Identifying Disease Treatment Relations using F.P Tree

Pratik K. Bobade[1] Vikrant G. Rekulwad[2] Vikram S. Chaudhari[3] Shweta Barshe[4]
[1, 2, 3]Student [4]Professor
[1, 2, 3, 4]Computer Engineering Department
[1, 2, 3, 4]Bharati Vidyapeeth College Of Engineering Navi Mumbai, India

Abstract--- The Machine Learning (ML) field has been one of the most important tool in the health care field. Automated learning is used in tasks such as medical decision support, medical picturing, 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 reports a ML-based methodology for creating a program that is capable of finding and extracting healthcare information. It allows researcher to add medical papers that mention diseases and treatments, and identifies relations that exist between diseases and treatments. Our proposed methods obtains satisfactory outcomes that could be integrated in an application to be used in the medical care domain. The potential value of this paper is in the settings that we propose.

Keywords: automatic learning; medical care; machine learning;

I. INTRODUCTION

People are worried about their health and want to be, now more than ever, take hold of their health and healthcare. Life has become busy than it ever had 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 researches as well. Applications that can help us manage and better keep track of our health such as Google Health and Microsoft HealthVault 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, dedication management and decision support. It is better to identify and eliminate first the sentences that do not contain relevant information, and then classify the rest of the sentences by the relations of interest, instead of doing everything in one step by classifying sentences into one of the relations of interest plus the extra class of uninformative sentences.

II. RELATED WORK

The most relevant related work is the work done by Rosario and Hearst [25]. 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 Medline 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: subcellularlocation (Craven, [4]), gene-disorder association (Ray and Craven, [23]), and diseases and drugs (Srinivasan and Rindflesch, [26]). 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, rulebased 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. [14] and Stapley and Benoit [27]. In biomedical literature, rule-based approaches have been widely used for solving relation extraction tasks. The main sources of information used by this technique are either syntactic: part-of-speech (POS) and syntactic structures; or semantic information in the form of fixed patterns that contain words that trigger a certain relation. One of the drawbacks of using methods based on rules is that they tend to require more human-expert effort than data-driven methods (though human effort is needed in data-driven methods too, to label the data). The best rule-based systems are the ones that use rules constructed manually or semiautomatically—extracted automatically and refined manually. A positive aspect of rule-based systems is the fact that they obtain good precision results, while the recall levels tend to be low. Syntactic rule-based relation extraction systems are complex systems based on additional tools used to assign POS tags or to extract syntactic parse trees. It is known that in the biomedical literature such tools are not yet at the state-of-the-art level as they are for general English texts, and therefore their performance on sentences is not always the best (Bunescu et al. [2]). Representative works on syntactic rule-based approaches for relation extraction in Medline abstracts and full-text articles are presented by Thomas et al. [28], Yakushiji et al. [29], and Leroy et al. [16]. Even though the syntactic information is

the result of tools that are not 100 percent accurate, success stories with these types of systems have been encountered in the biomedical domain. The winner of the BioCreative II.57 task was a syntactic rule-based system, OpenDMAP described in Hunter et al. [13]. A good comparison of different syntactic parsers and their contribution to extracting protein-protein interactions can be found in Miyao et al. [19]. The semantic rule-based approaches 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. [22], and on abstracts by Rindflesch et al. [24]. 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. [20]. 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. [5]) and (Mitsumori et al. [18]). Other researchers combined the bagof- words features, extracted from sentences, with other sources of information like POS (Bunescu and Mooney [1]). Giuliano et al. [9] used two sources of information: sentences in which the relation appears and the local context of the entities, and showed that simple representation techniques bring good results. Various learning algorithms have been used for the statistical learning approach with kernel methods being the popular ones applied to Medline abstracts (Li et al. [17]). The task of identifying informative sentences is addressed in the literature mostly for the tasks of summarization and information extraction, and typically on such domains as newswire data, novels, medical, and biomedical domain. In the later mentioned domains, Goadrich et al. [11] used inductive logic techniques for information extraction from abstracts, while Ould et al. [21] experimented with bag-of-word features on sentences. Our work differs from the ones mentioned in this section by the fact that we combine different textual representation techniques for various ML algorithms.

III. PROPOSED METHOD

In our proposed method we are requesting different researchers to provide us with their knowledge about different diseases, in the predefined format, the administrator will then verify this and has the authority to accept or reject his research. If accepted the provided symptoms will be added into the database under the particular disease. Every time such research is added into the database a new frequent pattern tree will be generated with the updated information. Each symptom will have its frequency evaluated every time new update is made. Based on these frequency frequent pattern tree will be generated. The methods that were used previously mostly considered using sequential searches to map symptoms to disease. Instead we proposed to use a frequent pattern tree to map symptoms to diseases, which overcomes the disadvantages of sequential search. Sequential search will fail if the user don't enter all the symptoms he/she is having returning no result. Our method goes one more step ahead and will act like a doctor i.e. it will provide user with other optional symptoms until the system has enough symptoms to conclude the disease based on the confirmed symptoms. Frequent pattern tree is better than apriori algorithm were large database is concerned. Once the disease is mapped to the symptoms information about the particular disease will be displayed, this includes the most frequent cause of the disease, its prevention and most common cure method.

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 distribute healthcare information. The first task is to collect or extracts information 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 relate 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. We focus on three relations: Cure, Prevent, and Cause, 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 [25], 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 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.

