DOCUMENT EXTRACTION FOR MEDICAL DISEASE TREATMENT USING DATAMINING

Ms.P.Sathyasutha (AP/CSE)	Gnanamani college of Technology,Namakkal
C.Seenivasan (Final/CSE)	
K.Kiran (Final/CSE)	
T.Uvantha (Final/CSE)	

***Abstract*— In our proposed system is identifying reliable information in the medical domain stand as building blocks for a healthcare system that is up-to-date with the latest discoveries. By using the tools such as NLP, ML techniques. In this research, focus on diseases and treatment information, and the relation that exists between these two entities. The main goal of this research is to identify the disease name with the symptoms specified and extract the sentence from the article and get the Relation that exists between Disease-Treatment and classify the information into cure, prevent, side effect to the user.This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document**.

1. INTRODUCTION

People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been, the medicine that is practiced today is an EBM in which medical expertise is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as Google Health and Microsoft HealthVault are reasons and facts that make people more powerful when it comes to healthcare knowledge and management. The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. EHR are becoming the standard in the healthcare domain. Researches and studies show that 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 second task is focused on three semantic relations: Cure, Prevent, and Side Effect. The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. People want fast access to reliable information and in a manner that is suitable to their habits and workflow. Medical care related information (e.g., published articles, clinical trials, news, etc.) is a source of power for both healthcare providers and laypeople. Studies reveal that people are searching the web and read medical related information in order to be informed about their health. Ginsberg et al. show how a new outbreak of the influenza virus can be detected from search engine query data.

Healthcare providers need to be up-to-date with all new discoveries about a certain treatment, in order to identify if it might have side effects for certain types of patients. We envision the potential and value of the findings of our work as guidelines for the performance of a framework that is capable to find relevant information about diseases and treatments in a medical domain repository. The results that we obtained show that it is a realistic scenario to use NLP and ML techniques to build a tool, similar to an RSS feed, capable to identify and disseminate textual information related to diseases and treatments. Therefore, this study is aimed at designing and examining various representation techniques in combination with various learning methods to identify and extract biomedical relations from literature. The contributions that we bring with our work stand in the fact that we present an extensive study of various ML algorithms and textual representations for classifying short medical texts and identifying semantic relations between two medical entities: diseases and treatments. From an ML point of view, we show that in short texts when identifying semantic relations between diseases and treatments a substantial improvement in results is obtained when using a hierarchical way of approaching the task (a pipeline of two tasks). It is better to identify and eliminate first the sentences that do not contain relevant information, and then classify the rest of the sentences by the relations of interest, instead of doing

everything in one step by classifying sentences into one of the relations of interest plus the extra class of uninformative sentences.

2. RELATED WORK

In order to embrace the views that the EHR system has, the potential benefits of having an EHR system are: Health information recording and clinical data repositories immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions; Medication management—rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc;

World need better, faster, and more reliable access to information. In the medical domain, the richest and most used source of information is Medline database of extensive life science published articles. All research discoveries come and enter the repository at high rate , making the process of identifying and disseminating reliable information a very difficult task. one task is automatically identifying sentences published in [1] medical abstracts (Medline) as containing or not information about diseases and Treatments and automatically identifying semantic relations that exist between diseases and treatments.

Bunescu R, Mooney R et. Al [2] proposed supervised machine learning methods have been used with great success in this task but they tend to suffer from data sparseness because of their restriction to obtain knowledge from limited amount of labeled data. We use feature coupling generalization (FCG) – a recently proposed semi-supervised learning strategy – to learn an enriched representation of local contexts in sentences from 47 million unlabeled examples and investigate the performance of the new features on AIMED corpus. The approach provides theoretically well-founded solutions to the problems of under- and over fitting. Secondly it allows learning from structured data, and has been empirically demonstrated to yield high predictive performance on a wide range of application domains. However, this approach is critical & challenging problem to develop user friendly natural language to computer interface.

M. Craven [3] examined the problem of distinguishing among seven relation types that can occur between the entities "treatment" and "disease" in bioscience text, and the problem of identifying such entities. They compare five generative graphical models and a neural network, using lexical, syntactic, and semantic features, finding that the latter help achieve high classification accuracy. The scheme was the correct management of word position information, which may be critical in identifying certain relationships. In this approach that facilitates the automatic recognition of relationships defined between two different concepts in text. However, this task involves the manual tuning of domain-dependent linguistic knowledge such as terminological dictionaries, domain specific lexico-semantics, and extraction patterns, and so on.

