ISSN: 2321-7782 (Online)
Volume 2, Issue 3, March 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

# A System for Extraction of Semantic Biomedical Relations Using Multinominal Naive Bayes Algorithm

**Ancy Sudhakar**[1]
M-tech Computer Science and Engineering
KMCT College of Engineering Calicut
Calicut University
Calicut City – India

**Merin Meleet**[2]
Assistant Professor, Deptartment of CSE
KMCT College of Engineering Calicut
Calicut University
Calicut City – India

***Abstract:*** ***The machine learning field is used recently as the most efficient tool in various medical domains. In this paper better machine learning algorithms and techniques are used for extracting disease treatment relations from various medical related articles. Multinominal Naive Bayes algorithm is used for this extraction purpose. The three semantic relations such as cure prevent and side effect associated with diseases and treatments are also extracted. Other than this the symptoms related to the corresponding disease are also efficiently extracted. This system provides the user only informative sentences regarding diseases and treatments by avoiding all uninformative sentences. For making better medical decisions we can make use of this proposed technique.***

***Keywords:*** ***Machine Learning, Disease Treatment Extraction, Stemming Algorithm, Medline, and Multinominal Naive Bayes algorithm.***

## I. INTRODUCTION

Now a day's people are more concerned about their health. There are lots of tools available today that helps them in managing their health. But these tools have several disadvantages such as it finds difficulties in extracting information from clustered form of data, they were far behind in classification performance and also they do not provide much reliable information. The proposed technique provides more reliable access to information and classification performances are very much improved. The proposed approach provides the users especially doctors in making better medical decisions.

To improve the quality of result of user query and also to obtain informative data we have to make use of the text mining approach. In this project we use various medical databases especially the Medline database which provides the most relevant information regarding a disease and its treatment. Medline is the database which includes the latest medical discoveries. To read complete articles published in these databases to know about a particular disease and its treatment is a tedious work. So to avoid such problems we extract informative sentences regarding a disease, its treatment, its three semantic relations such as cure, prevent and side effect and also the symptoms related to the disease specified.

In this project we make use of certain data representation techniques and algorithms for efficient extraction of informative sentences. Common users and doctors can without wasting their time gain information regarding a disease, its treatment and its related symptoms. In this system we make use of the natural language processing and machine learning approach. By using this method of information extraction we can gain the information that whether a treatment is beneficial for a particular disease. The doctor can gain the knowledge that a particular treatment can cause side effect to certain people with some medical disorders.

The classification algorithm used in this project is very efficient and it easily helps in identifying the disease treatment relations in short texts. This project will be more useful for common users who find difficulty in reading medical related articles.

 

## II. LITERATURE SURVEY

The most relevant work is the work done by Rosario and Hearst [1] and they are the ones who have distributed the datasets used in our projects. They make use of Hidden Markov Model for entity recognition and maximum entropy model for relation identification. In this research work they focus on seven semantic relations. For mapping the words into semantic categories they used medical subject headings. In this work they compare different generative models and graphical models. They make use of recall, precision and F-measure values for evaluation of role extraction task. The Naive Bayes algorithm is used here for extracting semantic relations.

In the work of M.Craven [2] two methods of information extraction are done such as information extraction via text classification and information extraction via relational learning. The methods for decreasing the cost of information extraction process are specified in this work. A statistical method is used for classifying sentences and specifically a Naive Bayes classifier with bag-of-words representation is used. In the approach specified by Ray and Craven [3] they deal with sub cellular localization relation also the gene disorder association relation. From this relation we can identify the location of particular protein with in a cell and also the disorders associated with a particular gene. The training set and test set used here are individual sentences. Semi supervised machine learning approach is used in this work for relation extraction.

Various concept pairs are extracted via medical subject heading [4] and semantic representation. The relation extracted by these methods will be useful for healthcare practitioners or researchers. Experimental results suggest that for relation extraction it is better to use AdaBoost classifier [5] with binary bag-of-words representation. Sum of two area under curve measures is used here as an evaluation score. As a baseline a random classifier with 0.5 area under curve value is used. In this paper they deal with Compliment Naive Bayes and Naive Bayes classifier. One problem with the use of AdaBoost classifier is that they consider whole abstract as the source of information while here we deal with sentences. Other than this various methods were used for identification of relations from sentences. It includes pre-processor, parser and error recovery components [6]. The pre-processor is used for identifying sentences, words and phrases. The parser determines the grammar specified with a sentence and it helps in identifying the relation between two entities. The segments of sentences are parsed by the error recovery components by using various strategies.

