String Kernel Approach for Efficient Extraction of Medical Relations

Vrinda. V
PG Scholar, CSE Department
PSG College of Technology
Coimbatore

R. Venkatesan, PhD
Professor & HOD, CSE Department
PSG College of Technology
Coimbatore

ABSTRACT
This paper presents a methodology for building an application that can classify healthcare information. It extracts informative sentences from medical papers that mentions about diseases and treatments, and then recognizes semantic relations that exist between the entities in the informative sentences. Support Vector Machine algorithm with String Kernel is used for relation identification. The proposed system avoids unnecessary information and gives the user disease and Treatment related sentences from medical pages. Evaluation results for this approach show that the proposed methodology obtains reliable results. The string kernel approach is also compared with naïve bayes approach and it was found that string kernel approach outperforms naïve bayes method for classification. This technique can be integrated with any medical management system to make good decisions and in patient management system by automatically mining the biomedical information from digital repositories. This system enables easy access to medical information in rural areas where there is a relative shortage of physicians.

General Terms
Classification, Machine learning.

Keywords
SVM, String kernel, Naïve Bayes, Relation extraction.

1. INTRODUCTION
Compared with people living in more populated urban areas, residents in rural environments may experience greater isolation. According to National institutes of health, rural people may have little access to health care and service of physicians. There is a relative shortage of physicians in rural areas. Residents of rural area need to be provided with more reliable access to information regarding health care in a faster way. The work presented in this paper focuses on extracting medical relations from documents and providing fast and reliable access to the information regarding health care.

The Machine Learning field has become popular in all domains of research and it has recently become a good performing tool in the medical domain [1]. Machine learning means the construction and study of systems that can learn from data. It improves predictions or behaviors based on some data by teaching the computers. Machine learning is a huge field with hundreds of different algorithms for solving different problems [2]. Machine learning poses challenging problems in terms of data representation, algorithmic approach and computational efficiency of the resulting program.

Relation extraction from natural language text is a long standing research area. Format of biomedical information and its updates are usually in natural language text. Every day, the information is growing considerably. All research discoveries are stored in the database at high rate, which makes the process of recognizing and disseminating reliable information from the repositories a very difficult task. Manually inspecting such large amounts of data will be very difficult and also time consuming. Since the information is in natural language text which is not structured, efficient natural language processing methods need to be applied in order to map the information into structured useful format.

The proposed work extracts the useful disease related information from medical articles using machine learning. The tasks in this work focuses on identifying sentences published in medical papers that contain information about diseases and treatments, and then identifying semantic relations that exist between the entities, as expressed in these texts. All the unnecessary contents are removed and extraction of information or sentences related to user specified disease is performed. From the informative sentences that were extracted, sentences having relations to disease like causes, symptoms and treatment is taken and then displayed to the user. The resulting document can be used in health care domain. The doctor can gain information about medicine that are effective for some patient but causes side effect to patient with some additional medical disorder.

The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. For information extraction and relation extraction, representation and machine learning algorithm combination is being used.

People need fast and better access to reliable information in a manner that suits their habit 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. People are searching the web and read medical related information in order to be informed about their health. With the increased amount of biomedical publications, it is becoming more and more challenging to access useful and relevant information about a specific topic manually [1].

There are various machine learning techniques that can be used to classify text data such as decision trees, Bayesian, SVM, and nearest neighbor classifiers, which are briefly described in the following paragraphs.

1.1 Decision Tree classifiers
Decision tree classifier arranges the data in a tree like structure. The division of data is done recursively until the leaf nodes contain a certain amount of records or meet a requirement. The key problem in decision tree classifier is to build an optimal decision tree. As the size of search space is exponential, finding the optimal tree is computationally not feasible. The algorithms that develops decision tree usually

employ a greedy strategy where decision tree is constructed by making a series of locally optimum decisions as to which attribute to be used for dividing the data. Some of the decision tree induction algorithms are Hunt's algorithm, C4.5, ID3, SPRINT, CART.

