International Journal of Innovative Research in Advanced Engineering (IJIRAE)
Volume 1 Issue 3 (May 2014) SPECIAL ISSUE

Review On A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts

Jagtap J.J.
ME computer Engineering
Pune University
jagtapjayanti@gmail.com

Kulkarni N.J.
ME computer Engineering
Pune University
nikita.kulkarni @zealeducation.com

Abstract— The Machine learning used in many application in almost every area. In this electronic world, everything has to be automated to improve quality, time complexity, accuracy, efficiency etc. In the biomedical area, information is mainly in natural language text format. The biomedical researchers need fast information accessing tools for extracting useful information from lager amount of biomedical repositories. This paper provides a survey of machine learning and various natural language text representations method that are used currently. The main purpose of this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques used for representation of information and what classification algorithms are suitable for identifying and classifying relevant biomedical information in short texts. This work focuses on retrieval of updated, accurate and relevant information from Medline datasets using Machine-Learning approach. It is also used to identifying relationship extraction from biomedical text. This will help to understand appropriate feature representation and machine learning technique suitable for the medical domain.

Keywords— Healthcare, machine learning, natural language processing, Disease Treatment Extraction, Medline

I. INTRODUCTION

Machine Learning is used in all domains of research and medical fields. Machine learning that are used for the construction and study of systems that can learn from data. It is a method of teaching computers to make and improve predictions or behaviors based on some data. Machine learning is a huge field with have hundreds of different algorithms for solving different problems. Machine learning seems challenging problems in terms of algorithmic approach, data representation, computational efficiency, and quality of the resulting program.

Relation extraction from natural language text is huge research area.Heathcare information are stored in natural language text format. Information is growing considerably every day. With the increased amount of biomedical publications, it is becoming more and more challenging to access useful and relevant information about a specific topic. 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. Manual inspection of such large amount of data will be very difficult and time consuming.

Natural Language Processing (NLP) and Machine Learning (ML) techniques is to show what representation of information and algorithms used to identify and provide relevant biomedical information in short texts. This technique is used to identify all medical related information and to built healthcare system that is up-to-date with all latest discoveries. The main objective is to focus on all information related to disease.

ML techniques the information is shown in short texts when identifying relations between diseases and treatment. There is improvement in solutions when using a pipeline of two tasks. It is better to identify and remove the sentence that does not contain information relevant to disease or treatments.

II. LITERATURE SURVEY

In Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger ,"**Tackling The POOR Assumption Of Naïve Bayes Text Classifier"** this paper described text classification by using naïve bayes text classifier but this text classifier does not give precision 100% for output. Sometimes prediction of classifier may be incorrect. [1]

In T.Mouratis, S.Kotsiantis, "**Increasing The Accuracy Of Discriminative Of Multinomial Bayesian Classifier In Text Classification",** there were introduced use of classifier that increased precision of output but the main problem in that work was at the time of classification it doesn't identify the verbs, nouns, adjectives,pharse properly so many time it may provided wrong data.[2]

In B.Rosario and M.A. Hearst**, "Semantic Relation in Bioscience Text"** In this paper author used Hidden Markov models for entity recognition. This includes mapping medical information into structural representation. It converts natural language text into structural format. Also they use machine learning for information extraction.

Text classification is used for the extraction of biomedical abstract. Semantic lexicons of words labeled with semantic classes so associations can be drawn between words which helps in extracting the necessary sentences related to the query. In this paper sentence co-occurrence and navie bayes algorithm are used for extracting semantic relation like Gene-Protein from Medline abstract, the precision and recall of the result obtained are shown in the graph but due to use of only one navie bayes algorithm it do not get good precision of output and it doesn't used bag of words to find adjective, verbs nouns phrase while doing classification.[3]

In M.Craven, **"Learning To Extract Relations from Medline"** In this research paper the individual sentences are considered as instances that are to be processed by the navie bayes classifier. Here each sentence is considered as positive training set. Relation extractions are made through relational learning. Extraction of words from Medline abstract has been done by using navie bayes and CNB algorithms. It also used bag of words during classification but it is not used natural language processing due to this performance of output degrades.[4]

In L. Hunter and K.B. Cohen**, "Biomedical Language Processing: What's Beyond Pubmed"** this system is used natural language processing for processing of biomedical words. In this work it takes the name of disease and give the solution which has been stored in local database of that disease by parsing user statement using natural language processing but it does not do diagnosis of disease.[5]

In Jeff Pasternack, Don Roth **"Extracting Article Text From Web With Maximum Subsequence Segmentation"** It involves to extract word according to occurrence of that word in article if no of word occur by no of time mentioned then extract that word from the web here author used bag of word to remove verbs and adjective from the article but it doesn't use Natural language processing while extracting.[6]

In Abdur Rehman, Haroon.A.Babri, Mehreen saeed, **"Feature Extraction Algorithm for Classification of Text Document"** It involves automatic extraction of semantic relation between medical related points. A dictionary of biomedical terms is used for sentence classification. The sentences are automatically parsed using semantic parser by using four classification algorithm such as NB, CNB, Decision tree, Adaptive, SVM etc while extracting word but it doesn't provided the information regarding diagnosis of disease.[7]

In Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl**., "Efficient Text Extraction Algorithm Using Color Clustering For Language Translation In Mobile Phone"** Auther used AdaBoost classifier is outperformed by other classifier**.** SVM classifier is always functions well when the information matches with the training set. Probabilistic model are used to perform text classification task. Bag of word technique is simple in nature but in many time it is hard to outperform it. Pipelining task is essential to obtain increased quality of result because majority functions may overcome the underrepresented ones. By using pipelining there is a balance between relevant and irrelevant data and the classifier has better chance to distinguish relevant and irrelevant data but it don't used Gennia tagger tool which is special parser for biomedical words.[8]

In Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "**A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts**" In this work it involves two task in pipelined manner for identifying and extracting the relationship between the given MEDLINE abstract. Firs task involves finding most suitable model for prediction, the second task is to find good data representation. To achieve this two task various predictive algorithm and textual representation techniques are considered. A set of six classification algorithms namely decision based models, probabilistic models, Adaptive learning, linear classifier like support vector machine and a classifier that always predicts the majority class in training data are used. Three representation technique namely Bag-Of-Word representation, NLP and Biomedical Concept representation and Medical concept representation are used to obtain the treatment relation from short text. Various experiments are conducted with the combination of the six classification algorithm and three representation techniques. The results are shown in bar chart form. As the result of the experiment it is concluded that bag-of-representation when combined with any of six classification algorithm produces better results but it does not give disease diagnosis as well information about particular disease by parsing statement.[9]

III. BIOMEDICAL RESOURCES

For biomedical information, the primary resource is MEDLINE. It is a bibliographic database of life sciences and biomedical information. This database contains more than 21 million records from over 5ooo selected publications. It includes bibliographic information for articles from academic journals that include medicine, nursing, pharmacy, dentistry, and health care. It is developed by the United States National Library of Medicine (NLM). MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrees system [12]. There are many sources for annotated MEDLINE abstracts.

IV. FEATURE REPRESENTATION

Feature selection is most important in machine learning approach. Important features are used for distinguish between the relation types or entity pairs. In this work, it is to be stated that, the more the features are included in the representation, then the more accurate will be the classification.

A. Bag of Words Representation

The bag-of-words representation is 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 to identify the most suitable words as features. After the feature 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 and frequency feature value. Binary feature value is used 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. Frequency feature value is used the value of the feature is the number of times it appears in an instance, or 0 if it did not appear

B. NLP and Biomedical Concepts Representation

This type of representation is based on syntactic information such as noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. Genia tagger tool is used to extract information. The tagger is specifically designed for medical text such as Medline abstracts the tagger analyzes English sentences and to generate the base forms, part-of-speech tags, chunk tags, and named entity tags. Genia tagger it extracts only nouns, phrases. verb and healthcare related concepts from each sentence of medical data set.

C. Medical Concepts (UMLS) Representation

UMLS is a knowledge source developed at the US National Library of Medicine and it contains over 1 million medical concepts and 5 million concept names. It contains a metathesaurus, a semantic network and a specialist lexicon for the biomedical domain. The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database and it contains information about biomedical and health related concepts, their various names, and the relationships among them.

V. RELATION CLASSIFICATION

Relation Classification task is used to identify occurrences of particular types of relationships between pairs of given entities and classify them according to the type of relationship. There are three approaches used in extracting relations between entities: co-occurrences analysis, rule based approaches, and statistical methods.

A. Co-Occurrences Analysis

The co-occurrences methods are based on lexical knowledge and words in context, and even though they tend to obtain good levels of recall, but their precision is low. Good representative examples of work on Medline abstracts include Jenssen et al. [11] and Stapley and Benoit [13].

B. Rule based approaches

In biomedical literature, rule-based approaches have been 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 these methods is that they tend to require more human-expert effort than data-driven methods. The best rule-based systems are the ones that use rules constructed semi automatically or manually that is extracted automatically and refined manually. A positive aspect of rule-based systems is that they obtain good precision results, while the recall levels tend to be low.

C. Statistical Methods

Statistical methods are 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 used representation technique is bag-of-words. It uses the words in context to create a feature vector and Other researchers combined the bag of- words features, extracted from sentences, with other sources of information like POS 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 statistical tools.

VII. CONCLUSION

Different machine learning method and their advantages and disadvantages are applied to biomedical literature are discussed here. Different feature representation techniques are also study and listed. For two class classification tasks SVM is found to be the best and for multiple classification problems, Naïve Bayesian techniques are performed well. We suggest multiple classifier system for improving the accuracy and efficiency of the system. In the multiple classifier system different member classifiers used should have different errors independent of each other and should have performance better than a minimum level. That is, for the method to be successful each member classifier should keep a minimum level of disagreement. By averaging the results of a large number of such classifiers, decision boundary can be approximated with some accuracy. That is uncorrelated errors of individual classifiers can be eliminated by averaging.

REFERENCES

[1] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, "Tackling the POOR Assumption of Naïve ayes Text Classifier", Proceedings Of The Twentieth International Conference On Machine Learning (ICML-2003), Washington DC, 2003.
[2] T.Mouratis, S.Kotsiantis, "Increasing the Accuracy of Discriminative of Multinomial Bayesian Classifier in Text Classification", Classification", ICCIT'09 Proceedings Of The 2009 Fourth International Conference On Computer Science and Convergence Information Technology
[3] B.Rosario and M.A.Hearst, "Semantic Relation 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.
[5] L.Hunter and K.B.Cohen, "Biomedical Language Processing: What's Beyond Pubmed?" Molecular Cell, Vol. 21-5 Pp. 589-594, 2006.
[6] Jeff Pasternack, Don Roth "Extracting Article Text from Webb with Maximum Subsequence Segmentation", bb, WWW 2009 MADRID.
[7] Abdur Rehman, Haroon.A.Babri, Mehreen saeed," Feature Extraction Algorithm For Classification Of Text Document", ICCIT 2012 .
[8] Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl.," Efficient Text Extraction Algorithm Using Color Clustering For Language Translation In Mobile Phone" , May 2012.
[9] In Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts"May2011
[10] 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,
[11] 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.
[12] National Center for Biotechnology Information. Entrez ProgrammingUtilities Help, 2010.
[13] 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.