MLA for Identifying Disease-Treatment

[1]Mr. Kunal M. Shirkande, [2] Ms. Pallavi B. Lamkane, [3]Prof.Prajakta A. Satarkar

***Abstract*—** The Machine Learning(ML) is just recently has become a reliable tool in the medical domain. The experimental domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein interaction, extraction of medical knowledge, and for overall patient management care. The Machine Learning is pictured as a tool by which computer-based systems can be unified in the healthcare field in order to get a better, more efficient medical care.

This paper describes a Machine Learning based methodology for building an application and scattering healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies relation between meanings that exist between diseases and treatments.

1. INTRODUCTION

1.1 Background and basics:

People are more concerned about their health than ever before. From their busy schedules they want each and everything to go in a good health. People want Fast access to reliable information and in a manner that is suitable to their habits and work. Medical field has grown to such an extent that the people practicing medicine should not only have experience but also information about latest discoveries. Electronic Health Record[HER] is becoming a standard in healthcare domain. Websites such as Google Health[9] and Microsoft Health Vault[10] make people to care deeply about their health.

To get the quality medical data for taking proper medical decisions. For this purpose we need a better, more efficient and reliable access to information.

If we think about researches people are searching the web to get informed regularly about their health. In medical domain the most used source of information is MEDLINE[12]. MEDLINE is database where all the research related new innovations come and enter at a high rate. From the busy schedules the experts are not getting time to read thousands of articles/resources therefore, there is a need to build a tool that will be enough to satisfy the need of the people.

Our objective is to work with ML and Neural Language Processing (NLP) is that the task of identifying and spreading reliable healthcare information, that will becomes easy and beneficial for the people. A hierarchical approach is used for performing the two tasks: The first is to identify and remove the sentences with no meaning and then second is to categorize the rest of the sentences by the relation of interest. By this a considerable improvement is shown getting the information.

Sentence selection - It identifies sentences from Medline published abstracts that talk about diseases and treatments. The task is same like to scan the sentences from the medicine abstracts which contains the disease treatment information which will be useful to the user.

Relation identification - it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first).

1.1.1 Business Context:

The proposed applications rely on personalizing the mass-media experience by matching ambient audio statistics. The applications provide the viewer with personalized layers of information, new avenues for social interaction, real time indications on show popularity and the ability to maintain a library of the favorite content through a virtual recording service. These personalization applications can be modified in order to provide the degree of privacy each viewer feels comfortable with. Similarly, the applications can vary according to viewer-specific technical constraints, such as bandwidth and CPU time.

1.2 Literature survey

The existing system:

1. In order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information.
2. In the medical domain, the richest and most used source of information is Medline, database of extensive life science published articles.
3. All research discoveries come and enter the repository at high rate (Hunter and Cohen, making the process of identifying and disseminating reliable information a very difficult task.
4. One task is automatically identifying Sentences published in medical abstracts (Medline) as containing or not information about disease Treatments and automatically identifying semantic relations that exist between disease and Treatments.
5. Use Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and the relation discrimination.

6. Their representation techniques are based on words 1in context, part of speech information, phrases, and a medical lexical ontology Mesh terms.

Drawbacks in existing system are:

a. People want Fast access to reliable information and in a manner that is suitable to their habits and workflow.

b. Reliable information is not received always by the system.

1.3 Relevance

This system is not the first of its kind, there does exist a system that has been designed to serve the same purpose. All of which are quite relevant to each other, in the context of the purpose.

We have done our best, to design an application which will be different from others and still serve our purpose. However, each system is different in the way it appears and functions and these variations depend upon their developer.

1.4 Scope of statement

HER are becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are3: Health information recording and immediate access to patient diagnoses and lab test results that enable better and time-efficient medical decisions; Medication management that is a rapid access to information regarding potential or highly effective drug reactions, immunizations, supplies, etc; Decision support—the ability to capture and use quality medical data for decisions in the workflow of healthcare; and Obtain treatments that are tailored to specific health needs—rapid access to information that is focused on certain topics. The work that we present in this paper is focused on two tasks: automatically identifying sentences published in medical abstracts (Medline) as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments, as expressed in these texts. The second task is focused on three semantic relations: Cure, Prevent, and Side Effect.

This describes testing plan for the developed system which identifies sentences from medical abstract by looking from already available abstracts we will call it as from different Medline database rather than the user requirments as defined. The purpose of this acceptance test is to make sure that the system developed using machine learning gives 100% correct/satisfied result as per the user requirement. This test executed in the testing phase of the project.

2. SOFTWARE ENGINEERING MODEL USED I.E. INCREMENTAL MODEL:

2.1 COMMUNICATION:

The software development starts with communication between customer and developer. In this phase we communicated with following principles of communication phase .We prepared before the communication i.e. we decide agenda of the meeting for concentrating on data mining. The team worked on all possible outcomes of project by studying the requirement from the user's point of view by considering what is the actual needed, what is output, output format of system.