1.2 Naïve Bayes classifiers

A naive Bayes classifier is a probabilistic classifier based on applying Bayes' theorem with naïve independence assumptions. A naive Bayes classifier assumes that the features are independent of each other i.e. value of a particular feature is not related to the occurrence of any other feature. For some categories of probability models, training of naive Bayes classifiers can be performed very efficiently in a setting of supervised learning. Usually in many applications, maximum likelihood method is used for parameter estimation by Naïve Bayes. One can work with the naive Bayes model without using any Bayesian methods or without accepting Bayesian probability.

1.3 Linear classifiers

Linear classifier tries to find the hyper plane between the different classes. Support Vector Machines are a form of linear classifiers that aim to find a linear separator between different classes. Text data is ideally suited for SVM classification because of the dispersion and high-dimensional nature of text, where they tend to be correlated with one another and can be organized into linearly separable regions.

Classification of data might be done by many hyperplanes. The hyperplane which provides largest separation of data or margin between two classes is the best one for classification of data. So the hyperplane is chosen in such a way that it maximizes the distance from it to the closest data point on both side. Such a hyperplane is known as the maximum-margin hyperplane. The linear classifier defined by the maximum-margin hyperplane is known as a maximum margin classifier.

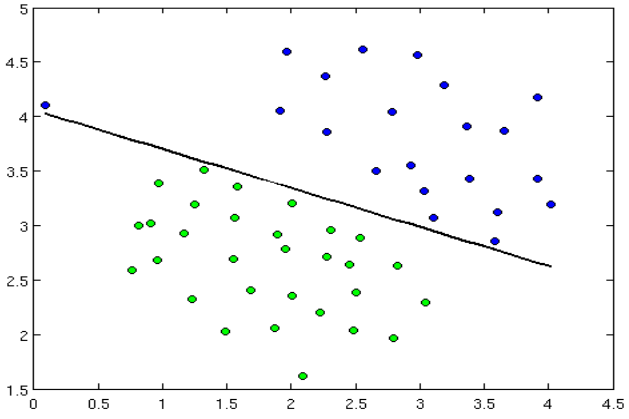

Fig 1 : SVM linear classifier

1.4 Proximity-based classifiers

Proximity-based classifiers uses distance measures to perform the classification. Documents that belong to the same class are likely to be closer to one another based on the dot product or cosine similarity. In K-Nearest neighbor method, classification of object is done based on closest training examples in the feature space. Based on a majority among the classes of the k nearest training points, the KNN algorithm assigns a point to a particular class.

2. LITERATURE SURVEY

B. Rosario and M.A. Hearst has proposed an approach for identifying relation between disease and treatment using graphical models and neural networks [3]. In graphical model, roles, relations and features are represented as nodes. The words and their features are the children of the role nodes, thus there are as many role states as there are words in the sentence. The aim was to recover the sequence of Role states, given the observed features. These models assume that with Markov assumption, it is possible to identify the ordering in the semantic roles and that the observations will be generated by the roles. The relation identification was also performed with feed forward neural networks. The result of neural networks outperformed graphical models. But neural network approach was slower and required a large amount of memory.

Thorsten Joachims analyzed the properties of training with text data and identified why support vector machines are suitable for this [4]. The goal of categorizing text was to classify documents into a fixed number of predefined categories. Text has properties like high dimensional input space and it has few irrelevant features which makes SVM better for text categorization. The results show that SVM consistently produce good performance on text categorization tasks and it outperforms existing methods substantially and significantly.

A Bayesian classifier has been developed to identify sentences in documents containing grant numbers, in the work proposed by Jongwoo Kim, Daniel X. Le, and George R. Thoma [5]. The Naive Bayes classifier depends on the several words in the zones. Therefore, it usually generates reasonable results that overcome situations such as typographic errors. But it had disadvantages like the methodology assumes full independence of the features and for problems that occur rare more training was required.