Razvan C et. al says that a new method for joint entity and relation extraction using a graph we call a “card-pyramid.” This graph compactly encodes all possible entities and relations in a sentence, reducing the task of their joint extraction to jointly labeling its nodes. We give an efficient labeling algorithm that is analogous to parsing using dynamic programming. These approaches assume that relations only exist within document, and classify them independently without considering dependencies between entities. However, this assumption does not hold in practice, and ignoring dependencies between entities may lead to reduced performance. Implicit relations can hardly be discovered in these models since they generally exist in cross document and they are only implied by the text. The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: subcellularlocation (Craven, [4]), gene-disorder association (Ray and Craven, [5]), and diseases and drugs (Srinivasan and Rindflesch, [6]). In these works , tasks often entail identification of relations between entitiesthat co-occur in the same sentence.

Heart disease is the leading cause of death all over the world. They have identifies gaps in the research on heart disease diagnosis and treatment and proposes a model to systematically close those gaps to discover if applying data mining techniques to heart disease treatment data can provide as reliable performances that achieved in diagnosing heart disease[14]. Various learning algorithms have been used for the statistical learning approach with kernel methods being the popular ones applied to Medline abstracts (Li et al. [13]).There are three major approaches used in extract in relations between entities: co-occurrences analysis, rule based approaches, and statistical methods. The co-occurrences methods are mostly based only on lexical knowledge and words in context, and even though they tend to obtain good levels of recall, their precision is low. Good representative examples of work on Medline abstracts include Jenssen et al. [7] and Stapley and Benoit [8]. Syntactic rule-based relation extraction systems are complex systems based on additional tools used to assign part of speech tags or to extract syntactic parse trees. It is known that in the biomedical literature such tools are not yet at the state-of-the-art level as they are for general English texts, and therefore their performance on sentences is

not always the best in Bunescu et al.[8]. Representative works on syntactic rule-based approaches for relation extraction in Medline abstracts and full-text articles are presented by Thomas et al. [9], Yakushiji et al. [10], Leroy et al. [11] and OpenDMAP described in Hunter et al. [12]. Even though the syntactic information is the result of tools that are not 100 percent accurate, success stories with these types of systems have been encountered in the biomedical domain.

From the literature point of view drawback of existing systems are: people cannot get the direct information about the disease because it displays history of disease at first. There is no reliable information.

3. METHODOLOGY

Proposed system consists of the Client Interface, Identify the Disease, Sentence Extraction and Classification.

The tasks that are available in the proposed system:

1. First task is automatically identifying sentences published in medical abstracts.
2. The second task is focused on three semantic relations: Cure, Prevent, and Side effect.

Client Interface: In this Module, develop a user page using Graphical User Interface which will be a media to connect User and Media Database and login screen where user can input his/her user name, password and password will check in database, if that will be a valid username and password then he/she can access the database . **Indentify the Disease: In** this module user is going to give the symptoms as an input and get the desired disease name. In this it will search as semantic word and give the output to the user. **Sentence Splitting:** n this stage user has to enter the symptom in a short text. Then taking out the human errors from the sentence typed by the user like comma, dot with space and without space.

Semantic Extraction: After removing the Human errors from the sentence we have to get the semantic words it means if user typed some wrong words then it will correct it with semantic words that is maintained in the database. **Removing unwanted words:** In this module we are concentrating on the unwanted words from the sentence typed by the user. It will be very tough task to implement with the sentence that talked about disease treatment relation. **Disease identification:** After eliminating words we are going to find the correct disease with High Priority and Low Priority. **Sentence Extraction:** In this module user to provide input as a disease. That means relevant to our article and extract the informative sentence from database. **Classification**: After extracting Sentence from the database we have to classify the relation for the Cure, Prevent and Side Effect. For classification navie baysean algorithms are used.

