IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 7, Issue 5 (Nov-Dec. 2012), PP 19-24
www.iosrjournals.org

A Machine Learning Approach for Classifying Medical Sentences into Different Classes

D. Nagarani[1] Avadhanula Karthik[2] G.Ravi[3]
[1,2] Student(M Tech(Software Engineering) Dept.of Computer Science & Technology, Sreenidhi Institute of Science & Technology/An Autonomous Institution, Hyderabad, AP, India)
3 Assistant Professor(M Tech(Software Engineering) Dept.of Computer Science & Technology, Sreenidhi Institute of Science & Technology/An Autonomous Institution, Hyderabad, AP, India)

Abstract: *The medicine that is practiced today is an Evidence-Based Medicine (EBM) in which medical expertise is not only based on years of practice but on the latest discoveries as well. All research discoveries come and enter the repository at high rate, making the process of identifying and disseminating reliable information a very difficult task. The work that we present in this paper is focused on two tasks: automatically identifying sentences published in medical abstracts as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments. The second task is focused on three semantic relations: Cure, Prevent, and Side Effect. The objective for this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques—what representation of information and what classification algorithms—can be used for identifying and classifying relevant medical information in short texts.*

Keywords: *Natural Language Processing; Machine Learning, Medical abstracts, semantic relations*

I. Introduction

Information is growing significantly every year. According to studies, the volume of medical knowledge doubles every five years , or even every two years . Despite the adoption rate of medical knowledge significantly improving, we face a new challenge due to the exponential increase in the rate of medical knowledge discovery. More than 18 million articles are currently catalogued in the biomedical literature.

Translating all of this emerging medical knowledge into practice is a staggering challenge. One hundred years ago, it is said that a physician might have reasonably expected to know everything in the field of medicine. Today, a typical primary care doctor must stay abreast of approximately 10,000 diseases and syndromes, 3,000 medications, and 1,100 laboratory tests. The ongoing information explosion has created a situation where it is now impossible for any one person to stay up-to-date with the changes in any topic area.

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. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, extraction of medical knowledge, and for overall patient management care. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, more efficient medical care. This review focuses on ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments.

II. Literature Survey

Reference paper:-
Alan R. Aronson Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap
Description:

Research has shown that MetaMap is an effective algorithm for discovering Metathesaurus concepts in text. But there are two areas in which MetaMap's performance requires improvement: first, detection of idiosyncratic text such as chemical names, acronyms and abbreviations, numeric quantities or similar constructs; and second, resolution of ambiguity. The first problem is being solved through the use of an extensible, hierarchical tokenization regime. The initial implementation of this regime includes detection of acronyms/abbreviations and chemical names. The problem of ambiguity is being investigated by developing a word sense disambiguation (WSD).

Reference paper:-
Anna Divoli and Teresa K. Attwood BioIE: extracting informative sentences from the biomedical literature
Description:

BioIE is a rule-based system that extracts informative sentences relating to protein families, their structures, functions and diseases from the biomedical literature. Based on manual definition of templates and rules, it aims at precise sentence extraction rather than wide recall. After uploading source text or retrieving abstracts from MEDLINE, users can extract sentences based on predefined or user-defined template categories.

Reference paper:-

Asma Ben Abacha, Pierre Zweigenbaum Automatic extraction of semantic relations between medical entities: a rule based approach

Description:

Knowledge and linguistic-based approach for the extraction of medical entities and the semantic relations linking them. This approach is based on two main steps: (i) the recognition of medical entities with an enhanced use of MetaMap and (ii) the exploitation of linguistic patterns taking into account the semantic types of medical entities. Using an external sentence segmenter and noun phrase chunker may improve the precision of MetaMap-based medical entity recognition. Our pattern-based relation extraction method obtains good precision and recall w.r.t related works.

III. The Proposed Approach

The two tasks that are undertaken in this paper provide the basis for the design of a framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments.

The problems addressed in this paper form the building blocks of a framework that can be used by healthcare providers (e.g., private clinics, hospitals, medical doctors, etc.), companies that build systematic reviews.The final product can be envisioned as a browser plug-in or a desktop application that will automatically find and extract the latest medical discoveries related to disease-treatment relations and present them to the user.