2.2 PLANNING:

It includes complete estimation and scheduling and risk analysis. In this phase we planned about when estimated to release the software, cost estimation, risk in the project regarding application etc. Finally in this phase we estimated cost of project including all expenditure of software according to user deadline with his participation.

2.3 MODELING:

It includes detail requirement analysis and project design, flowchart shows complete pictorial flow of program where the algorithm is step by step solution of problem .We analyze the requirement of the user according to that we draw the block diagram of the system. That is nothing but behavioral structure of system using UML 2.0 i.e. Class Diagram .Interaction Diagram and Use Case report.

2.4 CONSTRUCTION:

It includes Coding and Testing Steps:

a. Coding:

Design details are implemented using appropriate programming language. For coding we have choose the JAVA platform

b. Testing:

Testing is carried out by analyzing the system i.e., we first develop the prototype of the system and step by step find out input and output errors such as interface errors. Data structure errors, initialization errors etc. Therefore here Black Box testing strategy is useful.

2.5 DEPLOYMENT:

It includes software delivery, support and feedback from customer. If customer suggest some corrections, or demands additional changes then changes are required for such coercions and enhancement .Thus each iteration around the spiral leads to more completed version of software.

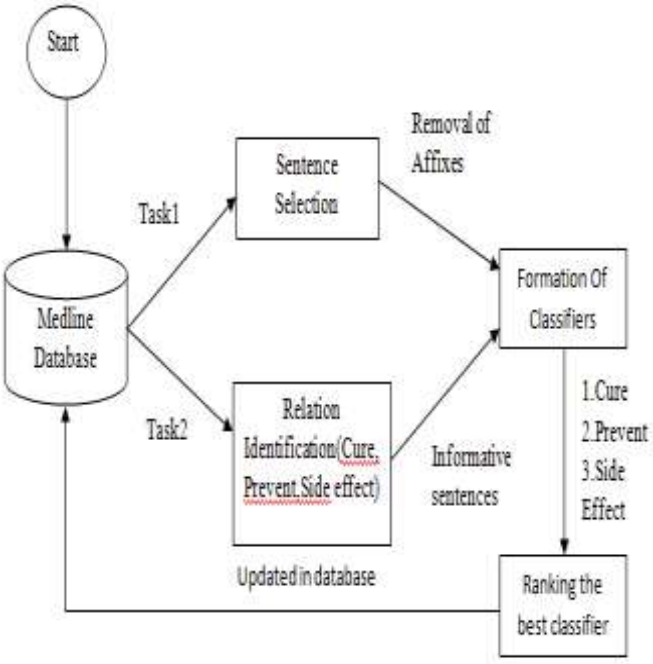
3. PLANNING AND MANAGEMENT

The following are used to process to estimate the total effort required for the project:

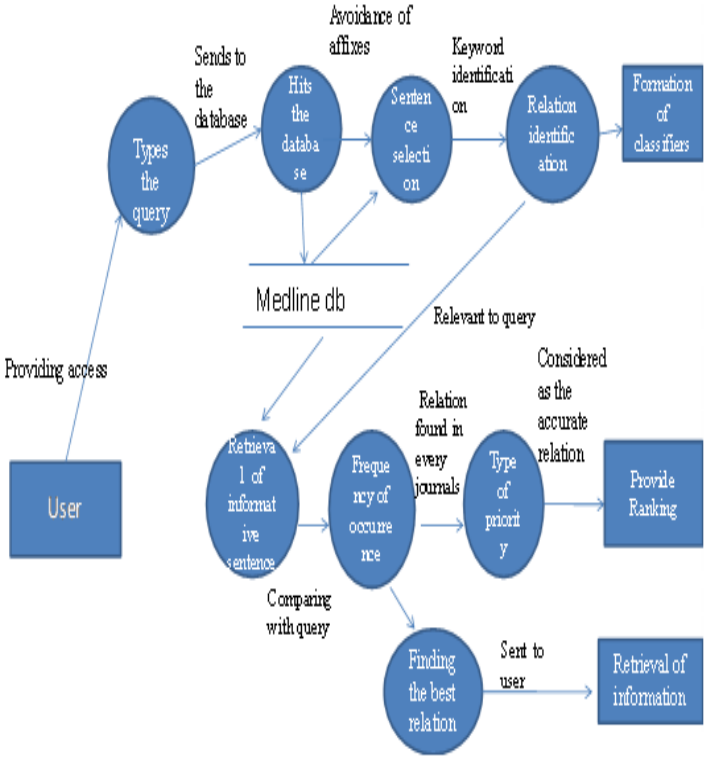
- 1Time and understand the work at a low level of detail time estimate was 10%.
- Add specialist resource hours for learning new technology were 20 % resources.
- For instance training specialists, procurement, legal, administrative, etc.
- Add project management time and proactively manage a project. 15% of the effort hours for project management are used.
- Contingency hours are used to reflect the uncertainty or risk associated with the estimate of 5%.
- Coding and Testing estimation is about 40 %.
- Review and adjust as necessary was about 5%.Document all assumptions was about 5-7%.