KNearest-Neighbour method was used for short text classification in the work proposed by Khushbu Khamar[6]. Using kNN approach Object is classified into the particular class which has maximum number of nearest instances. Distance function in KNN is computed based on distance between input test instance & training set instances. The distance/similarity function is important to compute the distance between instances. The advantage of KNN was that it is easy to implement and it is non parametric. But it had disadvantages like finding the optimal value of k is difficult. Also, KNN is sensitive to local structure of data etc.

In the work proposed by R. Bunescu and R. Mooney [7], a method for relation extraction is presented. The information needed to identify a relationship between two or more entities in the sentence is typically captured by the shortest path between those entities in the dependency graph. Using this method, it represents sentence as a chain of dependencies. But it assumes that named entities are already known and precision obtained was only 71.

Radiomics have been applied to select 3-D features from CT images of the lung toward providing prognostic information [8]. Several classifiers were used for predicting survival. The best performance was given by decision tree classifier. Using this approach, it was easy to handle missing values. But the tree constructed may become too large and it has Oversensivity to training set. Practical machine learning approaches for identification of MEDLINE documents and Swiss-Prot/TrEMBL protein records have been described in [9]. This also involves incorporation of those into TCDB which is a biological repository for transport proteins.

Learning with examples that are unlabelled have been done using biased classifier, and transductive SVM. This approach had the advantage that it considers all the points other than labeled ones. But no predictive model was built and If the number of features exceeds the number of samples, the approach might give poor performance. It was found that machine learning approaches outperform rules generated by a human expert.

In the work proposed by Rohini L. Damahe and Veena Kulkarni [10], XML data is processed using Eclat algorithm. It was observed that Eclat processing works on vertical pattern, so it produces less amount of candidates and so it takes less amount of time. But it was not so efficient when itemset was small. In the work by S. Gowrishanthi and Dr. Antony Selvadoss Thanamani, FP-Growth was used for web page categorization [11]. This approach needs Only 2 passes over the dataset and no candidate generation was required. But in some situations Fp-tree may not fit into memory and is expensive to build. In the work by Jiri hynek and Karel Jezek[12], document classification using Apriori rule was performed. Apriori approach is easy. It can also be parallelized. But it requires many data scans.

The proposed work differs from the ones mentioned in this section by the fact that SVM with String kernel approach was used for classification. This has an added advantage in the medical domain as most data are available in the form of text.

3. PROPOSED SYSTEM

Medical data is input to the system. After text preprocessing, Natural Language processing is done to extract the features. Relations between disease and treatment are identified using Support Vector Machine Algorithm with String kernel approach. The Proposed System is shown in figure 2.

3.1 Collection of Medical Data Set

The medical related data set is the input to the system. The medical data is taken from the URL http://www.nlm.nih.gov/medlineplus/all_healthtopics.html

3.2 Text Preprocessing

In text processing, stop words are words which are of less importance and are thus removed before processing of natural language data (text). Any set of words that need not be considered for processing can be chosen as the stop words. The stop words considered for removal are some of the most common, short words, such as is, the, at, which, and on. The stop words to be considered for removal are stored in database and are used for text preprocessing.

3.3 Bag of word representation

The bag-of-words model is a simple model used in retrieval of information and natural language processing. In this representation, a text is represented as the bag of its words, without considering the grammar and even word order but considering multiplicity. The below example models a text document using bag-of-words. Consider the following two simple text documents after text preprocessing:

John asked get vaccine chickenpox.

People who get vaccine will not get chickenpox.

Based on these two text documents, construction of a dictionary is done as follows:

```
{
   "John": 1,
   "asked": 2,
   "get": 3,
   "vaccine": 4,
   "chickenpox": 5
   "people": 6,
   "who": 7,
   "will": 8,
   "not": 9,
}
```

It has 9 different words. So in this case each of the document is represented as a 9-entry vector:

[1, 1, 1, 1, 1, 0, 0, 0, 0]

[0, 0, 2, 1, 1, 1, 1, 1, 1]

where each value of the vectors represents the frequency of the entry in the dictionary. A hash map is used for constructing the dictionary where key is the word and term frequency is the value.