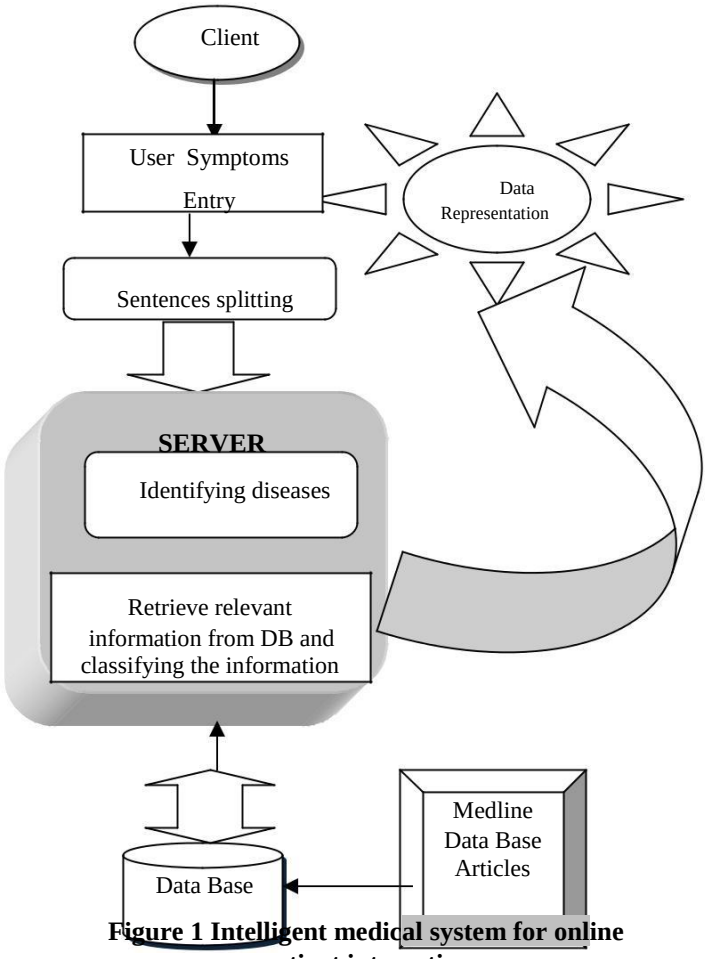

Figure 1 Intelligent medical system for online patient interaction

4. RESULT AND DISCUSSION

For example user can Enter the symptoms like "I have a head ache stomach pain". In this sentences splitting module splitting the sentences with the space and removable of the human errors. Example of the input system and output of the splitting task is shown in Table 1. In table 2 interprets the semantic extraction of sentence. Table3 shows the removing unwanted word in the symptoms. Table 4 presents user Symptoms after Removing human errors and semantic extraction. Figure 4 and 5 shows the output of the classification of automated medical system.

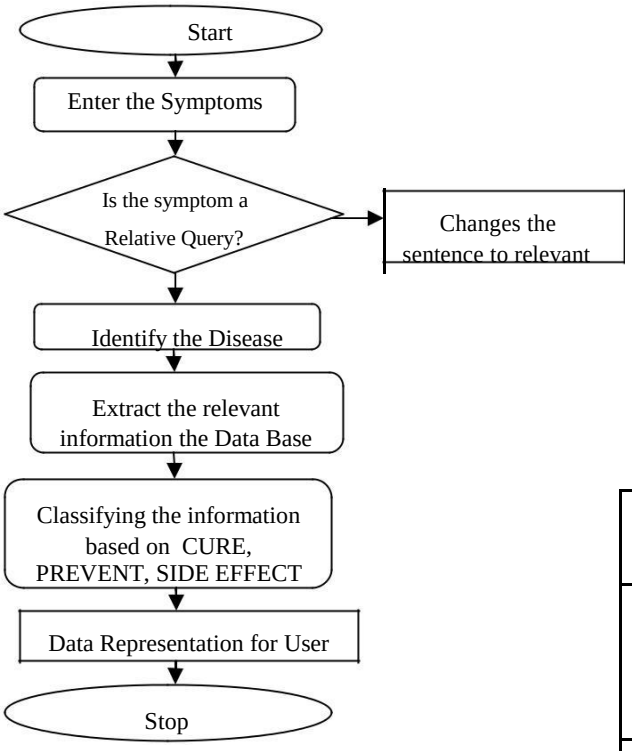

Figure 2 Flow chart for medical disease classification

Table 1 Input of the medical system and splitting sentence output

	Input of the Symptoms	Output of the splitting sentence
1	I have a headache and stomach pain	I Have A Headache and Stomach pain
2	I have cheast pain	I have cheast pain
3	I have high fever	I Have high fever

