Medical Diagnosis by Using Relation Extraction

M.Syed Rabiya,
[1,]*Asst professor, M.E ComputerScience and engineering,*
Sethu Institute of Technology .

ABSTRACT

The healthcare information system extracts the sentences from published medical papers that mention group of diseases and treatments, and identifies semantic relations that exist between diseases and treatments.The extracted information is less accurate. My proposed methodology obtains reliable outcomes that could be integrated in an application to be used in the medical care main. The potential value of my paper stands in the ML settings that I propose and in the fact that would outperform previous results on the same data set. The same data set to provide the fact.

INDEX TERMS: *Healthcare, machine learning, natural language processing*

I. INTRODUCTION

Datamining is the process of analyzing the data from different perspectives and summarizing it into useful information.The main aim of the project is Kernel-Based Learning for Biomedical Relation Extraction for helping health care and clinical data repositories. The designing and representation techniques in combination with various learning methods to identify and extract biomedical relations. 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 : immediate access to patient diagnoses, allergies,and lab test results that enable better and time-efficient medical decisions;

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.In order to embrace the views that the EHR system has,we need better, faster, and more reliable access to information. Medline(medical literature analysis and retrieval system online) a database of extensive life science published articles. 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 texts. The second task is focused on three semanticrelations: 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 as stand as building blocks for a healthcare system that is up-to-date with the latest discoveries.

II. RELATED WORK

Prior work is based on entity recognition for diseases and treatments. The data set consists of sentences from Medline abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. prior representation techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology—Mesh6 terms. Compared to this work, my work is focused on different representation techniques,different classification models, and most importantly generates improved results with less annotated data.The task addressed is information extraction and relation extraction.Information extraction is the process of extracting the information from the database..Relation extraction is the process of detecting and classifying semantic relations by given text. Various learning algorithms have been used for the statistical learning approach with kernel based learning is the popular ones applied to Medline abstracts.

III. THE PROPOSED APPROACH

The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments.The first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. 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. Decided to focus on these three relations because these are most represented in the corpus. Table 1 presents the original dataset. 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. The data sets contain sentences that are annotated with the appropriate information. the annotations of the data set are used to create a different task (task 1). 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 NLP and ML techniques can cope with the task of identifying informative sentences, or in other words, how well they can weed out sentences that are not relevant to medical diseases and treatments.

Relationship	Definition and Example
Cure 810 (648, 162)	TREAT cures DIS *Intravenous immune globulin for recurrent spontaneous abortion*
Only DIS 616 (492, 124)	TREAT not mentioned *Social ties and susceptibility to the common cold*
Only TREAT 166 (132, 34)	DIS not mentioned *Flucticasome propionate is safe in recommended doses*
Prevent 63 (50, 13)	TREAT prevents the DIS *Statins for prevention of stroke*
Vague 36 (28, 8)	Very unclear relationship *Phenylbutazone and leukemia*
Side Effect 29 (24, 5)	DIS is a result of a TREAT *Malignant mesodermal mixed tumor of the uterus following irradiation*
NO Cure 4 (3, 1)	TREAT does not cure DIS *Evidence for double resistance to permethrin and malathion in head lice*
Total relevant: 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	Treat and DIS not present *Patients were followed up for 6 months*
Total: 3495 (2793, 702)	

TABLE 1 Dataset description

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. The objectives are to build models that can later be deployed on other test sets with high performance. The task of identifying the three semantic relations is addressed in two ways:

Setting 1. Three models are built. Each model is focused on one relation and can distinguish sentences that contain the relation from sentences that do not. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question (Positive label) or with non relevant information (Negative label);
Setting 2. One model is built, to distinguish the three relations in a three-class classification task where each

sentence is labeled with one of the semantic relations.