The classification algorithms such as Support Vector Machines, Naive Bayes, Nearest Neighbor, and Decision trees and various representation techniques are specified for identification of the disease treatment relation [7]. In this work different bag-of-words representation techniques its advantages and disadvantages is specified. Weighted bag-of-words, latent semantic indexing and locality preserving indexing are specified here. The direct and indirect comparison between different classifiers is done. All the classifiers are compared with the baseline classifiers.

## III. PROPOSED SYSTEM

Different methods are undertaken in this paper for easily identifying and extracting the healthcare information's published in various medical related articles. We know that in the medical databases such as Medline and all there are lots of medial related articles. The main problem here is that to know about a particular disease and its treatment people have to read the entire article. So in order to avoid such difficulties we provide them with an easy method of extracting only relevant or informative sentences from the articles published in these medical databases. So here people are provided with the information regarding a particular disease , its treatments , symptoms and in addition to this its gives information regarding whether a particular treatment cause any side effect or whether it cures or prevents the disease.

For removing the unwanted information from the articles we use different methods. Firstly we remove the stop words form the articles and then by using the stemming algorithm we remove the repetition of words and after that with the help of Multinominal Naive Bayes algorithm and semantic probability calculations extract the informative sentences. The application used is designed using java. The button named relation finder finds the relation between diseases and treatments and also

 

provides us other information's. This button is placed near to a particular disease which the user or doctor is interested to know. Whenever the button is pressed the user or doctor obtains the relevant information regarding that particular disease. The system architecture of proposed system is shown in the figure1below:

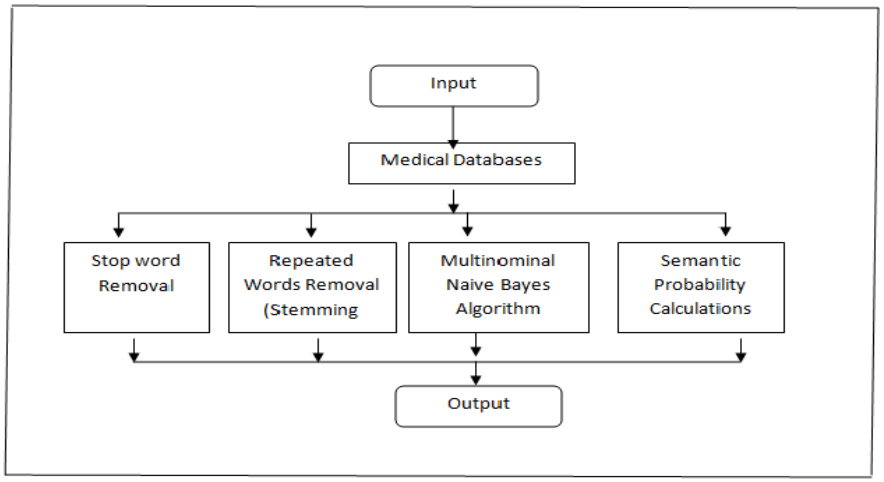


Figure 1: System Architecture of Proposed

In order to improve the quality of the result the process are performed in a pipelined manner. First we are having with us the text file which contains both informative and uninformative sentences. To avoid uninformative sentences we first perform the stop word removal process. In this removal process we remove stop words such as a, an, is, any, about, of, if, in etc from the text file. There are about 174 English stop words and we remove the entire stop words from the text file so that we can improve the quality of the result. By stop word removal content is reduced but quality is improved to a greater extend.

Next step is removal of repeated words. We know that after the stop word removal process the remaining text file contains repeated words such as expressed, expressing etc. The stem of such words for example express is same for two words we combine both of them to one word so that the repetition can be avoided. Likewise all the repeated words are removed. This removal of words increases the quality of result to a much higher level. For this removal we use the stemming algorithm. We know that the stemming algorithms are of different types. Out of this different stemming algorithm here we make use of the suffix stripping algorithm. By reducing the word count we can describe the information to the user in simple terms.

From the remaining text document we have to find the disease treatment relations. In addition to the three semantic relations cure, prevent and side effect we also find the symptoms associated with the disease. For finding the semantic relations here the Multinominal Naive Bayes algorithm is used. This algorithm easily finds the relation and we can easily display it to the user. The Multinominal Naive Bayes is used because it overcomes the drawbacks associated with the Naive Bayes algorithm.

