proposed tasks is based on NLP and ML techniques.

II. PROBLEM DEFINITION

A. Problem statement

To develop a system for identifying disease, symptoms and treatment for cardiovascular diseases using data mining and machine learning approach

1) Scope

- The input dataset comprises of the papers published in Journals of Cardiovascular Disease Research website.

- The end user of the proposed system would be a medical practitioner.
- The outcome of the system is based on information present in medical papers and does not take patient's medical records into consideration.

2) Design and Implementation Constraint

- The data source for the project comprises of papers from Journals of Cardiovascular Disease and Research. Hence the constrain of authenticate data source is fulfilled as all the papers under JCDR are approved by group of research scientists, vascular disease experts and cardiologists coming from North America, Asia and Europe etc.
- This also guarantees accurate and optimised results.

3) Assumptions

1. The user should have internet connection for using our application software.
2. The user should be authorized and have the experience in medical field.

4) User classes and Characteristics

This system is designed from point of view of Doctor as the end user.
Characteristics and Requirements-

- The doctor should have the basic knowledge to operate the Computer.
- The doctor would be an authenticated user with an exclusive username and password.

The queries entered should be in English language only.

III. LITERATURE SURVEY

In Jason D. M. Rennie, Lawrence Shih, Jaime Teevan,David R. Karger ,"Tackling The POOR Assumption O Naïve Bayes Text Classifier" naïve based text classifier were used, but it didn't gave 100% precision output. Sometimes prediction of classifier may be incorrect[1].

In T.Mouratis, S.Kotsiantis, "Increasing The Accuracy Of Discriminative Of Multinominal Bayesian Classifier In Text Classification", in this paper classifier was used correctly and it gave high precision output, but it was not able to classify the of verbs, nouns, adjectives correctly, so sometimes it may give wrong output [2].

In B.Rosario and M.A. Hearst,"Semantic Relation in Bioscience Text" here for entity recognition there was use of Hidden Markov models. Natural language is converted into text biomedical data is mapped into structural form. Information extraction was used for machine learning. In this paper Gene-Protein from Medline abstract is used, and the results are shown in graphical representation [3].

In Oana Frunza.et.al, "A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts" In this paper for classification of sentences a medical dictionary is used. Then these sentences are parsed in semantic parser. In this process few phenomenon are included such as extraction ,alteration, validation process are applied in order to differentiate the actual semantic relations in order to be extracted, but problem exists that only one algorithm of machine learning may not prove beneficial or provide better output[4].

In L. Hunter And K.B. Cohen, "Biomedical Language Processing: What's Beyond Pubmed" sentence processing Natural language processing is used for biomedical sentences processing. It actually matches the disease name from the query in the database as specified in the query and when the particular match is found ,then finally the solution is provided ,but the disease may not be identified automatically[5].

IV. PROPOSED METHODOLOGY

A. Functionality Requirements

1) Log In

Description: If the user is a registered member then Login is the feature of the system which allows the user to enter into system and will provide him access to the system.

2) Register

Description: If the user is new to the system then he need to register first for getting access to the system. Register will provide user a option for registration.

3) Select the mail server

Description: On logging in User is shown the home page for his profile with username and password option.

4) Edit personal details

Description: If user wants to update any of the personal information then he can edit it.

B. Operating Environment

TABLE I OPERATING ENVIRONMENT

Front End	JAVA
Back End	MYSQL Database 5.6
Tools	Netbeans IDE 7.2.1, OpenNPL 1.5.3
Server	1) Apache Tomcat 8.0.0-RC3 (alpha)
Operating system	Windows XP and above

Other coexisting software	HTTrack website copier,
System	Pentium IV 2.4 GHz and above
Hard disk	40 GB (minimum)
RAM	256 MB (minimum)

