IJDCST @November Issue- V-1, I-7, SW-45
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Natural Language Processing and Machine Learning Techniques for Identifying Disease-Treatment Relations

[1]Praneeth Mareedu, [2]S. Phani Praveen , [3]U.Tulasi ,

[1] Student, PVPSIT, KANURU, VIJAYAWADA, KRISHNA DIST.

[2] Assistant Prof, PVPSIT,KANURU,VIJAYAWADA,KRISHNA DIST.

[3] Assistant Prof, PVPSIT,KANURU,VIJAYAWADA,KRISHNA DIST.

Abstract: The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. Empirical domain of automatic learning is used in tasks such as medical decision support, protein-protein interaction, medical imaging, and extraction of medical knowledge. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better and more efficient medical care. A ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. Due to advancements in medical domain automatic learning has gained popularity in the fields of medical decision support, complete health management and extraction of medical knowledge. The main objective of this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques used for representation of information and what classification algorithms are suitable for identifying and classifying relevant medical information in short texts. This paper describes how ML and NLP can be used for extracting knowledge from published medical papers. It acknowledges the fact those tools capable of identifying reliable information in the medical domain stand as building blocks for a healthcare system that is up-to-date with the latest discoveries. Our research focus on the diseases and treatment information and the relation that exists between these two entities.

Index Terms: Automatic Learning, Natural Language Processing, Machine Learning, Medical Decision Support, Healthcare, Classifiers.

I. INTRODUCTION

People are more concerned about their health than ever before. Medline is the richest and most used source of information. Database that has enormous articles regarding life sciences. People want Fast access to reliable information and in a manner that is suitable to their habits and workflow. The medical field has grown to such an extent that the people practicing medicine should not only have experience but also information about latest discoveries. Mainly paper focuses on the two problems:

- *Automatically identifying sentences from Medline*

  It provides the relation between the diseases and treatments.

- It is based on the three relations.
  - ✓ Cure
  - ✓ Prevent
  - ✓ Side effects

These tasks helped to develop the Information Technology Framework that provides all the healthcare information. The healthcare related information is a source for both healthcare providers and people.

Natural Language Processing (NLP) and Machine Learning (ML) are the techniques that are used here. The aim m is to show what representation

of information and algorithms used to identify and provide relevant healthcare information in short texts. Tools that are used to identify all medical related information are the building blocks in medical domain that has all latest discoveries up-to-date. The main aim is to focus on all information related to disease. This paper provides the foundation for development of technology framework that makes easy to find all the relevant information regarding treatment and diseases. This work presents various Machine Learning (ML) and information for classifying short texts and relation between diseases and treatments.

ML techniques the information is shown in short texts when identifying relations between two entities such as diseases and treatment. There is improvement in solutions when using a pipeline of two tasks. It is better to identify and remove the sentence that does not contain information relevant to disease or treatments.

II. RELATED WORK

This work presents various Machine Learning (ML) and information for classification of short texts and finds the relation between diseases and treatments. By the improving in the Hierarchical way of approaching it is better to identify and remove the sentence that does not contain information relevant to disease or treatments. It will be very complex to identify the exact solution if everything is done in one step by classifying sentences based on interest and also including the sentences that do not provide relevant information.

The data set used in this work are created and distributed. The data set contains information from medline with all relevant information including diseases, treatments and eight relations between diseases and treatment. The entity recognition is focused mainly for diseases and treatments. The models are Hidden markov model and maximum entropy model. Representation of information depends on medical lexical ontology, phrases and words in context. One of the key tasks in natural language processing is that of Information Extraction (IE) that is traditionally divided into three sub problems:

a. Coreference resolution
b. Named entity recognition
c. Relation extraction

The main task in this work is to extract healthcare information and the relation details. There are three approaches for the extraction of the relations. They are:

- Co-occurrences analysis

  It is based on words in context and lexical knowledge.
- Rule based approaches

  It is used in solving relation between entities.
- Statistical methods

Rule based approach use syntactic: part-of-speech and syntactic structure or information in the form of fixed patterns that has word that describes the relation between the entities. Syntactic rule based systems are very complex since it depends on some more tools used to assign POS tags. It is clearly mentioned that such tools are not yet reached the state of art level since they are for designed only for common English texts and therefore it does not provide better solutions.