In text classification we make use of this Multinominal Naive Bayes algorithm due to its simplicity and computational advantage. This algorithm is a specialized version of Naive Bayes .The Naive Bayes algorithm is not used here because it suffers from some problems. The main problem is that it assumes that the attributes of a given class are independent of each other. This is not always true because in some cases the attributes are related to each other. For example consider the classifier for in the case of assessing the risk of issuing a credit card. Here for a worthy customer it will not be true to assume that there is no dependency or relation between that customer's age, income and education level. To avoid this problem we prefer Multinominal Naive Bayes algorithm. In this we calculate the semantic probability, which helps in easily identifying the semantic relations between these entities.

The above described method of finding disease treatment relation can be used in various other applications. The quality of the result can be found out with the help of recall, precision and f-measure values. This project helps in saving the time of various users especially doctors by easily extracting the informative sentences from the medical related articles. There are various important modules used to perform these task and they are described as follows:

 

*A. Extraction of Informative Sentences from Medical Articles*

As the first module here we have the extraction of informative sentences. From the medical databases such as Medline we collect various medical related articles which give information about a particular disease. The bag-of-words representation is used for this text classification process. In the bag-of-words representation we consider each word as a feature for training the classifier which we use.

*1) Stop Word Removal Process:*

As the first process we remove the stop words associated with each sentence. The stop word removal reduces the content size but improves the quality of the document. All the stop words present in the document are removed. There are about 174 English stop words and all these when present in the document are successfully removed.

*2) Removal of Repeated Words:*

After the removal of stop words the remaining documents contains combination of repeated words and these words have to be removed from the document in order to improve the quality of the result. So for this here we make use of the stemming algorithm. It is a very useful algorithm for removal of such words. The suffix stripping algorithm is used here instead of the look up table. In suffix stripping we have smaller set of rules which provides the algorithm a definite path in order to find its root form.

*B. Identifying the Sentences and Relationship Extraction Process*

In the sentence identification task the sentences which contain information about diseases, its treatments and the symptoms for the corresponding disease are identified. For finding the three semantic relations such as cur, prevent and side effect associated with a disease and its treatment the Multinominal Naive Bayes algorithm is used. The semantic probabilities associated with these sentences are also calculated. Multinominal Naive Bayes algorithm takes into account the number of times a word occurs in a document or a text file. It never considers that the probability of a word is independent of the context of the document. For the relation identification task a training set and test set are required. The training set used for training the algorithm and the test set is used for testing how the algorithm performs well when given this test set.

*C. Output Performance Evaluation*

The output performance of the proposed system is evaluated for various documents. The documents are articles published in various medical databases such as Medline. The result we obtained shows informative sentences regarding diseases, treatments, the three disease treatment relations and symptoms related to the disease.

## IV. CONCLUSION

The proposed work provides us only informative sentences and removes uninformative sentences from the medical related articles in a pipelined manner. This system helps users especially doctors in saving their time and they can know easily about a disease its treatment and symptoms and can analyze more about a various treatments associated with a particular disease. This system will be more useful to common users who want to know more about a disease in simpler manner. In various healths care domains and all we can make use of this method.

## Acknowledgement

I express my sincere gratitude to my guide Mrs. Merin Meleet, Assistant Professor of Computer Science and Engineering Department, and I express my sincere gratitude to Mr.Pratap.G.Nair, Dean of Computer Science & Information Technology Departments, KMCT College of Engineering for their valuable guidance. Above all, I thank the almighty for enabling me to be what I am.

 

## References

1. B. Rosario and M.A. Hearst, "Semantic Relations in Bioscience Text," Proc. 42nd Ann. Meeting on Assoc. For Computational Linguistics, vol. 430,2004.
2. M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.
3. 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.
4. P. Srinivasan and T. Rindflesch, "Exploring Text Mining from Medline," Proc. Am. Medical Informatics Assoc. (AMIA) Symp., 2002.
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. 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. Oana Frunza.et.al, "A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts", May 2011.

## AUTHOR(S) PROFILE

**Ancy Sudhakar** received her B-tech degree in Computer Science and Engineering from College of Engineering Vadakara, Cochin University, Kerala, India and is presently doing her final year Master of Technology from KMCT College of Engineering Calicut, Calicut University, Kerala, India. Her research interest includes Data Mining and Knowledge Engineering.

**Merin Meleet** received her ME degree in Computer Science and Engineering from PSG College of Technology Coimbatore , India and is presently working as Assistant Professor in Computer Science and Engineering Department, KMCT College of Engineering, Calicut, India.

 