National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications (NCACSA 2012)
Proceedings published in International Journal of Computer Applications® (IJCA)

Extraction of Semantic Biomedical Relations from Medline Abstracts using Machine Learning Approach

Suchitra A
M.E Computer and Communication
Department of Information Technology
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India.

Sudha R
Associate Professor
Department of Information Technology
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India.

ABSTRACT
Machine Learning (ML) is a natural outgrowth of the intersection of Computer Science and Statistics. Machine Learning has now become a reliable tool in the medical domain. ML which act as a tool by which computer-based systems could be integrated in the healthcare field in order to get a better and efficient health care. This methodology for building an application is capable of identifying and extracting healthcare information. The proposed system focuses on two main tasks. The first task identifies the sentences which are published in Medline abstracts. This task is similar to the task of sentence scanning contained in the medical abstract of an article in order to present to the user-only sentences that are identified as containing relevant information. The second task has a deeper semantic dimension and it focus on identifying semantic relations exists between disease-treatment. It focuses on three relations: Cure, Prevent, Side Effect and also focuses a subset of the eight relations that the corpus is annotated with. The proposed methodology obtains reliable outcomes and that could be integrated in an application to be used in the medical care domain. The framework's capabilities can be used in a commercial recommender system and it is integrated in a new Electronic Health Record system.

Keywords
Healthcare, machine learning, natural language processing.

1. INTRODUCTION
People care deeply about their health and they want fast access to reliable information which is suitable to their habits and workflow. The medicine that is practised today is an Electronic Health Records (EHR) and Evidence Based Medicines (EBR). These two systems are now becoming standard in the health care domain. EBR is only based on several years of practice but not on the latest discoveries as well. The challenging issues confronting EHRs is the fact that physicians must be the users of the system performing data entry as well as information retrieval if they are to realize the benefits of interactive on-line decision support.
There are several security technologies available that will help prevent unauthorized access to protected health information. Some of these technologies in health care include firewalls; passwords and properly designed and monitored audit trails can enhance user accountability by detecting and recording unauthorized access to confidential information. The enormous obstacle in the implementation of an Evidence Based Medicine is the lack of standardized terminology, lack of evidence and lack of skills. Another main obstacle in EBR is applying the older professionals to modern healthcare and costs. Although the benefits that support implementation of an EHR and EBR are clear there are still barriers too therefore the concept is still not accepted.
Tools that can help us to manage and better keep of our health using search engines such as Google Health 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. Using of search engines results in drawbacks such as poor precision, poor recall, varied document quality and in varied indexing path. The main drawback of using search engines involves a learning curve. Many beginning internet users because of these disadvantages become discouraged and frustrated. Regardless of the growing sophistication many well thought-out search phrases produce list after list of irrelevant web pages.
The typical search still requires sifting through dirt to find the gems. The information which are related to medical care is a source power for both healthcare providers and the people. People are searching the web and reading medical related information in order to be informed about their health. In order to eliminate the need of the physicians to monitor the user databases here we are moving to the machine learning approaches and it also help us to provide the latest medicine technologies since it keep track of up to date medical information from latest published medical abstracts. In order to overcome the drawbacks in the EHR system we need better, faster and more reliable access to information. Medline is a database of extensive life science published articles which is most used source in medical field. All research discoveries will come and enter the repository at high rate making the process of identifying and disseminating reliable information a very difficult task.
The proposed system focused on two tasks. The first tasks will automatically identifying sentences published in medical abstracts as containing or not information about diseases and treatments and also automatically identify semantic relations that exist between diseases and treatments as expressed in these texts. 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.
The framework's capabilities can be used in a commercial recommender system and it is integrated in a new Electronic Health Record system. Our objective for this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques are suitable to use for identifying and classifying medical information in short texts. These tools are capable of identifying reliable information in the medical

domain which stands as building blocks for the healthcare system that is up-to-date with the latest discoveries. We focus on diseases and treatment information and the relation that exists between these two entities.
Since 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. The results that we obtained show that it is a realistic scenario to use NLP and ML techniques to build a tool which is capable of identifying and disseminating information which are related to diseases and treatments. Therefore this process is aimed at designing as user friendly, eliminate the need of physicians and also examining various representation techniques used in combination with various learning methods to identify and extract biomedical relations.

2. RELATED WORKS

Three major approaches which are used in relation extraction between the entities are co-occurrences analysis, rule based approaches and statistical models. Under each type methods vary in how they utilize the lexical, syntactic and semantic information in texts.
Co-occurrence Analysis: Co-occurrence analysis identifies relations between biomedical entities based on their probabilities of occurrence in the article Stapley and Benoit [9]. These approaches are based on the assumption that if two entities are both mentioned in the same article there is an underlying biological relationship. In most cases only lexical information (i.e., words) is needed for co-occurrence analysis. Due to their simplicity and flexibility these approaches have been widely used for relation extraction and can achieve high recall. Since it can capture little syntactic or semantic information, co-occurrence analysis cannot distinguish relation types and often achieves low precision.