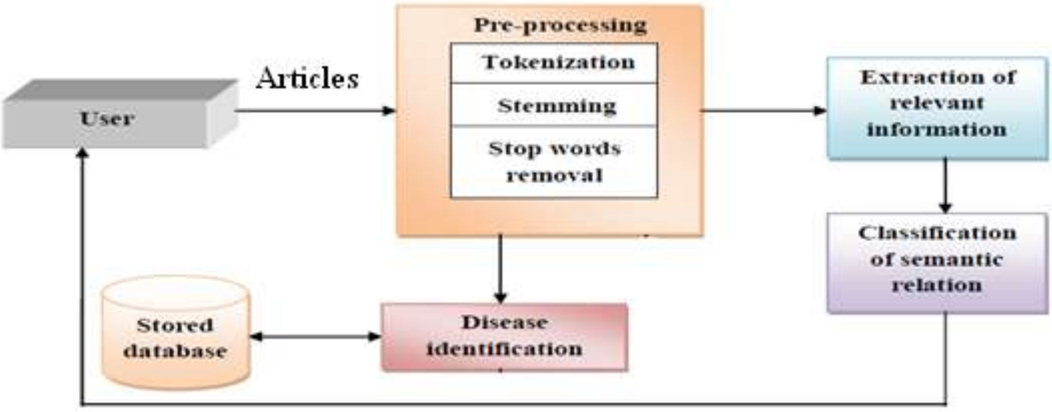

Figure 1: Architectural design of the proposed system

1.1 Sentence Extraction

The first task (sentence selection) identifies sentences from published abstracts that talk about diseases and treatments. The task is similar to a scan of sentences contained in the abstract of an article in order to present to the user-only sentences that are identified as containing relevant information (disease treatment information).

Semantic Category Recognition:

For Identifying treatments,diseases and many other medical entities from raw sentences, an algorithm is used called MetaMap.

The MetaMap Algorithm

MetaMap is a algorithm used for medical entity recognition which allows mapping medical text to UMLS concepts. Using MetaMap therefore provides a strong baseline to start with. MetaMap is able to identify most concepts in the titles of articles from medline.

By “medical entity”, we refer to an instance of a medical concept such as Disease or Drug. Medical entity recognition consists in: (i) identifying medical entities in the text and (ii) determining their categories. For instance, in the following sentence “ACE inhibitors reduce major cardiovascular disease outcomes in patients with diabetes.”, the medical entity ACE inhibitors should be identified as a treatment and the medical entity cardiovascular disease outcomes should be identified as a problem.

Steps to follow in Metamap algorithm

1. Parse the text into noun phrases and remove stop words.
2. Generate the variants for each noun phrase where a variant essentially consists of one or more noun phrase words together with all of its spelling variants, abbreviations, acronyms, synonyms, inflectional and derivational variants.
3. Form the candidate set of all Metathesaurus strings containing one of the variants.
4. Supply the candidate set to the Metamap through java API

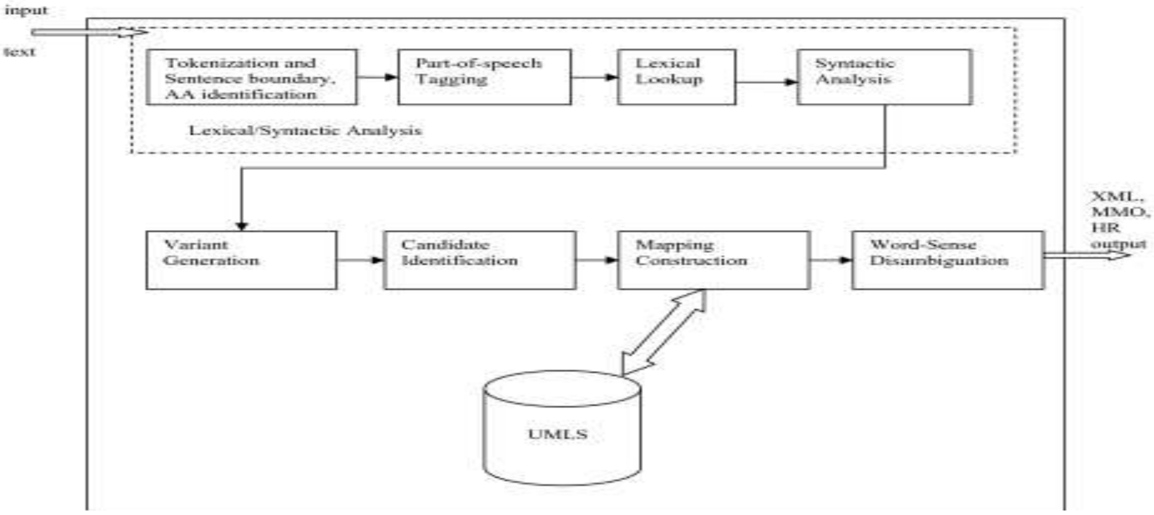

Figure 2: METAMAP PROCESSING

Description of each step in Detail:

1. **Parsing** : Arbitrary text is parsed into (mainly) simple noun phrases; this limits the scope of further processing and thereby makes the mapping effort more tractable. Parsing includes two different subtasks.

