E. Madhusudhana Reddy et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 1, March 2018, pp. 223-226

Able Machine Learning Method for classifying Disease- Treatment Semantic Relations from Bio-Medical Sentences

E. Madhusudhana Reddy[1], P. Bhaskar[2]

Professor Dept. of CSE[1], DRK College of Engineering & Technology, Hyderabad[1]

***Abstract*— 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, 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, more efficient medical care. This paper 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.. 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. 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.**

***Index Terms*—Healthcare, machine learning, natural language processing.**

I. INTRODUCTION

People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been, the medicine that is practiced today is an Evidence-Based Medicine (hereafter, EBM) 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 Health1 and Microsoft HealthVault2 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 are3: 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 health care

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 [1]), 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.

II. RELATED WORK

The most relevant related work is the work done by Rosario and Hearst [3]. 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: subcellular location (Craven, [4]), gene-disorder association (Ray and Craven, [5]), and diseases and drugs (Srinivasan and Rindflesch, [6]). 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.

III. THE PROPOSED APPROACH

A. Tasks and Data Sets

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 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 reviews8 (hereafter, SR), or laypeople who want to be in charge of their health by reading the latest life science published articles related to their interests. 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. The product can be developed and sold by companies that do research in Healthcare Informatics, Natural Language Processing, and Machine Learning, and companies that develop tools like Microsoft Health Vault. The value of the product from an e-commerce point of view stands in the fact that it can be used in marketing strategies to show that the information that is presented is trustful (Medline articles) and that the results are the latest discoveries. For any type of business, the trust and interest of customers are the key success factors. Consumers are looking to buy or use products that satisfy their needs and gain their trust and confidence. Healthcare products are probably the most sensitive to the trust and confidence of consumers. The first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. 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 [3], 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.

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

B Classification Algorithms and Data Representations

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks. 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. 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. As classification algorithms, we use a set of six representative models: decision-based models (Decision trees), probabilistic models (Naive Bayes (NB) and Complement Naive Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive learning (Ada-Boost), a linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier that always predicts the majority class in the training data (used as a baseline). We decided to use these classifiers because they are representative for the learning algorithms in the literature and were shown to work well on both short and long texts. Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts. Probabilistic models, especially the ones based on the Naive Bayes theory, are the state of the art in text classification and in almost any automatic text classification task. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets. SVM- based models are acknowledged state-of-the-art classification techniques on text. All classifiers are part of a tool called Weka.9 One can imagine the steps of processing the data (in our case textual information—sentences) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection (sentences in our case). The most difficult and important step is to identify.

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

Fig 1: Data Sets Used for the Second Task

C 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. 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.

D 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 sochamaejasmin." The noun and verb-phrases identified by the tagger are features used for the second representation technique.

E Medical Concepts (UMLS) Representation

In order to work with a representation that provides features that are more general than the words in the abstracts (used in the BOW representation), we also used 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 met thesaurus, a semantic network, and the specialist lexicon for biomedical domain. The met thesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts.

F. EHR System

An electronic health record (EHR) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations. It is a record in digital format that is theoretically capable of being shared across different health care settings. In some cases this sharing can occur by way of network-connected enterprise-wide information systems and other information networks or exchanges. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal stats like age and weight, and billing information.

G EHRs Respond to the Complex AMC Environments

The major value of integrated clinical systems is that they enable the capture of clinical data as a part of the overall workflow. An EHR enables the administrator to obtain data for billing, the physician to see trends in the effectiveness of treatments, a nurse to report an adverse reaction, and a researcher to analyze the efficacy of medications in patients with co-morbidities. If each of these professionals works from a data silo, each will have an incomplete picture of the patient's condition. An EHR integrates data to serve different needs. The goal is to collect data once, then use it multiple times. EHRs are used in complex clinical environments. Features and interfaces that are very appropriate for one medical specialty, such as pediatrics, may be frustratingly unusable in another (such as the intensive care unit). The data presented, the format, the level of detail, and the order of presentation may be remarkably different, depending on the service venue and the role of the user. Scot M. Silverstein, MD, of Drexel University, stated "Clinical IT projects are complex social endeavors in unforgiving clinical settings that happen to involve computers, as opposed to IT projects that happen to involve doctors."

H. Physicians, Nurses, and Other Clinicians

EHR workflow implications for healthcare clinicians (physicians, nurses, dentists, nurse practitioners, etc.) may vary by type of patient care facility and professional responsibility. However, the most cited changes EHRs foster involve increased efficiencies, improved accuracy, timeliness, availability, and productivity (See references 1, 8, and 9 in the References section).

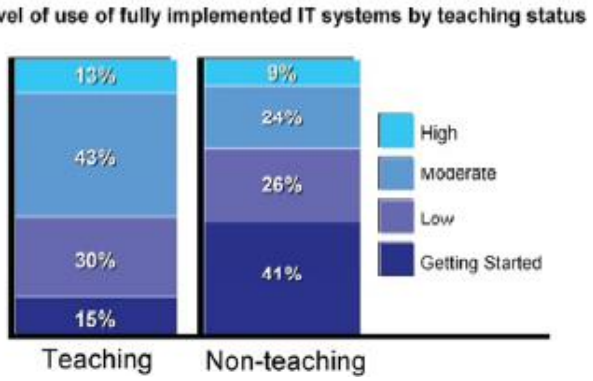

Fig 2: Level of HER related technology used by teaching

Hospitals Clinicians in environments with EHRs spend less time updating static data, such as demographic and prior health history, because these data are populated throughout the record and generally remain constant. Clinicians also have much greater access to other automated information (regarding diseases, etc.), improved organization tools, and alert screens. Alerts are a significant capacity of EHRs because they identify medication allergies and other needed reminders. For clinical researchers, alerts can be established to assist with recruitment efforts by identifying eligible

research participants. Workflow redesigns to ensure increased efficiencies, to generate improvements in quality of care, and to realize the maximum benefits of an automated environment. NIH Challenges that EHRs may present to workflow processes include: increased documentation time (slow system response, system crashes, multiple screens, etc.), decreased interdisciplinary communication, and impaired critical thinking through the overuse of checkboxes and other automated documentation. System crashes are particularly problematic because clinicians, particularly at in-patient facilities, will not know what treatments are needed or if medications are due. Interestingly, the national attention and rapid adoption of EHRs come at a time when the nursing industry is experiencing a substantial decrease in workforce and an increase in workload. To help compensate for this workforce discrepancy, EHR implementations must coincide with

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. The first task that we tackle in this paper is a task that has applications in information retrieval, information extraction, and text summarization. We identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts. We show that 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. The second task that we address can be viewed as a task that could benefit from solving the first task first. In this study, we have focused on three semantic relations between diseases and treatments. Our work shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task. Also, to perform a triage of the sentences (task 1) for a relation classification task is an important step. In Setting 1, we included the sentences that did not contain any of the three relations in question and the results were lower than the one when we used models trained only on sentences containing the three relations of interest. These discoveries validate the fact that it is crucial to have the first step to weed out uninformative sentences, before looking deeper into classifying them. Similar findings and conclusions can be made for the representation and classification techniques for task 2. The above observations support the pipeline of tasks that we propose in this work. The improvement in results of 14 and 18 percentage points that we obtain for two of the classes in question shows that a framework in which tasks 1and 2 are used in pipeline is superior to when the two tasks are solved in one step by a four-way classification. Probabilistic models combined with a rich representation technique bring the best results.

REFERENCES

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

[2] J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb. 2009.

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

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

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

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

[8] I. Donaldson et al., "PreBIND and Textomy: Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine," BMC Bioinformatics, vol. 4, 2003.