Rule-Based Approaches: The rule-based system suffers from lexicons that change from the domain to domain because the new rules to be created each time the domain changes. Semantic rules are applied to full-text articles which are described in Friedman [6].

Statistical Learning: The Natural Language Processing tasks can be solved by statistical methods. The statistical methods can perform well even with the little training data. Statistical learning can be categorized into feature based methods and kernel-based methods. Features are defined and selected to capture the data characteristics. Rosario and Hearst [10] compared generative graphical and discriminative models for relation extraction using both word and role features. Various supervised statistical algorithms have been used with the kernel methods are applied to medical abstracts.

Inductive Logic Techniques: Goadrich [11] used inductive logic methods for the extraction of information from the medical and biomedical domain.

Support Vector Machine: A machine learning technique gaining increasing recognition and popularity in recent years is the support vector machines (SVMs). SVM is based on statistical learning theory that tries to find a hyper plane to best separate two or multiple classes. This statistical learning model has been applied in different applications and the results have been encouraging.

3. PROPOSED APPROACH

3.1 Data Sets and the Proposed Tasks

The two tasks that are taken in this paper provide the design of an information technology framework which is capable to identify and extract health care information. The first task identifies and extracts sentences that mention on diseases and treatments topics. The second task performs a classification of these sentences according to the semantic relations that exists between diseases and treatments.
The first task identifies sentences from Medical abstracts that mention about diseases and treatments. The task is similar to a sentences scanning contained in the abstract of an article in order to present to the user-only sentences that are identified as containing relevant disease-treatment information.
The second task has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences which are selected as informative. Here we focus on three main relations: Cure, Prevent, Side Effect and a subset of the eight relations that is annotated with.
The NLP and ML based techniques are used to solve the two proposed tasks. In a standard machine learning which are under supervision, a training set and a test set are required. The training set is used to train the machine learning algorithm and the test set to test its performance. It identifies informative sentences that contain information about diseases and treatments and semantic relations between them versus non informative sentences. This allows us to see how well the natural language processing techniques and Machine learning based techniques can cope with the task of identifying informative sentences.

TABLE 1Sentence Selection Task

Label	Sentence
Informative sentence	Urgent colonoscopy for the diagnosis and treatment of severe diverticular haemorrhage.
Non-informative Sentence	In all cases a copra parasitological study was performed.

The algorithm uses a linear piecewise method for free surface reconstruction, coupled to a unique fully multidimensional method of cell boundary flux integration. The main words are called as key words are filtered using stemming process. This process will remove all the verbal words which are used understanding a sentence. We use Bloom filter for the removal of unwanted words so as to fetch only the important words.
For the first task the data sets are annotated with a label which indicating that the sentence is informative or a label indicating that the sentence is not informative. Table1 gives an example of labelled sentences. For the second task the sentences which have annotation information that states if there any relation that exists between the disease and treatment such as Cure, Prevent or Side Effect. These are the relations that are more represented in the original data set.

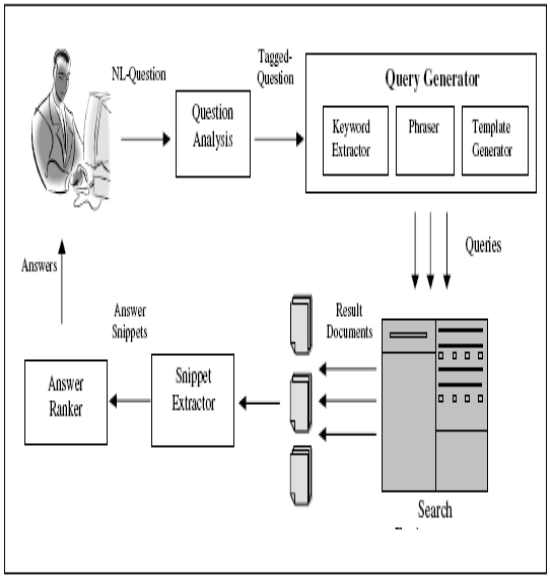

Figure 1. Question Answering System.

Figure1 shows the architecture of question answering system. The system takes in a natural language (NL) question in English from the user. This question is then passed to a Part-of-Speech (POS) tagger which parses the question and identifies POS of every word involved in the question. This tagged question is then used by the query generators which generate different types of queries, which can be passed to a search engine. These queries are then executed by a search engine in parallel.

The search engine provides the documents which are likely to have the answers we are looking for. These documents are checked for this by the answer extractor. Snippet Extractor extracts snippets which contain the query phrases/words from the documents. These snippets are passed to the ranker which sorts them according to the ranking algorithm. After getting the corresponding weight of the term of the particular symptoms, side effects of all the values are tabled for further processing. The scores are compared with each other to fetch the top values and it is made in the ascending order. The top valued books are kept in the order.