For the work presented in this paper, the data sets contain sentences that are annotated with the appropriate information. Unlike in the work of Rosario and Hearst [25], in our research, 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

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: Data Set Description, Taken from Rosario and Hearst (’04)

informative sentences, or in other words, how well they can weed out sentences that are not relevant to medical diseases and treatments. Extracting informative sentences is a task by itself in the NLP and ML community. Research fields like summarization and information extraction are disciplines where the identification of informative text is a crucial task. The contributions and research value that are brought with this task stand in the usefulness of the results and the insights about the experimental settings for the task in the medical domain. For the first task, the data sets are annotated with the following information: a label indicating that the sentence is informative, i.e., containing disease-treatment information, or a label indicating that the sentence is not informative. For the second task, the sentences have annotation information that states if the relation that exists in a sentence between the disease and treatment is Cure, Prevent, or Cause. These are the relations that are more represented in the original data set and also needed for our future research. We would like to focus on a few relations of interest and try to identify what predictive model and representation technique bring the best results. 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. Up to this point, we presented the two tasks separately as being two self-defined tasks since they can be used for other more complex tasks as well. From a methodological point of view, and, more importantly, from a practical point of view, they can be integrated together in a pipeline of tasks as a solution to a framework that is tailored to identify semantic relations in short texts and sentences, when it is not known a priori if the text contains useful information. The proposed pipeline solves task 1 first and then processes the results in task 2, so that in the end, only informative sentences are classified into the three semantic relations. The logic behind choosing to experiment with and report results for the pipeline of tasks is that we have to identify the best model that will get us closer to our main goal: being able to identify and classify reliably medical information. Using the pipeline of tasks, we eliminate some errors that can be introduced due to the fact that we would consider uninformative sentences as potential data when classifying sentences directly into semantic relations. We will show that the pipeline achieves much better results than a more straightforward approach of classifying in one step into one of the three relations of interest plus an extra class for uninformative sentences.

The pipeline is similar to a hierarchy of tasks in which the results of one task is given as input to the other. We believe that this can be a solution for identifying and disseminating relevant information tailored to a specific semantic relation because the second task is trying a finer grained classification of the sentences that already contain information about the relations of interest.

This framework is appropriate for consumers that tend to be more interested in an end result that is more specific, e.g., relevant information only for the class Cure, rather than identifying sentences that have the potential to be informative for a wider variety of disease-treatment semantic relations.

IV. Acknowledgment

Our most sincere appreciation are to all the people who has helped and inspired us throughout the working of this project. Firstly we are thankful to our principal Dr. M. Z. Shaikh for his help. We are extremely grateful for his friendly support and professionalism.

We express our heartfelt gratitude to our Head of Department Prof. Vidya Chitre and our seminar coordinator Prof. B. W. Balkhande for their help and support. This task would not have been possible without the help and guidance of our project guide Prof. Mrs. Shweta Barshe.

We are also convening special thanks to all staff of Computer Engineering Department for their support and help. Last but not least, we are very much thankful to our friends who directly or indirectly helped us in completion of the technical paper.

V. 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 1 and 18 percentage points that we obtain for two of the classes in question shows that a framework in which tasks 1 and 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. 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

VI. REFERENCES

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

[2] 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. 171-178, 2006.

[3] 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.

[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] C. Giuliano, L. Alberto, and R. Lorenza, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics, 2006.

[10] 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.

[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.

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

[13] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson,P.V. Ogren, and K.B. Cohen, "OpenDMAP: An Open Source,Ontology-Driven Concept Analysis Engine, with Applications to Capturing Knowledge Regarding Protein Transport, Protein Interactions and Cell-Type-Specific Gene Expression," BMC Bioinformatics, vol. 9, article no. 78, Jan. 2008.

[14] T.K. Jenssen, A. Laegreid, J. Komorowski, and 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.

[15] R. Kohavi and F. Provost, "Glossary of Terms," Machine Learning, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, pp. 271-274, 1998.

[16] G. Leroy, H.C. Chen, and J.D. Martinez, "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text," J. Biomedical Informatics, vol. 36, no. 3, pp. 145-158, 2003.

[17] J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," J. Am. Soc. Information Science and Technology, vol. 59, no. 5, pp. 756-769, 2008.

[18] T. Mitsumori, M. Murata, Y. Fukuda, K. Doi, and H. Doi, "Extracting Protein-Protein Interaction Information from Biomedical Text with SVM," IEICE Trans. Information and Systems, vol. E89D, no. 8, pp. 2464-2466, 2006.

[19] M. Yusuke, S. Kenji, S. Rune, M. Takuya, and T. Jun'ichi,"Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction," Bioinformatics, vol. 25, pp. 394-400, 2009.

[20] S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, A Natural Language Processing Engine for MEDLINE Abstracts," Bioinformatics, vol. 19, no. 13, pp. 1699-1706, 2003.

[21] M. Ould Abdel Vetah, C. Ne´dellec, P. Bessie`res, F. Caropreso, A.-P. Manine, and S. Matwin, "Sentence Categorization in Genomics Bibliography: A Naive Bayes Approach," Actes de la Journe´e Informatique et Transcriptome, J.-F. Boulicaut and M. Gandrillon, eds., Mai 2003.

[22] J. Pustejovsky, J. Castan˜ o, J. Zhang, M. Kotecki, and B. Cochran, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," Proc. Pacific Symp. Biocomputing, vol. 7, pp. 362-373, 2002.

[23] 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.

[24] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter, "EDGAR: Extraction of Drugs, Genes, and Relations from the Biomedical Literature," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 514-525, 2000.

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

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

[27] 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.

[28] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic Extraction of Protein Interations from Scientific Abstracts," Proc. Pacific Symp. Biocomputing, vol. 5, pp. 538-549, 2000.

[29] Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event Extraction from Biomedical Papers Using a Full Parser," Proc. Pacific Symp. Biocomputing, vol. 6,pp.408-419,2001.