C. System Features

1) Provide diagnosis result

TABLE III FONT PROVIDE DIAGNOSIS RESULT

Requirement ID	1
Requirement Category	The system provides diagnosis result based on the query given in the form of symptom or disease.
Requirement Importance	Mandatory
Requirement Description	This is an actual task intended.
Priority	Highest
Difficulty	High

2) Update Dataset

TABLE IIIII UPDATE DATABASE

Requirement ID	2
Requirement Category	The medical data used for this system is updated regularly so, as to keep the system up to date.
Requirement Importance	Mandatory
Requirement Description	It is for updating of medical papers.
Priority	Medium
Difficulty	High

3) Optimize search result

TABLE IVV OPTIMIZE SEARCH RESULT

Requirement ID	3
Requirement Category	The system provides Optimized search results.
Requirement Importance	Medium
Requirement Description	If user enters input as symptom, there is possibility that this symptom can be present in more than one disease. Hence the system provides the feature of optimize search results for effective diagnosis.
Priority	Medium
Difficulty	High

V. SYSTEM ARCHITECTURE

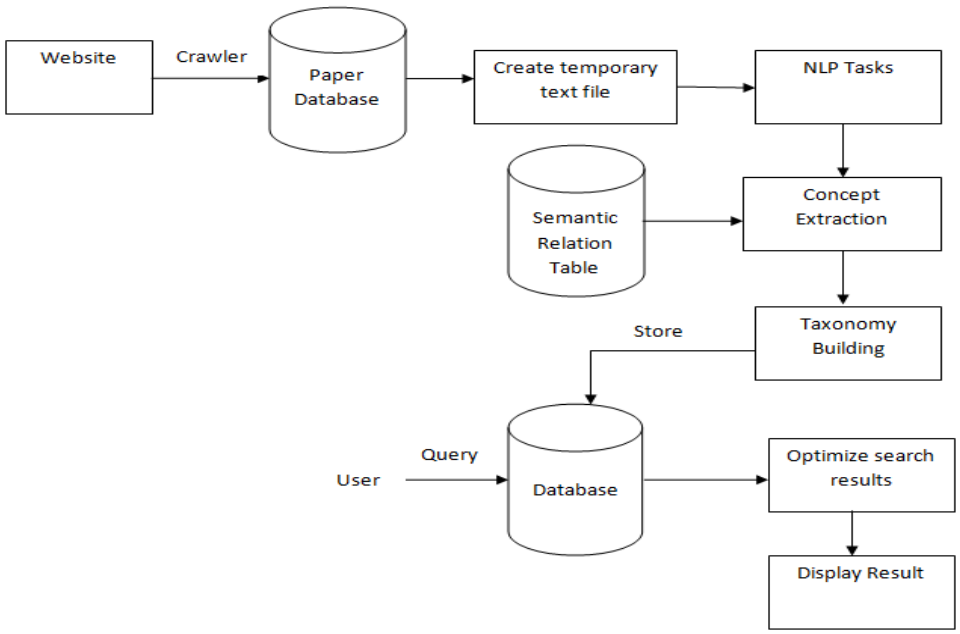

Fig. 1 System Architecture

To understand the user query in a more better way we are using Natural Language Processing and for data mining purpose we are using the Machine Learning Algorithms.

- Module 1:-HTML to text

Here the relevant web pages are downloaded from web and then converted to zip folders and using XML parsers later converted toXML files and finally gets converted to text document.

- Module 2:-Data obtained is to be processed

After converting file we eliminate the irrelevant data using Sentence Detection, tokenization, POS tagging and Named-entity detection.

- Module 3 :-Semantic extraction

Here in semantic extraction it needs to recognize the data from different websites of medical domain and then segregate th data accordingly.The mission is parallel to a scan of sentences contained in the abstract of an article in order to present to the user only sentences that are identified as containing relevant information on the basis of cure,prevent and side-effects and also prioritization is done.

- Module 4:-Data collection

After receiving the separated information as per the semantic extractions, we can store thes data in database.

- Module 5:-Analysis

Finally we have to analyze the results and also optimize the results and finally display it to the user.

