Identification and Classification of Disease and its Treatment using MEDLINE Literature

Aditi A. Ghive
(P. G. Student)
Department of Computer Science
R. C. Patel Institute of Technology
Shirpur, India

D. R. Patil
(Associate Professor)
Department of Computer Science
R. C. Patel Institute of Technology
Shirpur, India

ABSTRACT
In this era of information technology everyone needs information at the fingertip. With the development of IT various applications are developed in the medical field. This will make revolution in the medical field. Also, it will help doctors to know the recent developments and research carried out in the world. This will update the information of each and every doctor using such applications. Engineers are designing and developing best MEDLINE applications which are beneficial not only to doctors but also to the common peoples. All applications are user friendly. This will strengthen the present healthcare system.

This paper introduces novel approach for identification and classification of disease and its treatment. We have also discusses all possible identification and classification approaches in this paper. The proposed approach uses the MEDLINE literature to extract all important biomedical information related to the respective disease and finding its treatment. Natural language processing (NLP) and data mining techniques are used to automatically extract information from MEDLINE literature. A software system automatically identifies disease related terms from the MEDLINE abstracts.

Index Terms
MEDLINE application, biomedical information, heath care system

1. INTRODUCTION
With the development of IT sector information related to all important process in the world are recorded in the form of text, graphs, pictures and tables. It is not necessary that all recorded information is useful every time still the record is maintained. Such a recorded data is available in the biomedical field. This data is recorded for a long period, but it is not used in a more extend till now. But, with the development in the computer domain this biomedical data is now used for automating and improving the medical domain. Nowadays some software applications are being developed which uses this recorded biomedical information to find the diseases and their treatments. By using different tools in computer domain one can make abandon used of this information for developing more and more applications which are beneficial for the society.

In this era life is hectic and due to busy schedules, people want each and everything to go with a good flow. Everyone cares for health and wants to be always fit and good health. People want quick access to reliable information. The traditional healthcare system needs to be modernized. Diagnosis of various diseases is now a day's carried out by an advance healthcare system which involves gathering clinical information, extraction of useful data, identification of diseases and finding relations for treating various diseases. Also the researches in medical domain and pharmaceutical field can be made available to everyone. Recent developments of drugs on various diseases can easily available to doctors over the globe. Some diseases are new to the doctors, so, treating patients with new, unknown diseases will be possible by developing a computer application which can use research abstracts and finds a relation of disease for proper treatment. This application will be an important approach for modernization of the traditional health care system. Various algorithms can be developed which can identify and classify the medical information in text form.

2. RELATED WORK
T. Sakthimurugan et al [1] this work shows on retrieval of updated, accurate and relevant information from Medline datasets using Machine Learning approach. This system uses keyword searching algorithm for extracting relevant information from Medline datasets and *K*-Nearest Neighbor algorithm (KNN) to get the relation between disease and treatment. Since, improvement of patient care achieved effectively.

C. B. Sivaparthipan et al [2] shows unique search engine which makes use of a neural network that is adequately trained with a number of journals that have been published in the medical field. In this approach, the patient searches for the necessary treatment/medicine/expert advice for a particular disease and the search engine, in turn, provides the appropriate treatment/medicine/expert advice is presented in this paper

Vikas Chaurasia et al [3] proposes an approach for research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The research work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, thirteen attributes were involved in predicting the heart disease. Three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes.

Won Kim et al [4] presents research methodologies to show a supervised learning approach for identifying high quality phrases. It gets a set of known well-formed useful phrases from an existing source and labels these phrases as positive. Then it extracts from MEDLINE a large set of multiword strings that do not contain stop words or punctuation. We believe this unlabeled set contains many well-formed phrases. Our goal is to identify these additional high quality phrases.

3. PROBLEM DEFINITION

In this section, we have explained about the problem faced by the existing disease related treatments identification system. First problem is that once the disease is identified the extracted result gives us multiple treatments for a single disease. Out of this multiple treatments the system doesn't identify the best treatment. Second problem is that the data set used for this identification and classification purpose is a standard data set which needs to be updated time to time. To overcome this the paper presents a new approach which uses the data set as MEDLINE abstracts and extracts disease related information from it. All information is extracted using the keyword searching algorithm and Classifiers are used to classify the semantic relation that exists between disease and treatment. Data mining technique can be applied to find the best treatment.