Table 2 Semantic Extraction Task

	Sentence Splitting	Semantic Extraction
1	I have a fevar	I Have A Fever
2	I have a stamach pain	I have a stomach pain

Table 3 Removing unwanted words

	Sentence with unwanted words	Output of the Removing unwanted words
1	I have a fever	fever
2	I have cheast pain	cheast pain
3	I Have high fever	high fever

Table 4 User Symptoms after Removing human errors and semantic extraction

	User Symptoms after Removing human errors and semantic extraction	Output of the Removing unwanted words
1	Chest pain	High Priority:Heart Attack Low Priority:Myo Cardial Infraction
2	Head ache	High Priority:Brain tumer Low Priority:fever

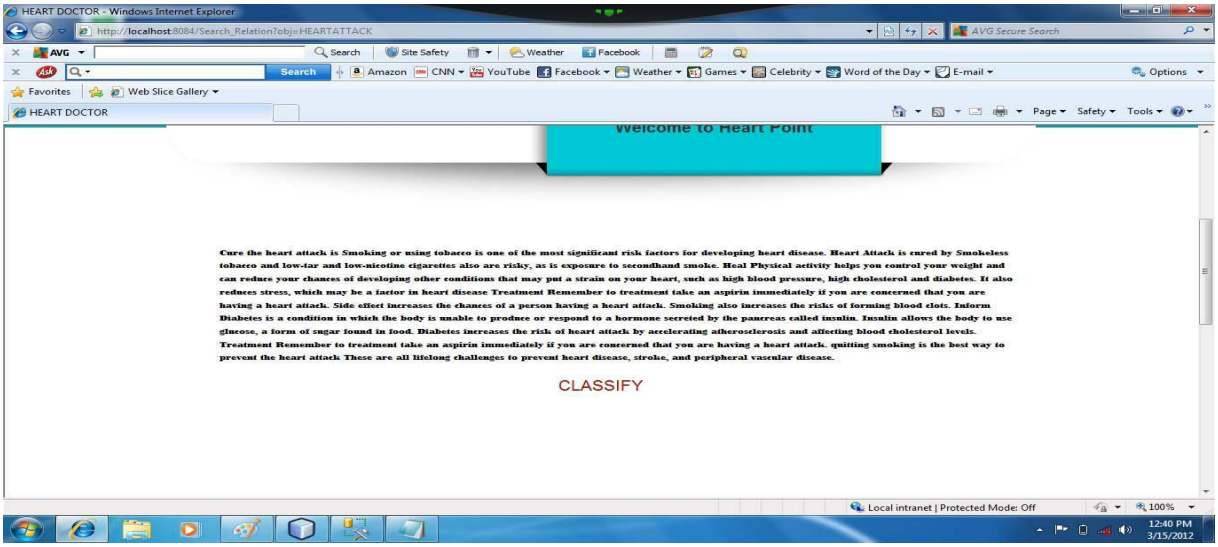

Figure :3 Classification information

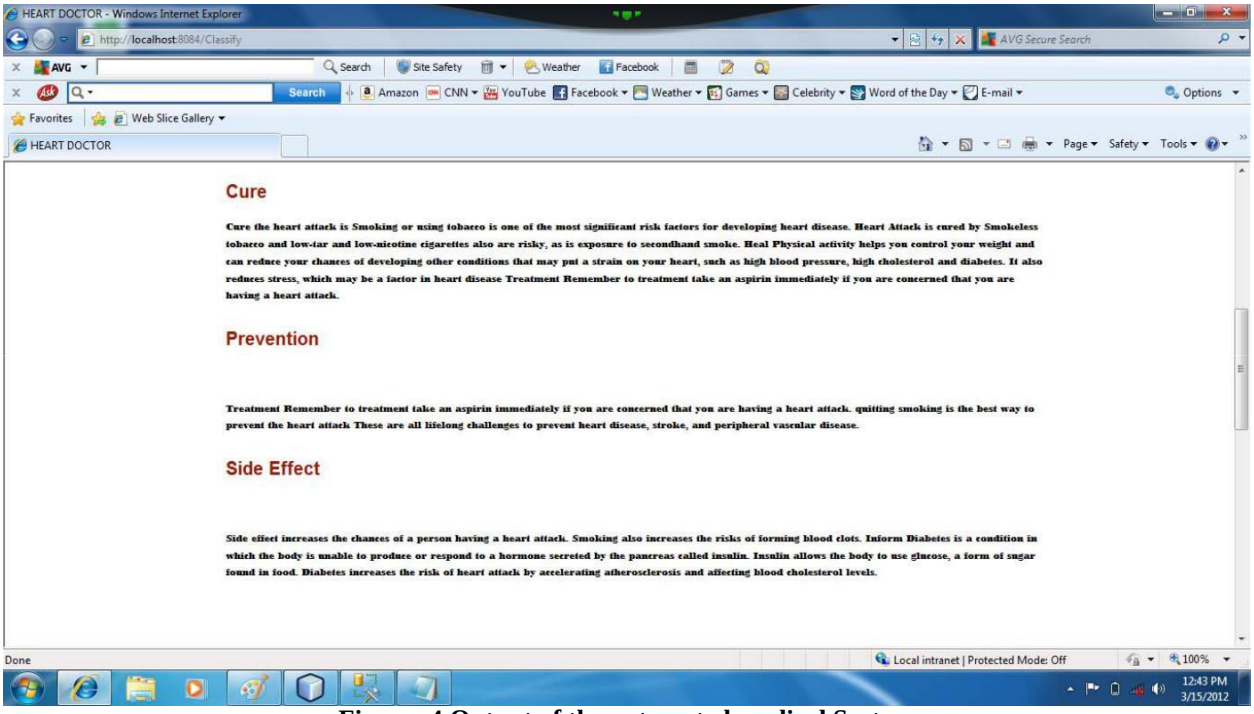

Figure :4 Output of the automated medical System

5. CONCLUSION AND FUTURE ENHANCEMENT

The conclusions of our study suggest that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results. The source data is from the wed and identifying then classifying the data on the web is a challenge but bringing valuable information in future it has the capability in framework model.

REFERENCES

1. S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, A Natural Language Processing Engine for MEDLINE Abstracts," Bioinformatics, vol. 19, no. 13, pp. 1699-1706, 2003.
2. 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, 2006.
3. M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.
4. M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.
5. 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.
6. P. Srinivasan and T. Rindflesch, "Exploring Text Mining from Medline," Proc. Am. Medical Informatics Assoc. (AMIA) Symp., 2002.