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.
Stop words are words which are filtered out prior to, or after, processing of natural language text. There is not one definite list of stop words which all tools use. Any group of words can be chosen as the stop words for a given purpose.

2. **Variant Generation :** A variant consists of a phrase word (called a *generator*) together with all its acronyms, abbreviations, synonyms, derivational variants, meaningful combinations of these, and finally inflectional and spelling variants. A variant generator is any *meaningful* subsequence of words in the phrase where a sub-sequence is meaningful if it is either a single word or occurs in the SPECIALIST lexicon.

3. **Candidate Identification:** The Metathesaurus candidates for a noun phrase consist of the set of all Metathesaurus strings containing at least one of the variants computed for the phrase.The candidates for the noun phrase *ocular complications* are given below

Complications

complications Eye Optic Ophthalmia

4. **Mapping Construction**: The final step in the mapping algorithm is straightforward. It consists of examining combinations of Metathesaurus candidates which participate in matches with disjoint parts of the noun phrase.

Keyword searching algorithm:

After the diseases and treatments are identified, then sentences containing that diseases and treatments are extracted from database. Then these extracted sentences are further classified into informative sentences which contain relevant keywords and non informative sentences which contain irrelevant keywords.

Keyword searching algorithm
Input: Articles extracted from database corresponding to the disease identified.
Output: Relevant semantic keywords.
Step 1: Extract the input articles.
Step 2: Preprocess all the extracted input articles.
Step 3: First split the paragraph into sentences using delimiter.
Step 4: Next, split each sentence into word using the Stanford POS tagger tool which also creates tags and words are enclosed with these tags.
Step 5: Obtain the meaning of each word (using parts of speech).
Step6: Retrieve the informative sentences using the relevant semantic keywords.

1.2 Classification

The second task (task 2 or relation identification) has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first). We focus on three relations: Cure, Prevent, and Side Effect.

After extracting informative Sentence containing input disease name as template, we can classify the sentences and label them with three relations Cure, Prevent and Side Effect.

- How we can classify sentences? Sentences are classified using generated attributes or labels (here cure, prevent, side affects).
- Based on what attributes are generated? By selecting features which best characterize a document or sentences.
- What features are selected from documents or sentences? Text Representation:

Text document is represented by the words (features) it contains and their occurrences

Bag-of-words representation

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which the features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped into this feature representation by giving values to each feature for a certain instance. Two feature value representations are the most commonly used for the BOW representation: binary feature values – the value of a feature is 1 if the feature is present in the instance and 0 otherwise, or frequency feature values – the feature value is the number of times it appears in an instance, or 0 if it did not appear.

***Classifier**:*

We use a classifier to automatically generate labels (attributes) from the features we feed into it. Two Approaches are used to Feature Selection:

1. Select features before using them in a classifier
   –Requires a feature ranking method.
   –Many choices.

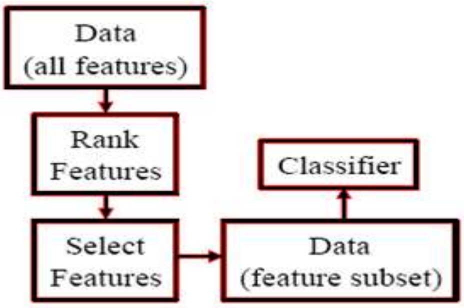

Figure 3 : process of Select features before using them in a classifier

2. Select features based on how well they work in a classifier
–The classifier is part of the feature selection method.
–Often an iterative process

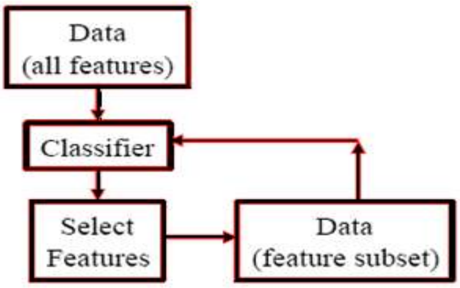

Figure 4 : process of Select features based on how well they work in a classifier

Feature Selection from corpus using Bag of words approach

The classification of documents by machine learning techniques requires a representation of each document as a set of features; almost without exception, these features are based on the presence, absence, or frequency of words in the text. This 'bag-of- words' model is popular both due to its ease of implementation, and its excellent performance on many tasks.

Problems With the Bag of Words Approach

- Words that appear in *testing* documents but not in *training* documents are completely ignored by the BOW approach.
- Words that appear infrequently in the training set, or appear just once, are mostly ignored even if they are essential for proper classification.