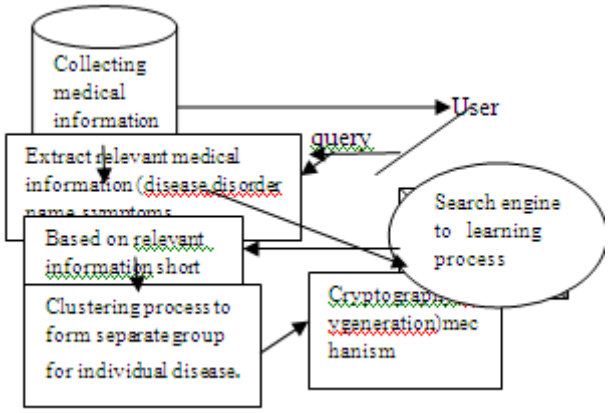
	Informative sentences	Non-informative sentences
Training set	1225	1176
Test set	612	591

TABLE 2 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

TABLE 3 Data Sets Used for the Second Task.
As a result of task2 only informative
sentences are classified into the three semantic relations.

IV. SYSTEM ARCITECTURE



The Medical care related information to collect (e.g., published articles, clinical trials, news, etc.) is a source of power for both healthcare providers and laypeople. The relevant information extracted, which the information is Disease name, disorder name, and symptoms. To identify sentences from Medline published abstracts that talk about diseases and treatments. We are introducing new techniques for learning approach. 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. (For example: disease and treatment information). Semantic model is used to identify the 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 medicine, Prevent medicine, and Side Effect medicine, a subset of the eight relations that the corpus is annotated with.

The people are searching the web and read medical related information in order to be informed about their health.according to user query search engine searches the data and extract relevant information from the database.Search engine mine the data available from the database and return that information to user. classification algorithms are suitable to use for identifying and classifying relevant medical information in short texts. k means clustering is machine learning algorithm used to cluster the observations into groups of related observations without prior knowledge. Support vector machine algoritm used to classify the text into some predefined categories. Advanced encryption standard algorithm(AES) used to generate the random key for the user.In AES algorithm encryption and decryption uses the same key.authentication algorithm provide random key for each user that secure the data for each individual user through that the treatment should not be changed to anyother user. The biomedical system in the architecture used to form the short text for each individual records.If any user enter the query,searches the relevant information according to the query.this relevant

information is applied to the biomedical system to form the short text,based on that information search enginedisplays the result to the user.

V. DATA REPRESENTATIONS

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.

5.1 Bag-of-Words 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.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.

5.2 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. 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.

5.3 Medical Concepts (UMLS) Representation

The Unified Medical Language system12 (hereafter, UMLS) concept representations. UMLS is a knowledge source developed at the US National Library of Medicine (hereafter, NLM) and it contains a metathesaurus, a semantic network, and the specialist lexicon for biomedical domain. The metathesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts. UMLS contains over 1 million medical concepts, and over 5 million concept names which are hierarchical organized.All concepts are assigned at least one semantic type from the semantic network providing a generalization of the existing relations between concepts. There are 135 semantic types the knowledge base that are linked through 54 relationships.

5.4.Evaluation Measures

The most common used evaluation measures in the ML settings are: accuracy, precision, recall, and F-measure. The formulas for the evaluation measures are: Accuracy =the total number of correctly classified instances; Recall = the ratio of correctly classified positive instances to the total number of positives. This evaluation measure is known to the medical research community as sensitivity. Precision = the ratio of correctly classified positive instances to the total number of classified as positive. F-measure = the harmonic mean between precision
and recall.

VI. CONCLUSION

This paper provides the framework for ML,NLP techniques to classify the text. In NLP and ML community, BOW is a representation technique that even though it is simplistic, most of the times it is really hard to outperform. we outperform it when we combine it with more structured information such as medical and biomedical concepts. the simple BOW approach, well known to give reliable results on text classification tasks, can be significantly outperformed when adding more complex and structured information from various ontologies.

REFERENCES

[1] 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.

[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.(TREC), 2004.

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

[4] 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.

[5] R. Bunescu, R. Mooney, Y. Weiss, B. Scho¨ lkopf, and J. Platt, "Subsequence Kernels for Relation Extraction," Advances in Neural Info, 2006.