7. T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," Nature Genetics, vol. 28, no. 1, pp. 21-28, 2001.
8. B.J. Stapley and G. Benoit, "Bibliometrics: Information Retrieval Visualization from Co-Occurrences of Gene Names in MEDLINE Abstracts," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 526-537, 2000.
9. 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, 2006.
10. J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic Extraction of Protein Interations from Scientific Abstracts," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 538-549,2000.
11. A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event Extraction from Biomedical Papers Using a Full Parser," Proc. Pacific Symp. Biocomputing, vol. 6, pp. 408-419, 2001.
12. G. Leroy, H.C. Chen, and J.D. Martinez, "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text," J. Biomedical Informatics, vol. 36, no. 3, pp. 145-158, 2003.
13. L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson, P.V. Ogren, and K.B. Cohen, "OpenDMAP: An Open Source,
14. Ontology-Driven Concept Analysis Engine, with Applications to Capturing Knowledge Regarding Protein Transport, Protein Interactions and Cell-Type-Specific Gene Expression," BMC Bioinformatics, vol. 9, article no. 78, Jan. 2008.
15. J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," J. Am. Soc. Information Science and Technology, vol. 59, no. 5, pp. 756-769, 2008.
16. Mai Shouman, Tim Turner and Rob Stocker, (2012), Using Data Mining Techniques In Heartdisease Diagnosis And Treatment, Japan-Egypt Conference on Electronics, Communications and computers, Vol.4, issue 10 page 189-193.