To obtain good results semantic and syntactic based systems are combined so that they provide flexibility of syntactic information and good precision of semantic rule. The statistical approaches are used to solve various tasks. This approach is used to solve various NLP tasks. Concerning relation extraction the rule checks whether the text information contains any relation or not. Some researchers combined this technique with POS which provides two sources of information such as relation between entities and their specific contexts.

III. ARCHITECTURE

The server will extract the information from various articles related to those symptoms in the case of the work user can give their symptoms. Then it classifies that information based on the symptoms and then provides the cure, preventive measures and side-effects for those symptoms.

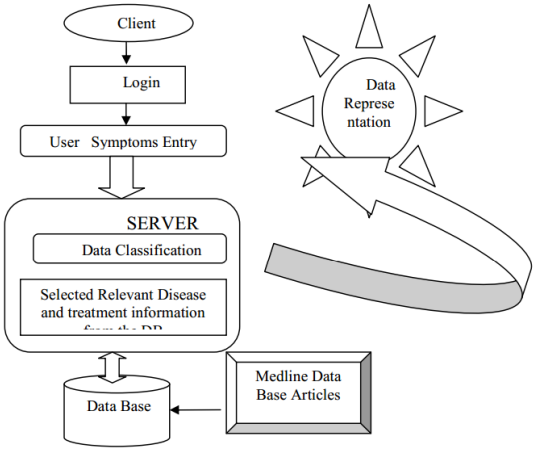

Figure 1.System Architecture

The main task in this work is to extract healthcare information and the relation details. It focuses on diseases and treatment information, and the relation that exists between these two entities. User interests are in line with the tendency of having a personalized medicine. It is not enough to read and know only about one study that states that a treatment is beneficial for a certain disease.

IV. PROPOSED APPROACH

Tasks and Data Sets: The two tasks used in this paper are the basis for the development of information technology framework. Our framework helps to identify the medical related information from abstracts. First task deals with extraction all information regarding diseases and treatments while the task deals with extraction of related information existing between disease and treatments. The future product can be provided with browser plug-in and desktop application so that it helps the user to get all information related to diseases and treatments and also the relation between those entities. It is also be useful to know more about latest discoveries related to medicine. The develop tools like Microsoft Health Vault and Google Health care product developed by the Natural Language Processing (NLP), Machine Learning (ML), and medical care domain. This product is valuable in e-commerce fields by showing the statistics that the information provided here are accurate and also provide all the recent discoveries related to health care. It is the key factor foe any company to make product successful. Coming to health care products it should be more trust worthy since it is dealing with health related problems. Natural Language Processing (NLP) and Machine Learning (ML) are used to extract accurate information or it can also say that it perfectly removes the unwanted information which is not related to disease or treatment. They itself involve in extracting informative sentences. It has difficult task to identify the informative sentences in fields such as summarization and information extraction.

In the first task data set are provided with information such as label containing that the extracted sentence is informative or label containing that the extracted sentence is not accurate.

In the second task, it is provided with information that shows if the relation between disease and treatment is prevent, cure or side effects. The three relations are mentioned in original data set and they are needed for future reference.
The most important thing is that they can be combined in pipeline so that it provides solution to framework which identifies all related information.

Classification Algorithms and Data Representations: In Machine Learning the expertise and previous research provides the guidance to solve new tasks. This model should be able to identify and provide informative sentences and relation between entities. While working with ML technique two challenges should be considered. ML provides a better model that can be used. The model one should rely on empirical studies and should gain knowledge in healthcare domain. Next one is to provide better data representation and to do feature engineering because feature increase the performance of the model. These two challenges can be achieved by using various algorithms and textual representation that suits for the tasks.