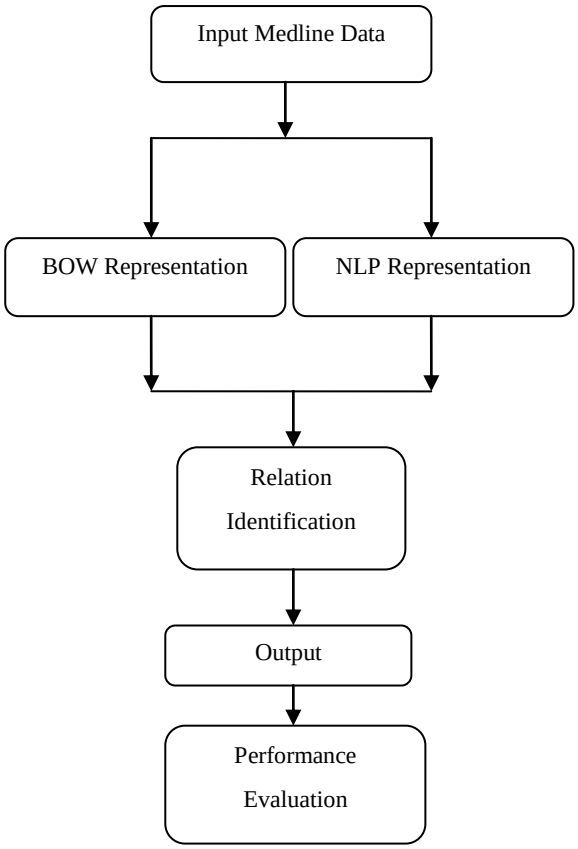

Fig 2: The flow graph of Proposed system

3.4 Natural language processing

Nouns are extracted from sentences in a document. The opennlp packages were used for this purpose. The jars added include: opennlp-maxent.jar, opennlp-tools.jar, opennlp-uima.jar. A parser en-parser-chunking.bin and a POS tagger en-pos-maxent.bin were used for noun extraction. A hash set was used to store the nouns extracted from each document by nlp. The steps involved in noun extraction are as follows:

1. Load chunking model

2. Create parse tree

3. Call subroutine to extract noun phrases

4. Recursively loop through the tree, thus extracting noun phrases

Medical terms were also extracted from sentences in a document. The extracted medical terms were stored in a hash set and added to database

3.5 Relation Identification using SVM with String kernel

Classification has been done using support vector machine with string kernel function. Support vector machines are supervised learning models in machine learning with learning algorithms that analyzes data and identify patterns using the training set. SVM is used for classification purpose. When a set of training examples for medical data set are given, each marked as belonging to one of the categories i.e symptoms, causes and treatment, an SVM training algorithm builds a model using the training set and it assigns new test samples into one category or the other. For identifying the nouns and medical terms, the categorization done in the previous step was being used. Here, nonlinear classification is done by using the kernel trick. SVM with string kernel is used for this purpose.

A string kernel is a kernel function [13] that can be applied on strings. This kernel applies on finite sequences that can be of different lengths. String kernels are functions that measures how much two strings are similar. More the similarity between two string x and y, higher will be the value of string kernel K (x, y). Kernel function for two strings x and y can be expressed as

$$K(x, y) = \Phi(x).\Phi(y) \dots\dots\dots (1)$$

Where Φ maps the arguments into an inner product space.

The idea used is feature map based on spectrum of sequences. The k-spectrum of a sequence is the set of all k-length contiguous subsequences that it contains. For each training set and test set, the k-3 sequences were generated and measure of similarity of test set with each training set is done. The test set was classified to the one with higher measure of similarity. An example of sequence generated for k=3 is shown in figure 3.

When user searches for information of a particular disease, semantic matching is done and information related to disease like treatment, symptom, cause etc are retrieved.

