Efficient Data Integration in Finding Ailment-Treatment Relation

[1]A. Nageswara Rao, [2]G. Venu Gopal, [3]K. V. R. Chandra Mouli
[1,3]Dept. of CSE, Pragati Engineering College, Surampalem, Peddapuram, AP, India
[2]Dept. of CSE, Narayana Engineering College, Gudur, AP, India

Abstract
To automatically analyze medical narratives, one needs linguistic and conceptual resources which support capturing of important information from texts and its representation in a structured way. Thus the conceptual structures encoding domain concepts and relations are crucial for the development of reliable and high-performance information extraction system. Machine Intelligence plays a crucial role in the design of expert systems in medical diagnosis. In India most of the people suffering from some sort of diseases like asthma, diabetics, cancer and many more. The Machine Learning field has gained its thrust in almost any domain of research and just recently has become a reliable tool in the medical domain. The experiential domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction, 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, well-organized medical care. It describes a 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. Our evaluation results for these tasks show that the proposed methodology obtains reliable outcomes that could be integrated in an application to be used in the medical care domain. The potential value of this paper stands in the ML settings that we propose and in the fact that we outperform previous results on the same data set.

Keywords
K-Means, Healthcare, Electronic Health Records Machine Learning, Natural Language Processing

I. Introduction

Life is more hectic than has ever been, the medicine that is practiced today is an Evidence-Based Medicine in which medical expertise is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as Google Health and Microsoft Health Vault 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. Electronic Health Records (hereafter, EHR) are becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are:

- Health information recording and clinical data repositories
- Medication management
- Decision support
- Obtain treatments that are tailored to specific health needs.

EHRs can increase the efficiency of your practice and improve quality of care. They can also help maximize reimbursement, and assist in educating and motivating patients. Here's how:

A. Practice Management

1. Integrated scheduling systems (especially useful for practices with multiple health care providers or multiple locations) link appointments directly to progress notes.
2. The health care provider's documentation of the patient's visit automatically generates a list of codes for billing purposes. EHRs then submit and manage claims electronically.

B. Chart Management

1. No more time spent looking for charts or missing information.
2. Multiple staff members with appropriate access privileges can view and modify a single patient's chart simultaneously. No one has to wait for a chart to become available.
3. Centralizing all information in the patient's record can reduce redundant testing.

C. Communication

1. Patients' health information can be accessed from outside the office, which is especially useful in emergencies.
2. Practices can send messages electronically and assign patient-related tasks to other staff members.
3. Staff can submit, track, and receive information from referrals and hospitals.

In order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information. In the medical domain, the richest and most used source of information is Medline,4 a database of extensive life science published articles. All research discoveries come and enter the repository at high rate (Hunter and Cohen [5]), 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 (Medline) as containing or not information about diseases and treatments, and automatically identifying 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.

Our 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—are suitable to use for identifying and classifying relevant medical information in short texts. We acknowledge the fact that 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. In this research, we focus on diseases and treatment information, and the relation that exists between these two entities. Our interests are in line with the tendency of having a personalized medicine, one in which each patient has its medical care tailored to its needs. It is not enough to read and know only about one study that states that a treatment is beneficial for a certain disease. 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.

II. Related Works

The traditional healthcare system is also becoming one that hug the Internet and the electronic world. Electronic Health Records (EHR) is becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are:

- Health information recording and clinical data repositories immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions.
- Medication management rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc.
- Decision support the ability to capture and use quality medical data for decisions in the workflow of healthcare.
- Obtain treatments that are tailored to specific health needs—rapid access to information that is focused on certain topics.