VI. TEST CASES

- The input given to the system in the form of medical papers taken from: www.medline.com
- This site contains journals for human diseases only.
- It is an official publication for medical researchers of Scibiolmed.org. It is a non-profit Bangalore based organization dedicated to research in the field of Science, Biology and Medicine.
- And hence contains authentic data as they are being examined by the experts before being published.
- The journal is indexed with Google Scholar, Health and Wellness Research Center, PubMed, PubMed Central Caspur, Chemical Abstracts etc.

A. Evaluation of confusion matrix:

Let us take an example:

		Predicted	
		Negative	Positive
Actual	Negative	a (30)	b (10)
	Positive	c (5)	d (40)

Accuracy (AC) = (a+d) / (a+b+c+d).
= (25+32) / (30+40+5+10)
= 82 %

True Positive/Recall (TP) = d / (c+d)
= 40 / (5+40)
= 88.8 %

False Positive (FP) = b / (a+b)
= 10 / (30+10)
= 25%

True Negative (TN) = a / (a+b)
= 30 / (30+10)
= 75%

False Negative (FN) = c / (c+d)
= 5 / (45)
= 11.12%

Precision (P) = d / (b+d)
= 40 / (10 + 40)
= 80 %

F-Measure (FM) = (2P*TP) / (P+TP)
= (2*80*88.8) / (80+88.8)
= 84.20%

Hence, in this manner we can apply the method of confusion matrix to test our results. We can calculate the accuracy and precision of our system using confusion matrix.

VI. CONCLUSIONS

The results that we have obtained are highly reliable as the papers are extracted from a very trusted website of medical world. On these obtained data or papers various processing task is done such as downloading HTML pages,converting to zip then to XML and ultimately to word format .And later on Natural Language Processing is done to obtain relevant information and to obtain the results as per the semantic relations we use machine learning algorithms for that.
Hence we have observed that all the hectic work has been done by the machine itself no human intervention is required.

ACKNOWLEDGMENT

First and foremost I offer my sincerest gratitude to my college, JSPM's BSIOTR and my department of Computer Science and Engineering which has provided the support and equipment I have needed to complete my work. I thank the Medline repository which provides the datasets for preceding my work. I extend my heartfelt gratitude to my guide, Prof. A.G.Baviskar and Prof.Bharat Burghate, who has supported me throughout our research with their patience and knowledge.

REFERENCES

[1] Russell S, Norvig P, (1994), Artificial Intelligence: A Modern Approach. New Jersey ,Prentice-Hall, Inc.,A Simon & Schuster Company, Englewood Cliffs, New Jersey 07632, ISBN 0-13-103805-2.

[2] O. Frunza, D.Inkpen and T. Tran,"Machine learning approach for identifying disease-treatment relations in short text", IEEE Computer Society, June 2011, pp-no. 1041-4347.

[3] L.Hunter,B.Cohen," Perspective Biomedical Language Processing :What's Beyond PubMed? ", Molecular Cell 21, March 2006, pp-no. 589–594

[4] H.Yana, Y.Jiangb, J.Zhenge, C.Pengc,Q.Lid ,"A multilayer perceptron-based medical decision support system for heart disease diagnosis", Expert Systems and Applications 30, May 2006,pp-no. 272–281

[5] O. Frunza and D. Inkpen, "Textual Information in Predicting Functional Properties of the Genes," Proc. Workshop Current Trends in Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Linguistics (ACL '08), 2008.

[6] B. Rosario and M.A. Hearst, "Semantic Relations in Bioscience Text," Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, vol. 430, 2004.

[7] Jeff Pasternack, Don Roth "Extracting Article Text From Web With Maximum Subsequence Segmentation", WWW 2009 MADRID

[8] Abdur Rehman, Haroon.A.Babri, Mehreen saeed," Feature Extraction Algorithm For Classification Of Text Document",ICCIT 2012