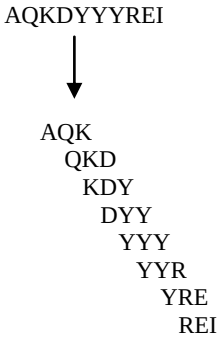

Fig 3: Kernel with k=3

3.6 Relation Identification using Naïve Bayes

This class of techniques treats a document as a set of distinct words with no frequency information, in which an element (term) may be either present or absent. A Naive Bayes Classifier predicts a class for the item, given a set of attributes for the item [14, 15].

For each known value, the following are done for classification using Naïve Bayes classifier.

1. Calculate conditional probabilities for each attribute.

2. Use the product rule in order to calculate the joint conditional probability for the attributes.

3. Derive conditional probabilities by using Bayes rule.

The Naive Bayes classifier outputs the class with highest probability value.

For training Naïve Bayes classifier with labeled data, first the conditional probability of each attribute was calculated from the training data. If X is data tuple with attributes {x1, x2….xn}, then

$$P(Ck \mid X)=(P(Ck)P(X \mid Ck))/P(X) \dots\dots\dots (2)$$

Where,

P(ck|x) is the class's posterior probability given attribute

P(ck) is the prior probability of class.

P(x|ck) is the probability of the attribute x belonging to ck.

P(x) is the probability of attribute.

For each of the three classes say Symptom, Cause and treatment, the posterior probability of each attribute belonging to these three classes was calculated in the training phase. To illustrate this concept, consider the following frequency table shown in Table 1.

Table 1: Frequency table for attributes

Attribute	Symptom	Cause	Treatment
A	5	2	1
B	2	7	3
C	1	2	5

Conditional probability of A is calculated as follows:

P(A/Symptom)=5/8

P(A/Cause)=2/11

P(A/Treatment)=1/9

P(Symptom)= 8/28

P(Cause)=11/28

P(Treatment)=9/28

P(A)=8/28

P(B)=12/28

P(C)=8/28

In the classification phase, the total probability of a sentence belonging to a class is found. The class C that maximizes the

probability is selected. Probability of a sentence with features x1, x2…..xk belonging to a class C is computed as follows:

$$P(C|x1,x2....xk)\alpha P(C)\prod_{i=1}^{k} P(\frac{Xi}{C})) \ldots\ldots\ldots\ldots (3)$$

Probability of each sentence belonging to the classes Symptom, Cause and Treatment is found out. Then sentence is assigned to the class with maximum probability.

4. SYSTEM ENVIRONMENT

The system environment used for the application is shown in the below table.

Table 2: System Requirements

Operating system	Windows 7
RAM	2 Gigabyte
Hard disk space required	100 Gigabytes
Libraries	Opennlp-maxent-3.0.3.jar Opennlp-tools-1.5.3.jar Opennlp-uima-1.5.3.jar

5. RESULTS

Results were evaluated using precision, recall and F-measure. Recall is the ratio of correctly classified positive instances to the total number of positives. Precision is the ratio of correctly classified positive instances to the total number of classified as positive. F-measure is the harmonic mean between precision and recall. Results obtained with SVM using string kernel classification and naïve Bayes classification are shown in figure 4.

From the results, it can be observed that for classification using SVM with string kernel, the precision obtained is 86%. Recall is 73% and f-measure is 79%. For Naïve Bayes classification, Precision and recall is 74%. Even though recall was a little high for Naïve Bayes, precision and Fmeasure was very high for SVM with string kernel approach compared to Naïve Bayes.

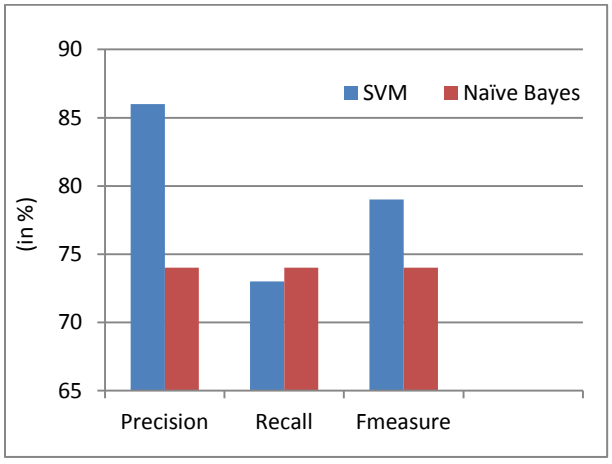

Fig 4: Results

Higher precision for SVM with string kernel approach indicates that out of the total instances retrieved, most of them were relevant. Hence, this approach can be used for efficient extraction of medical relations.