As the above EHR system has some of the flaws in order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information. 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 most relevant related work is the work done by Rosario and Hearst [9]. The authors of this paper are the ones who created and distributed the data set used in our research. The data set consists of sentences from Medline5 abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. The main focus of their work is on entity recognition for diseases and treatments. The authors use Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and the relation discrimination. Their representation techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology—Mesh6 terms. Compared to this work, our research is focused on different representation techniques, different classification models, and most importantly generates improved results with less annotated data. The tasks addressed in our research are information extraction and relation extraction. From the wealth of research in these domains, we are going to mention some representative works. The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: sub cellular location (Craven, [4]), gene-disorder association (Ray and Craven, [4]), and diseases and drugs (Srinivasan and Rindflesch, [1]). Usually, the data sets used in biomedical specific tasks use short texts, often sentences. This is the case of the first two related works mentioned above. The tasks often entail identification of relations between entities that co-occur in the same sentence.

There are three major approaches used in extracting relations between entities: co-occurrences analysis, rule based approaches, and statistical methods. The co-occurrences methods are mostly based only on lexical knowledge and words in context, and even though they tend to obtain good levels of recall, their precision is low. Good representative examples of work on Medline abstracts include Jenssen et al. [2] and Stapley and Benoit [3].

In co-occurrences method the Information retrieval is often divided into two categories: searching and browsing. Searching implies that you have a good to- perfect idea of what you want. Browsing implies that you will be able to recognize what you want when you see it. To determine the various relationships between all of the terms returned from the searching and browsing, the analysis of co-occurrences will give a metric to determine the strength of an association. The strength of the co-occurring terms comes from the number of times two terms occur together within the collection. –T

In Rule-based approach suffer from the fact that the lexicon changes from domain to domain, and new rules need to be created each time. Certain rules are created for biological corpora, medical corpora, pharmaceutical corpora, etc. Systems based on semantic rules applied to full-text articles are described by Friedman et al. [6], on sentences by Pustejovsky et al. [7], and on abstracts by Rindflesch et al. Some researchers combined syntactic and semantic rules from Medline abstracts in order to obtain better systems with the flexibility of the syntactic information and the good precision of the semantic rules, e.g., Gaizauskas et al. [8] and Novichkova et al.

The Statistical methods tend to be used to solve various NLP tasks when annotated corpora are available. Rules are automatically extracted by the learning algorithm when using statistical approaches to solve various tasks. In general, statistical techniques can perform well even with little training data. For extracting relations, the rules are used to determine if a textual input contains a relation or not. Taking a statistical approach to solve the relation extraction problem from abstracts, the most used representation technique is bag-of-words. It uses the words in context to create a feature vector (Donaldson et al.) and (Mitsumori et al.).

A. Introduction to ML

Machine learning (ML) disciplines provide computational methods and learning mechanisms that can help generate new knowledge from large databases. Applications of ML are useful for constructing approaches to solving problems of classification, prediction, recognition patterns, and knowledge extraction, where the data take the form of a set of examples, and the output takes the form of prediction of new examples. In this sense, ML can provide techniques and tools that help solve diagnostic and prognostic problems in medical domains, where the input is a dataset with characteristics of the subjects, and the output is a diagnosis or prognosis of a specific disease. Although diagnosis and prognosis are relatively straightforward ML problems, clinical decision making using ML applications is not yet widely used by the medical community, because such a complex task requires not only accuracy, but also the confidence of physician specialists about the functional use of ML approaches in the medical field.

To successfully implement an ML application in problems related to clinical decisions, it is necessary to consider some specific requirements. For example, the prediction of disease progression is generally associated with the evolution of certain risk factors; in the case of some chronic diseases (e.g., cancer, cardiovascular diseases, and diabetes), the risk factors include non-changeable characteristics, such as age or gender. The use of such non-changeable qualities to predict the onset of a disease might not be as useful for avoiding evolution of the disease, because currently there is no medical treatment for modifying these biological characteristics. Thus, ML applications usually focus on changeable qualities, which make the prognostic task more difficult and complex.

Another important aspect to consider is the need to obtain interpretable approximations, in order to provide medical staff with useful information about the given problem. This is typically achieved using symbolic learning methods (e.g., decision trees and rules systems), which allow decisions to be explained in an easily comprehensible manner. However, the use of a symbolic

learning algorithm to obtain a more comprehensible model frequently sacrifices accuracy in the prediction.