The main objective of question analysis module is to derive the expected answer type from the question text. This is a crucial step of the processing since the Answer Extraction module uses a different strategy depending on the expected answer type. Another operation performed by this module is to analyse the query with the purpose of identifying the constraints to be used in the extraction phase.

The Classifiers are formed between the phrases of Cure, prevention and side effects. All these Classifiers are added in the main Database along with its Semantic Words. With these Classifiers we compare with the Medical Journals and we exact the High Relationship between these Classifiers. We also Rank the best and high Classifier occurrence in different Medical Journals and its Rank is also updated in the Data base. User query is passed to the main database and Query extraction is processed and results are tabulated to the user as per the high ranking information.

The results show that the proposed methodology could be integrated in an application to be used in the medical care domain. The framework's capabilities can be used in a commercial system and it is integrated in a new Electronic Health Record system. The proposed methodology will eliminate the need of the physicians to monitor the user databases here we are moving to the machine learning approaches and it also help us to provide the latest medicine technologies since it keep track of up to date medical information from latest published medical abstracts. Our experience with user interfaces and high-performance computing are ideally suited to help healthcare.

3.2 Classification Algorithms

A set of six algorithms are used for classification. The *decision trees* algorithm does not require any prior knowledge about data distribution which works well on noisy data. The decision algorithms are used which are similar to the rule-based approach are suitable for the classification of short texts. *Probabilistic models* which are based on the Naïve Bayes are used in the task of automatic text classification. *Adaptive learning Algorithms* are used to focus on the hard-to-learn concepts. The *Support Vector Machine* based models are used for the task of state-of-art classification on text.

The machine learning algorithms which are used for the representation of data should capture the correlation between the featured sentences. These experimental settings will stand in identifying the informative features in order to increase the chance of predicting the correct labels for the new texts which are to be processed in the future.

3.3 Representation of Data

Bag-of-Words Representation: This representation is commonly used for text classification. It is use ad to represent the features which are chosen among the words. There are two most common feature representations for bag-of-words representation. They are represented in binary feature values can be either 0 or 1, where 1 represent that the instance of the feature is present otherwise 0.

Natural Language Processing Representation (NLP): The NLP representation is based on the syntactic information such as noun-phrases, verb-phrases and biomedical concepts. The Genia tagger is used to extract this type of information. This tagger is specially used for the biomedical text. Here the Genia tagger will ran on the entire data set. Then we extract only the noun-phrases, verb-phrases and biomedical concepts.

Medical Concepts Representation: We use the Unified Medical Language system (UMLS) for the medical concept representation. UMLS is a source of knowledge which is developed at the US National Library of Medicine. UMLS contains about 1 million medical concepts and also about 5 million concepts which are organized hierarchical.

4. PERFORMANCE EVALUATION

The most used common evaluation measures in the machine learning based settings are accuracy, precision, recall and F-measure.

$$Accuracy = \text{total number of correctly classified Instances.}$$

$$Recall = \frac{\text{Correctly classified positive instances}}{\text{The total number of positives}}$$

$$Precision = \frac{\text{Correctly classified positive instances}}{\text{The total number of classified as positive}}$$

F-measure = the harmonic mean between Precision and recall

Figure2 represent the uses of Bag-of-words features, concepts of UMLS, phrases between the noun and verb. The results obtain the 90 percent of F-measure and 90.3 percent accuracy.

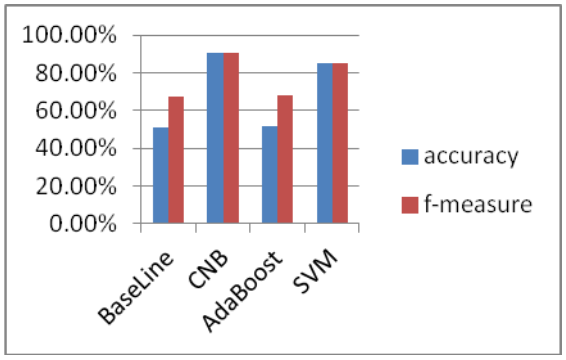

Figure 2. The result of accuracy and F-measure when using the combination of bag-of-words, NLP, biomedical and UMLS concept.

5. CONCLUSION

This paper will give the domain-specific knowledge which improves the results. It provides the models which are stable and reliable for tasks performed on short texts in the medical domain and the representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results.

The first task focused on information retrieval, information extraction and text summarization. Identifying potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts. The second task has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative. It focused on three semantic relations between diseases and treatments.

In order to eliminate the need of the physicians to monitor the user databases here we are moving to the machine learning approaches and it also help us to provide the latest medicine technologies since it keep track of up to date medical information from latest published medical abstracts. It also shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task.

6. FUTURE WORK

As future work we would like to focus the data which comes from the web. Extraction of medical related information from the web is a challenge which can bring the valuable information to the end users and the research community.

7. REFERENCES

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

[2] 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. 2004.

[3] 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. pp. 171-178, 2006.

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

[5] 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] 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.

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

[11] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," Proc. 14th Int'l Conf. Inductive Logic Programming, 2004.