4. METHODOLOGY

In this era of the system is to generate a Health information recording and clinical data reporting system which will give the doctor an easy and fast access to patient diagnoses, allergies, and lab test reports and results that enable time-efficient and better medical decisions; Decision support .It is the ability to collect and use quality medical data for precise decisions in the workflow of healthcare system, it also helps to obtain treatments information that are to collect specific health needs rapid access to information that is focused on certain topics. In this section, two tasks are present in this system that provides the design view of an information system framework which is capable to identify and extract health care information from medical publication. In first task is identification and extraction of sentences that mention on diseases and treatments topics for a given symptoms. The second task is to perform a classification of these sentences according to the semantic relations that exists between diseases and treatments.

The NLP and ML based techniques are used to solve the two proposed tasks. It identifies informative sentences that contain information about diseases and treatments and semantic relations between them versus non informative sentences. This allows us to see how well the natural language processing techniques and Machine learning based techniques can cope with the task of identifying informative sentences. NLP provides a means for analysis text. The role of NLP is to make computers analyze and understand the languages that humans use naturally. NLP can perform interaction between Computers-Humans.

These two tasks can be combined in pipeline approach. This proposed pipeline approach first performs task1 and then performs task2 so that it can give only informative sentences based on three relations. The logic behind using this pipeline approach is to identify the best model to identify and extract the reliable healthcare information. NLP perform:-

1. Sentence detection,
2. Tokenization,
3. Pos-tagging,
4. Parsing,
5. Named-entity detection.

4.1 Classification Algorithms and Data Representation

The model must be reliable to identify informative sentences and discriminating disease- treatment semantic relations. While working with ML techniques two challenges are encountered first one is to find the most suitable model for prediction. ML uses various predictive algorithms. The second one is to find a good data representation. These two challenges are addressed by trying various predictive algorithms and by using various textual representation techniques suitable for the task. We use a set of six representative models as classification algorithms: decision-based models (Decision trees), probabilistic models (Naive Bayes (NB) and Complement Naive Bayes (CNB), adaptive learning (Ada- Boost), a linear classifier (support vector machine with polynomial kernel), and a classifier that always predicts the majority class in the training data. These classifiers are used to work on long text and short texts and to learn more algorithms. Probabilistic models are used in automatic text classification tasks. Decision trees based on decision models are used in short texts. Adaptive learning algorithm is used in unbalanced data sets. The SVM based algorithm is involved in text classification technique. Following data sets are used for the above two tasks.

4.2 Bag-of-Words Representation

The bag-of-words (BOW) representation is generally used for text classification tasks. In this representation features are chosen among the words that are present in the training data. In order to identify the most suitable words as features selection techniques are used. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for Bag-of-Words (BOW) representation are: binary feature values or frequency feature values. The binary feature value is the value of a feature can be either 0 or 1, where 1 represents the act that the feature is present in the instance and 0 otherwise. The frequency feature value is the value of the feature is the number of times it appears in an instance or 0 if it did not appear. Here we use frequency feature value. There is not much difference between binary feature values and frequency feature values because there are only twenty words in each sentence of short texts. It is advantageous to use frequency feature values as the feature's value will be greater than other features since it captures the number of features appeared once in a sentence.

4.3 Natural Language Processing and Biomedical Concepts Representation

The second type of representation is based on syntactic information: noun-phrases, verb-phrases and biomedical concepts identified in the sentences. In order to extract this type of information the Genia11 tagger tool is used. It analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. It is specifically used for biomedical text such as Medline abstracts. The following preprocessing steps are applied in order to identify the final set of features to be used for classification: removing features that contain punctuations and considering lemma-based forms. Lemma is used because there are lots of words that has plural forms. Lemmatized form gives the base form of the word. Lemma forms remove the problem of representing only a few features in short text forms.

4.4 Medical Concepts (UMLS) Representation

We use the Unified Medical Language system (UMLS). UMLS is a knowledge source developed at the US National Library of Medicine (NLM). It contains a met thesaurus, a semantic network, and the specialist lexicon for the biomedical domain. It gives the relation between various concepts. NLM has

created a set of tools that will allow easier access to the useful information. UMLS contains over a million medical concepts and over five million concept names. Each concept is assigned to one semantic type from the semantic between concepts. It had created new tool called MetaMap which maps text to healthcare concepts in UMLS.