Collaboration defines an interaction and is a society of roles and other elements that work together to provide some cooperative behavior. So collaborations have structural as well as behavioral, dimensions. These collaborations represent the implementation of patterns that make up a system

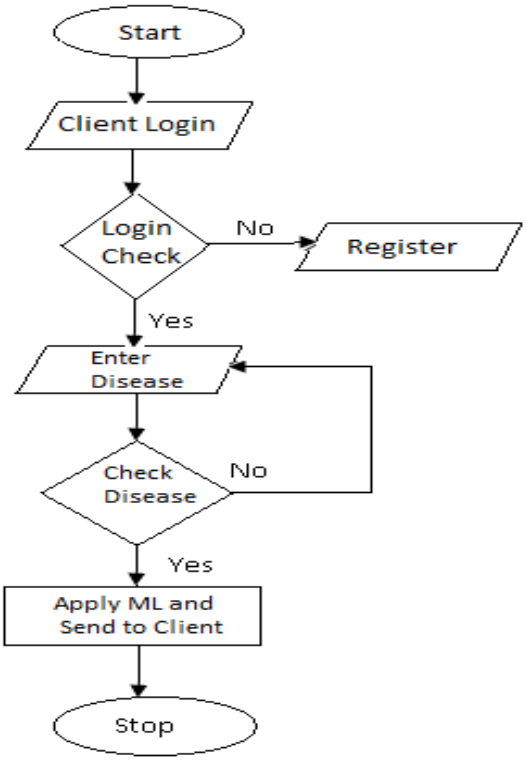
3.1 Transform flow & transition flow



3.2 Data Flow Diagrams:



3.3 Algorithms/flow charts



4. TOOLS USED

a. Weka tool:

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic purpose of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

b. Stanford Pos tagger tool :

This is a tool used for tagging the words in the sentence to check whether a noun, a pronoun, an adjective etc. Noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. In order to extract this type of information, we used the Stanford Pos-tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags.

For the purpose of identifying and disseminating healthcare information reliably we use naive bayes classification algorithm which is a part of weka tool. The naive bayes classifier with combinations of various representation features is the one that obtains the best result for all relations.

FUTURE WORK

This approach is very useful for everyone as it gives information only of the area of interest. The task is divided into two tasks The first task that we tackle in this paper is a task that has applications in information retrieval, information extraction, and text summarization. We identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts.

The second task that we address can be viewed as a task that could benefit from solving the first task first. In this study, we have focused on three semantic relations between diseases and treatments. Our work shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task.

CONCLUSION

This study is related to a particular field but the future scope of the paper lies in the fact that this can be extended to the information on the web. Identifying and classifying medical-related information on the web is a challenge that can bring valuable information to the research community and also to the end user. We also consider as potential future work.

REGERENCES

[1] Oana Frunza, Diana Inkpen, and Thomas Tran " A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" vol. 23, 2011.

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

[3] M. Craven, “Learning to Extract Relations from Medline,” Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.
[4] S. Ray and M. Craven, “Representing Sentence Structure in Hidden Markov Models for Information Extraction,” Proc. Int’l Joint Conf. Artificial Intelligence (IJCAI ’01), 2001.
[5] P. Srinivasan and T. Rindflesch, “Exploring Text Mining from Medline,” Proc. Am. Medical Informatics Assoc. (AMIA) Symp., 2002.
[6] R. Bunescu, R. Mooney, Y. Weiss, B. Scho¨ lkopf, and J. Platt,“Subsequence Kernels for Relation Extraction,” Advances in Neural Information Processing Systems, vol.18, pp. 171-178, 20
[7] R. Kohavi and F. Provost, “Glossary of Terms,” Machine Learning, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, pp. 271-274, 1998.
[8] http://nlp.stanford.edu/software/tagger.shtml.
[9] Google health report,Microsoft Health Vault, http://healthvault.com Health care tracker, http://healthcaretracker.wordpress.com/
[10] Medline Database, http://www.proquest.com/en US/catalogs/databases/detail/medline_ft.shtml
[11] Medical Subject Headings, http://www.nlm.nih.gov/mesh/meshhome.html.
[12] Weka tool, http://www.cs.waikato.ac.nz/ml/weka/.
[13] List of Stop words, http://www.site.uottawa.ca/~diana/csi5180/StopWords.