6. DISCUSSION AND CONCLUSION

The medical data were categorized into appropriate class using SVM with string kernel approach and Naïve Bayes approach. The ML algorithm results are influenced by representation techniques. Representations that consistently obtain best results are more informative.

The medical relation extraction performed can be used in health care sector where a doctor can analyze various treatments that can be given to patient for a particular disease. The doctor also gain information about medicines effective for some patient and causes side effect to some other patients. This system can also be used in rural areas for faster access to information.

SVM is one of the best known classifiers. String kernel approach enables to implement a kernel function that can be applied on strings. Another advantage is that kernel applies on finite sequences that can be of different lengths. But it had some limitation also. A practical disadvantage with string kernel approach is their computational expense. In general, these kernels rely on dynamic programming algorithms for which the computation of each kernel value K(x, y) is quadratic in the length of the input sequences x and y, that is, O(|x||y|) with constant factor that depends on the parameters of the kernel.

As future work, classification can be performed by a hybrid concept of association rule mining with machine learning approach

7. ACKNOWLEDGMENTS

Our sincere gratitude to Dr.R.Rudramoorthy, Principal, PSG College of Technology for providing us with necessary facilities for the work. We also thank National Institute of Health for providing open access to medline data.

8. REFERENCES

[1] Oana Frunza, Diana Inkpen, and Thomas 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

[2] Bertrand C., Ernest F. and Zhang H.H., Principles and theory for data mining and machine learning. USA: Springer, 2009, pp. 405-485.

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

[4] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with many relevant features", Dortmund, Germany, 1998

[5] Jongwoo Kim, Daniel X. Le, and George R. Thoma, "Naïve Bayes Classifier for Extracting Bibliographic Information from Biomedical Online Articles", Proc 2008 International Conference on Data Mining. Las Vegas, Nevada, USA. July 2008

[6] Khushbu Khamar, "Short Text Classification Using kNN Based on Distance Function", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013

[7] 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.

[8] Samuel h. hawkins, John n. korecki, Voganand balagurunathan, Vuhuagu, Virendrakumar, Satrajitbasu, Llawrence o. hall, Dmitry b. goldgof, robert a. gatenby, and robert j. gillies, "Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features", IEEE Transactions November 2014, volume 2

[9] Aditya Kumar Sehgal, Sanmay Das, Keith Noto, Milton H. Saier, Jr., and Charles Elkan, "Identifying Relevant Data for a Biological Database: Handcrafted Rules versus Machine Learning", IEEE/ACM transactions on computational biology and bioinformatics, vol. 8, no. 3, may/june 2011

[10] Rohini L. Damahe, VeenaKulkarni, " A Framework for Processing XML data Using Eclat Algorithm", International Journal of Emerging Technology and Advanced, Volume 4, Issue 8, August 2014

[11] S. Gowrishanthi, Dr. Antony Selvadoss Thanamani, " Web page categorization using web mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 7, September 2012 .

[12] Jiri hynek, Karel Jezek, "Document classification using itemsets", Proceedings of International Conference MOSIS 2000, Czech Republic, May 2000.

[13] Huma Lodhihuma, Craig Saunders craig, John Shawe-Taylor , Nello Cristianini , Chris Watkins, "Text Classification using String Kernels", Journal of Machine Learning, 2002.

[14] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[15] Wu X. and Kumar V., "The top ten algorithms in data mining" USA: CRC Press, 2009, p. 21, p. 93.