A Framework for Identifying Disease Treatment Relations Using Classification Algorithm

Mr.R.Srinivasan M.E.,
Assistant professor/CSE
Muthayammal College of Engineering

Mr.K.Hariprasath M.E.,
Assistant professor/CSE
Muthayammal College of Engineering

Abstract:- *The Machine Learning based methodology for building an application that is capable for identifying and disseminating health-care information.The first task identifies and extracts informative sentences on diseases and treatment topics,while the second one performs a finer grained classification of these sentences according to the semantic relations that exist between diseases and treatments.Evaluation results for these tasks show that the proposed methodology obtains reliable outcomes that could be integrated in an application to be used in medical care domain.*

Keywords: Bow, Healthcare, Machine Learning, Natural Language processing

1. 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.Researches and studies show that the potential benefits of having an Electronic Health Records(EHR) system in areas such as 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.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, a database of extensive life science published articles. 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. The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. People want fast access to reliable information and in a manner that is suitable to their habits and workflow. Medical care related information (e.g., published articles, clinical trials, news, etc.) is a source of power for both healthcare providers and laypeople.

2. Collection of Medical Related Data

The medical related data set will be the input of our project. We are going to collect the medical related data set from

http://www.nlm.nih.gov/medlineplus/all_healthtopics.html

Table 1: Data Set Description, Taken from Rosario and Hearst ('04)

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)	

3. Classification Algorithms

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. So we use prefer decision models in our project for data representations. The ML algorithms that are using this data representation to create predictive models should capture correlations between features, feature values, and labels, in order to obtain good prediction labels on future test data.

4. 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. 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).

Text Document Representation based on the BoW model

The text document representation based on the BoW model in NLP is reviewed first. Here are two simple text documents:

- John likes to watch movies. Mary likes too.
- John also likes to watch football games.

Based on these two text documents, a dictionary is constructed as:

Dictionary={1:"John", 2:"likes", 3:"to", 4:"watch", 5:"movies", 6:"also", 7:"football", 8:"games", 9:"Mary", 10:"too"},

which has 10 distinct words. And using the indexes of the dictionary, each document is represented by a 10-entry vector:

- [1, 2, 1, 1, 1, 0, 0, 0, 1, 1]
- [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

where each entry of the vectors refers to count of the corresponding entry in the dictionary (This is also the histogram representation).

5. Medical Concepts (UMLS) Representation

UMLS is a knowledge source developed at the US National Library of Medicine (hereafter, NLM) and it contains a Meta thesaurus, a semantic network, and the specialist lexicon for biomedical domain. The Meta thesaurus 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. For each of the noun-phrases that the system finds in the text, variant noun-phrases are generated. For each of the variant noun-phrases, candidate concepts (concepts that contain the noun-phrase variant) from the UMLS Meta thesaurus are retrieved and evaluated. The retrieved concepts are compared to the actual phrase using a fit function that measures the text overlap between the actual phrase and the candidate concept (it returns a numerical value). The best of the candidates are then organized according to the decreasing value of the fit function. We used the top concept

candidate for each identified phrase in an abstract as a feature.

6. Semantic Matching of Medical Concept

The focus for the second task is to automatically identify which sentences contain information for the three semantic relations: Cure, Prevent, and Side Effect. The reported results are based on similar settings to the ones used for the previous task. Since imbalanced data sets are used for this task, the evaluation measure that we are going to report is the F-measure. Due to space issues, we are going to present the best results obtained for all settings. The best results are chosen from all the representation techniques and classification algorithms that we also used for the first task. The labels on the x-axis stand for the name of the semantic relation, the representation technique, and the classification algorithm used.

7. Aiding Medical Decision Support System

For this module we are getting input from the user and according to the user input.Our decision support medical system will extraction of medical knowledge from our medical repository. The decision support is the ability to capture very quality medical data for medical repository. Our decision support fully depend on two tasks: automatically identifying sentences published in medical abstracts (medical repository) 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 and give a full reliable decision to the user.

8. Performance Comparison

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 project 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 ontology. 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 .So 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.

9. Conclusion

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.

10. References

[1]Oana Frunza, Diana Inkpen, and Thomas Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts," vol. 23, no. 6, June 2011

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

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

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

[6] http://medline.cos.com/