In the algorithm set of six representative models can be used. They are:

- Adaptive learning

  Adaptive learning algorithm is used to focus on hard concepts such as

unbalanced data sets and underrepresented in data.

- Decision-based models

  It based on decision models are used in short texts.
- Probabilistic models

  It based on Naive Bayes used in text classification and automatic text classification tasks.
- Complement Naive Bayes (CNB)

  It is adapted for text with imbalanced class distribution
- A linear classifier

  It support vector machine (SVM) with polynomial kernel

Bag-of-Words Representation: The bag-of-words is the name commonly used to classification of tasks. The selection techniques are used in order to identify the most suitable words as features. Each training and test instance is mapped to this feature representation by providing values to each feature for a specific instance. Representations for Bag-of-Words are expressed in the two common feature values. The binary feature value is the value of a feature can be either 0 or 1 that 1 represents the fact that the feature is present in the instance and 0 otherwise.

There is no that much difference between binary feature values and frequency feature values because there is only twenty words in each sentence of short texts. Advantage of using frequency feature values is that the feature's value will be greater than other features since it captures the number of feature appeared once in a sentence.

NLP and Biomedical Concepts Representation: A tool called Genia tagger is used to extract information. Tagger is specifically designed for biomedical text such as Medline abstracts. The phrases and nouns obtained by tagger are used in second representation method. Running the Genia tagger it extracts only nouns, phrases and healthcare related concepts from each sentence of data set.

Medical Concepts (UMLS) Representation: We use Unified Medical Language system (UMLS), is a developed at the US National Library of Medicine (NLM). The metathesaurus deals with concepts and meanings. Even it provides relation between various different concepts. NLM created a set of tools that allow easier access to the useful information. NLM created new tool called MetaMap that maps text to healthcare concepts in UMLS. This text is processed through entire data set and finally it provides ranked list of all possible concepts for particular noun-phrase. The best of the candidates are then organized according to the decreasing value of the fit function.

V. Evaluation Measures

The most common used evaluation measures in the ML settings are: precision, precision, accuracy, and recalls. All these measures are computed form a confusion matrix hat contains information about the actual classes. The test set on which the models are evaluated contain the true classes and the evaluation tries to identify how many of the true classes were predicted by the model classifier. For data sets that are highly imbalanced standard evaluation measures like accuracy are not suitable. We decided to report macro and not micro averaged F-measure because the macromeasure is not influenced by the majority class. The macromeasure better focuses on the performance the classifier has on the minority classes. Formulas for the evaluation measures are:

a. Accuracy ¼ the total number of correctly classified instances
b. Recall ¼ the ratio of correctly classified positive instances to the total number of positives
c. Precision ¼ the ratio of correctly classified positive instances to the total number of classified as positive
d. F-measure ¼ the harmonic mean between precision and recall

VI. CONCLUSION

This approach is very useful for everyone as it gives information only of the area of interest. The interests are in line with the tendency of having a personalized medicine that has one in which each patient has its medical care tailored to its needs. 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. The proposed system used the top concept candidate for each identified phrase in an abstract as a feature. We also consider as potential future work ways in which the framework's capabilities can be used in a commercial recommender system. . It extracts diseases and treatments given in and identifies the semantic relations between them.

VII. REFERENCE's

[1] P. Menaka, Prof.D.Thilagavathy , "Identifying Semantic Relations for the Disease Treatment in Midline," International Journal of Electronics and Computer Science Engineering, pp. 566-571.

[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] A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," Proc. 13th Text Retrieval Conf. (TREC), 2004.

[4] M. Craven, "Learning to Extract Relations from Medline," Proc.Assoc. for the Advancement of Artificial Intelligence, 1999.

[5] I. Donaldson et al., "PreBIND and Textomy: Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine," BMC Bioinformatics, vol. 4, 2003.

[6] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles," Bioinformatics, vol. 17, pp. S74-S82, 2001.

[7] 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.

[8] R. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," Bioinformatics, vol. 19, no. 1, pp. 135- 143, 2003.

[9] R. Bunescu and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/ EMNLP), pp. 724-731, 2005.