This text is processed through the entire data set and finally it provides a ranking list of all possible concepts for particular noun - phrases. For each noun phrase in text various noun-phrases are generated. For each variant noun-phrases UMLS met thesaurus are obtained and evaluated. The obtained phrase is compared with actual phrase using fit function. Fit function measures the text overlap between obtained concepts and actual phrase. The best of the candidates are then arranged according to the decrementing value of the fitness function.

5. PROPOSED ARCHITECTURE

The overall architecture of this application is as shown in figure.

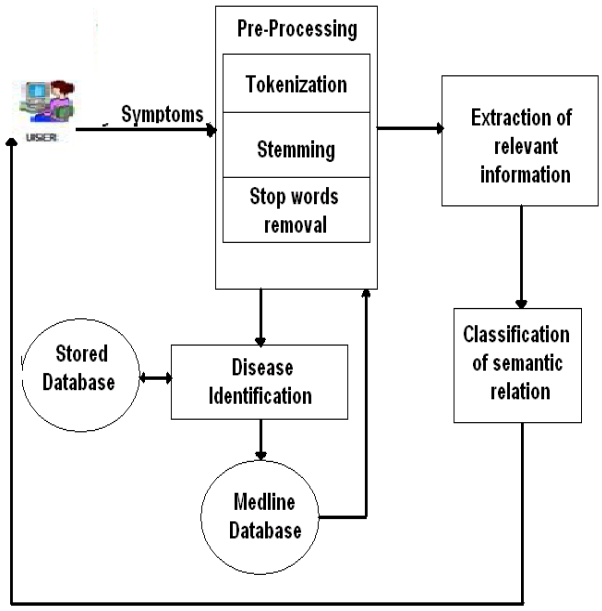

Fig 1. Architecture of proposed system

In the proposed work user will search for the disease summary (disease and treatment related information) by giving symptoms as a query in the search engine. These symptoms are preprocessed to make the further process easier to find the semantic keyword which helps to identify the disease quickly.

Then the semantic keyword is matched with the stored medical input database to identify the exact disease related to that keyword. Once the disease is identified, it is sent to the medical database to extract the articles pertaining to that disease. Now all these extracted articles are preprocessed to make the further process easier and more efficient. The preprocessing process involves being tokenization, removal of stop words and stemming. Followed by that, relevant information is extracted using the keyword searching algorithm. Finally, the relevant keywords are classified into four class labels, namely cure, no cure, prevent, and side effect.

6. CONCLUSION

The overall study carried out in this paper leads to the conclusion that computer systems are gaining more importance in the biomedical automation. Also it shows how Machine learning and Natural language processing techniques are helpful in MEDLINE applications. Traditional healthcare systems will get modernized with the development of applications related to the biomedical field. The application developed will help doctors to treat various diseases easily and the research work over the globe will be at their fingertips. In the future, development of more biomedical applications using different algorithms and different software tools which will move a one step ahead in biomedical automation. Also, comparative analysis can be done with other classification algorithms in order to provide better performance.

7. REFERENCES

[1] Won Kim and Lana Yeganova," Identifying well-formed biomedical phrases in MEDLINE text", Journal of Biomedical Informatics, pp. 1035–1041, 2012.

[2] Vikas Chaurasia and Saurabh Pal," Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 4, pp. 56-66, 2013.

[3] C. B. Sivaparthipan and V.Dheepa," Neural Network Based Online Learning and Identifying Disease-Treatment", International Conference on Computer Science and Information Technology (ICCSIT'2011), Dec. 2011

[4] T.Sakthimurugan1 and S.Poonkuzhali", An Effective Retrieval of Medical Records using Data Mining Techniques", International Journal of Pharmaceutical science and Health,2012

[5] O. Frunza, D. Inkpen and T. Tran, "A Machine Learning Approach for Identifying Disease Treatment Relations in Short Texts", *IEEE* Transactions on Knowledge and Data *Engineering*, Vol. 23, No. 6, June 2011

[6] Rosario and Hearst, "Machine Learning (ML) Approach for Identifying Disease -Treatment Relations in Short Texts", IEEE Transactions on Knowledge and Data Engineering, 2011