III. Proposed System

The propose system approach, this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques what demonstration of information and what classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts. We recognize the fact that tools able of identifying reliable information in the medical domain stand as construction blocks for a healthcare system that is up-to-date with the latest discoveries. In this examine, we focus on diseases and treatment information, and the relation that exists between these two entities. The approach used to solve the two proposed tasks is based on NLP and ML techniques. In a standard supervised ML setting, a training set and a test set are required. The training set is used to train the ML algorithm and the test set to test its performance.

Table 1: Sentence 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.

Companies that want to sell information technology healthcare frameworks need to build tools that allow them to extract and mine automatically the wealth of published research. For example, in frameworks that make recommendations for drugs or treatments, these recommendations need to be based on acknowledged discoveries and published results, in order to gain the consumers' trust. The product value also stands in the fact that it can provide a dynamic content to the consumers, information tailored to a certain user (e.g., a set of diseases that the consumer is interested in).

The first task (task 1 or sentence selection) identifies sentences from Medline 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).

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, a subset of the eight relations that the corpus is annotated with. We decided to focus on these three relations because these are most represented in the corpus while for the other five, very few examples are available. Table 1 presents the original data set, the one used by Rosario and Hearst [9], that we also use in our research. The numbers in parentheses represent the training and test set size. For example, for Cure relation, out of 810 sentences present in the data set, 648 are used for training and 162 for testing.

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks.

Table 2: Examples of Annotated Sentences for the Sentence Selection Task

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

The models should be reliable at identifying informative sentences and discriminating disease treatment semantic relations. The research experiments need to be guided such that high performance is obtained. The experimental settings are directed such that they are adapted to the domain of study (medical knowledge) and to the type of data we deal with (short texts or sentences), allowing for the methods to bring improved performance.

	Informative sentences	Non-informative sentences
Training set	1225	1176
Test set	612	591

Fig. 1: Data Sets Used for the First Task

	Training		Test	
	Positive	Negative	Positive	Negative
Cure	554	531	276	266
Prevent	42	531	21	266
SideEffect	20	531	10	266

Fig. 2: Data Sets Used for the Second Task

There are at least two challenges that can be encountered while working with ML techniques. One is to find the most suitable model for prediction. The ML field offers a suite of predictive models (algorithms) that can be used and deployed. The task of finding the suitable one relies heavily on empirical studies and knowledge expertise. The second one is to find a good data representation and to do feature engineering because features strongly influence the performance of the models. Identifying the right and sufficient features to represent the data for the predictive models, especially when the source of information is not large, as it is the case of sentences, is a crucial aspect that needs to be taken into consideration. These challenges are addressed by trying various predictive algorithms, and by using various textual representation techniques that we consider suitable for the task.

A. BOW Representation

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which 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 to this feature representation by giving values to each feature for a certain instance. Two most common

feature value representations for BOW representation are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear.

Inhibition	Inhibition	NN	B-NP	O
of	of	IN	B-PP	O
NF-kappaB	NF-kappaB	NN	B-NP	B-protein
activation	activation	NN	I-NP	O
reversed	reverse	VBD	B-VP	O
the	the	DT	B-NP	O
anti-apoptotic	anti-apoptotic	JJ	I-NP	O
effect	effect	NN	I-NP	O
of	of	IN	B-PP	O
isochamaejasmin	isochamaejasmin	NN	B-NP	O
.	.	.	O	O

Fig. 3. Example of Genia tagger output including for each word: its base form, its part-of-speech, beginning (B), inside (I), outside (O) tags for the word, and the final tag for the phrase

Because we deal with short texts with an average of 20 words per sentence, the difference between a binary value representation and a frequency value representation is not large. In our case, we chose a frequency value representation. This has the advantage that if a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capture this—the feature's value will be greater than that of other features. The selected features are words delimited by spaces and simple punctuation marks such as (,) , [,] ,. ,'. We keep only the words that appeared at least three times in the training collection, contain at least one alphanumeric character, are not part of an English list of stop words,10 and are longer than three characters. The frequency threshold of three is commonly used for text collections because it removes non informative features and also strings of characters that might be the result of a wrong tokenization when splitting the text into words. Words that have length of two or one character are not considered as features because of two other reasons: possible incorrect tokenization and problems with very short acronyms in the medical domain that could be highly ambiguous (could be an acronym or an abbreviation of a common word).