Feature Selection

This feature selection process is seen as the dimensionality reduction. This has been applied in a sequence of two steps. First eliminating words by document frequency and then applying an attribute selection measure like Information Gain. Eliminating words based on document frequency, refers to the procedure that we calculate the frequency of each word in all the documents and then retain only those words which satisfy certain threshold value.

Information Gain (IG):

The goal of applying the IG to the set of feature vector T is to find the best subset T' which maximizes the classification efficiency. Information for any attribute or feature is its measure of purity. It represents the amount of information that this feature carries and helps in classifying a new instance based on this word alone.

Actual Attribute Generation

Attributes generated are merely labels of the classes automatically produced by a classifier on the features that passed the feature selection process. The next step is to populate the database that results from above.

Machine Learning Approach

The goal of classification is to build a set of models that can correctly predict the class of the different sentences. The input to these methods is a set of sentences (i.e., training data), the classes which these sentences belong to (i.e., dependent variables), and a set of variables describing different characteristics of the sentences (i.e., independent variables). Once such a predictive model is built, it can be used to predict the class of the sentences for which class information is not known a priori.

Naïve Bayesian

algorithm has been widely used for text classification, and shown to produce very good performance. The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a sentence. NB algorithm computes the posterior probability that the sentence belongs to different classes and assigns it to the class with the highest posterior probability. The posterior probability of class is computed using Bayes rule and the testing sample is assigned to the class with the highest posterior probability. The naive part of NB algorithm is the assumption of word independence that the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category.

Naïve bayes is used as classifier for categorizing the sentences belonging to either cure or prevent or side affect based on features that are identified in training set.

Naïve-Bayes is one of the widely used techniques for text classification. The learning task for the Naïve-Bayes classifier is to use a set of training sentences to estimate the model parameters then use the estimated model to classify the new sentences. *bag-of-words* approach was used to represent the sentences

1. **Training:** For each class c_j of documents
   (i) Estimate $P(c_j)$
   (ii) For each word w_k estimate $P(w_k \mid cj)$

2. **Classify (sentence):**

Suppose we have a set of training sentences D and each sentence $d_i \in D$ is assigned to a category $c_j \in C$. The purpose of a Naïve-Bayes classifier to label a given test sentence t_i with a category $c_j \in C$ that produces highest $P(c_j \mid t_i)$ given the set of training sentences, D. $P(c_j \mid t_i)$ is calculated with the following formula:

$$P(c_j|t_j)=\frac{P(c_j)\prod_{k=1}^{|t_i|}P(w_k|c_j)}{\sum_{m=1}^{|C|}P(c_m)\prod_{k=1}^{|t_i|}P(w_k|c_m)}$$

assuming words are conditionally independent, given class where $P(c_j)$ is the class probability of c_j, w_k is a word that appears in t_i 's representation, and $P(w_k \mid c_j)$ is the probability of w_k given c_j. $P(c_j)$ and $P(w_k \mid c_j)$ are calculated with the following formulas:

$$P(c_j)=\frac{\sum_{i=1}^{|D|}P(c_j|d_i)}{|D|}$$

$$P(w_k|c_j)=\frac{1+\sum_{i=1}^{|D|}N(w_k,d_i)P(c_j|d_i)}{|V|+\sum_{l=1}^{|V|}\sum_{i=1}^{|D|}N(w_l,d_i)P(c_j|d_i)}$$

where $N(w_k, d_i)$ is the number of times word w_k appears in training sentence d_i and V is the set of distinct words that appear in the complete set of training sentences.

IV. Conclusion

The conclusions of our study suggest that domain-specific knowledge improves the results. Pipelining of two tasks gives effective results because classifier can focus only on informative sentences rather than all sentences including non informative sentences.

References

[1] Rindflesch,T. EDGAR: extraction of drugs, genes and relations from the biomedical literature.

[2] Asma Ben Abacha, Pierre Zweigenbaum Automatic extraction of semantic relations between medical entities: a rule based approach

[3] Anna Divoli and Teresa K. Attwood BioIE: extracting informative sentences from the biomedical literature

[4] Alan R. Aronson Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap

[5] Rindflesch TC and Aronson AR. Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus

[6] McCray AT, Srinivasan S and Browne AC. Lexical methods for managing variation in biomedical terminologies.

[7] Lee CH, Khoo C, Na JC: Automatic identification of treatment relations for medical ontology learning: An exploratory study.

[8] Ben Abacha A, Zweigenbaum P: Medical Entity Recognition: A Comparison of Semantic and Statistical Methods.