B. NLP and Biomedical Concepts Representation

The second type of representation is based on syntactic information: noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. In order to extract this type of information, we used the Genia11 tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags.

The tagger is specifically tuned for biomedical text such as Medline abstracts. Fig. 1, presents an example of the output of the Genia tagger for the sentence: "Inhibition of NF-kappa B activation reversed the anti-apoptotic effect of isochamaejasmin." The noun and verb-phrases identified by the tagger are features used for the second representation technique. We ran the Genia tagger on the entire data set. We extracted only noun-phrases, verb-phrases, and biomedical concepts as potential features from the output of each sentence present in the data set.

The following preprocessing steps are applied in order to identify the final set of features to be used for classification: removing features that contain only punctuation, removing stop words (using the same list of words as for our BOW representation), and considering valid features only the lemma-based forms. We chose to use lemmas because there are a lot of inflected forms (e.g., plural forms) for the same word and the lemmatized form (the base form of a word) will give us the same base form for all of them. Another reason is to reduce the data sparseness problem. Dealing with short texts, very few features are represented in each instance; using lemma forms alleviates this problem. Experiments are performed when using as features only the final set of identified noun-phrases, only verb-phrases, only biomedical entities, and with combinations of all these features. When combining the features, the feature vector for each instance is a concatenation of all features.

C. Algorithm

Here for collecting the data sets and aggregating those data sets we use better clustering algorithm to get the diseases and its treatment relations for performing the above tasks. In ML representation all the data sets (Disease-Treatment relations) are gathered here by using this algorithm and this step is before Bag Of Words(BOW) representation.

Input: A set, V, consisting of n points

Output: A single points x (cluster center) that minimizes the squared error distortion d(V,x) over all possible choices(i.e., a collection of data(disease-treatment relations) from a set of papers) of x

1-Means Clustering problem is easy.

However, it becomes very difficult (NP-complete) for more than one center.

Arbitrarily assign the k cluster centers

1. while the cluster centers keep changing

2. Assign each data point (i.e., disease treatment relation from a particular paper) to the cluster Ci corresponding to the closest cluster representative (center) ($1 \leq i \leq k$)

3. After the assignment of all data points, compute new cluster representatives according to the center of gravity of each cluster, that is, the new cluster representative is

$\Sigma v \setminus |C|$ for all v in C for every cluster C (i.e., disease treatment relation from a particular paper)

This may lead to merely a locally optimal clustering

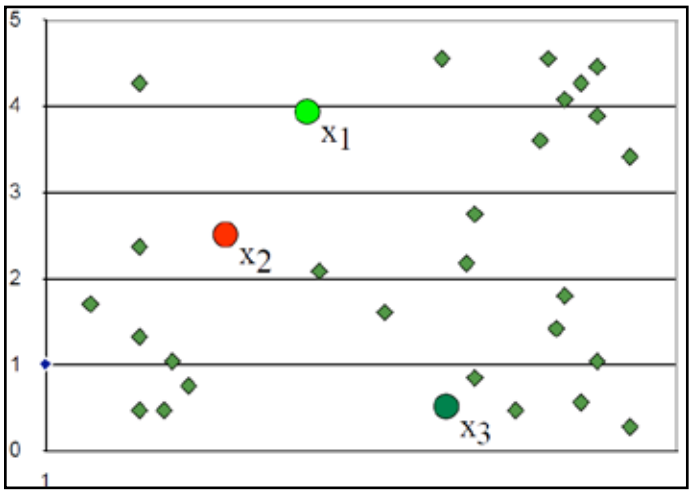

Fig. 4: Clustering of Datasets(i.e., Disease-Treatment Relations) from Different Papers

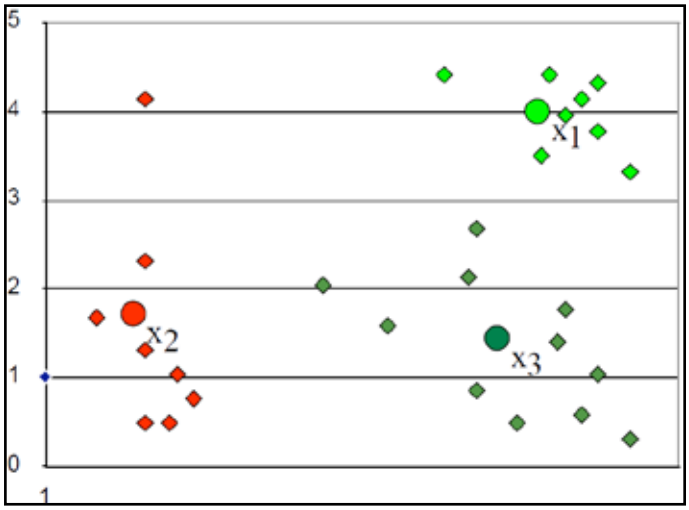

Fig. 5: Clustering of Datasets using K-Means alg.

IV. Conclusion

The conclusions of our study suggest that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results. As future work, we would like to extend the experimental methodology when the first setting is applied for the second task, to use additional sources of information as representation techniques, and to focus more on ways to integrate the research discoveries in a framework to be deployed to consumers. In addition to more methodological settings in which we try to find the potential value of other types of representations, we would like to focus on source data that comes from the web. Identifying and classifying medical-related information on the web is a challenge that can bring valuable information to the research community and also to the end user. We also consider as potential future work ways in which the framework's capabilities can be used in a commercial recommender system and in integration in a new EHR system. Amazon representative Jeff Bezos said: "Our experience with user interfaces and high-performance computing are ideally suited to help healthcare. We nudge people's decision making and behavior with the gentle push of data [. . .]".

References

[1] P. Srinivasan, T. Rindflesch,"Exploring Text Mining from Medline", Proc. Am. Medical Informatics Assoc. (AMIA) Symp., 2002.

[2] T.K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression", Nature Genetics, Vol. 28, No. 1, pp. 21-28, 2001.

[3] B.J. Stapley, 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.

[4] S. Ray, M. Craven,"Representing Sentence Structure in Hidden Markov Models for Information Extraction", Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '01), 2001.

[5] L. Hunter, K.B. Cohen,"Biomedical Language Processing: What's beyond PubMed?", Molecular Cell, Vol. 21-5, pp. 589-594, 2006.

[6] C. Friedman, P. Kra, H. Yu, M. Krauthammer, 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] J. Pustejovsky, J. Castan˜ o, J. Zhang, M. Kotecki, B. Cochran,"Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations", Proc. Pacific Symp. Biocomputing, Vol. 7, pp. 362- 373, 2002.

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

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

[10] T.K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression", Nature Genetics, Vol. 28, No. 1, pp. 21-28, 2001.

Mr.A.Nageswara Rao, well known Author and excellent teacher Received B.E(CSE) from CBIT at Hyderbad and M.Tech (CS) from Hyderbad Centeral University(HCU) is working as Associate Professor and HOD, Department of MCA, M.Tech Computer science engineering , PRAGATI Engineering College, He is an active life member of ISTE. he has 12 years of teaching experience in various engineering colleges. To his credit couple of publications both national and international conferences. His area of Interest includes Data Warehouse and Data Mining, Design and Analysis of Algorithms, Computer Networks.

Mr.K.V.R.Chandra Mouli is a student of PRAGATI Engineering College in Surampalem near Peddapuram.Presently he is pursuing his M.TECH from this college and He received his graduation from Jawaharlal Nehru Technological University In the year 2008.he has a member of CSI.

G.Venu Gopal is working as Associate Professor in NARAYANA Engineering College in gudur,Andhra Pradesh, India. His area of interest in Data Mining.