Building a Classifier for Identifying Sentences Pertaining to Disease-Drug Relationships in Tardive Dyskinesia

Xia Bi[1], Hongzhan Huang[1,3], Sherri Matis-Mitchell[2], Peter McGarvey[3], Manabu Torii[1], Hagit Shatkay[1], Cathy Wu[1,3]
[1]Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA
[2]AstraZeneca Pharmaceuticals, Wilmington, DE, USA
[3]Protein Information Resource, Georgetown University, Washington, D.C.
bix@dbi.udel.com, huang@dbi.udel.edu, sherri.matis@astrazeneca.com, pbm9@georgetown.edu, torii@cis.udel.edu,
shatkay@cis.udel.edu, wuc@dbi.udel.edu

Abstract—**In this paper, we attempt to build a pipeline that identifies and extracts disease-drug relationships via sentence classification, and demonstrate the feasibility and utility of our approach using tardive dyskinesia as a case study. We manually developed and annotated a biomedical training corpus for tardive dyskinesia. Using 10-fold cross validation, we tested and trained a naïve Bayes classifier to identify sentences pertaining to disease-drug relationships. Our precision, recall, and F-measure were all approximately 66%, and area under the ROC curve was over 80%. Our method helps to elucidate various drug effects on tardive dyskinesia and constitutes an initial effort toward the task of disease-drug relationship extraction.**

Keywords-Tardive dyskinesia; Biomedical text mining; Naïve Bayes model; Relationship extraction

I. INTRODUCTION

Advances in computational and biological methods have greatly accelerated the pace of scientific discovery and produced a tremendous amount of experimental and computational data in the biomedical domain. Given the wealth of information that is available both in scientific papers and electronic databases, one particular challenge in biomedicine is to detect disease-drug associations and to organize them in a meaningful way that will accelerate pharmacogenetic research [1]. The main motivation for this paper is to devise a method that assists researchers to quickly identify disease-drug relationships from the biomedical literature and to classify those relationships into specific categories to enable better understanding of various drug effects. Specifically, we use tardive dyskinesia (TD) as a case study for our approach.

TD is a serious, irreversible neurological disorder characterized by repetitive, involuntary, and purposeless movements of various body parts. Although the prevalence rates are difficult to estimate and have reportedly differed across studies, a meta-analysis including 39,187 subjects with antipsychotic disorders from 76 studies found an overall prevalence of 24.2% [2]. Current research suggests that TD may result primarily from neuroleptic-induced D2 receptor hypersensitivity in the nigrostriatal pathway [3]. People affected by TD exhibit signs of abnormal movements and are subjected to humiliation and embarrassment, which lead to social stigma and inability to lead a normal lifestyle. This paper uses TD as a case study to build a model that seeks to better understand TD-related drugs and other symptomatic observations in association with TD.

In order to develop such a model and a classification system, we first set out to manually assemble and annotate a biomedical training corpus concerning TD via sentence classification. To identify relevant sentences related to drug effects in TD, we employed a naïve Bayes modeling classifier using the WEKA implementation (Waikato Environment for Knowledge Analysis) [4]. To assess the system, we employed the 10-fold cross-validation method to evaluate using precision, recall, F-measure, and area under the ROC curve. Our weighted average for precision, recall, and F-measure were all approximately 66%, and area under the ROC curve was over 80%.

We organize the rest of the paper as follows: First, we describe current tools to mine disease-drug associations from the biomedical literature. Then, we explain our method to annotate the biomedical training corpus for TD. We continue with the development and evaluation of our classification model. Finally, we present a discussion of the results and future work.

II. BACKGROUND AND RELATED WORK

Given the vast bodies of phenotypic and pharmaceutical data that are available both in scientific papers and electronic databases, researchers now face the challenge of integrating this data to detect disease-drug associations and construct meaningful scientific queries to support knowledge discovery. Several text mining tools have been developed to facilitate this purpose. MedMiner [5] is a keyword-based system that requires the user or programmer to supply the drug and gene names. EDGAR [6], which stands for Extraction of Drugs, Genes and Relations, is a natural language processing system that extracts information about drugs and genes relevant to cancer from the biomedical literature. The system is still under development and its performance has not been quantitatively assessed [6].

Textpresso [7] supports full text literature search over categories of terms pertaining to several model organisms including C. elegans and Mouse. Some other biomedical text mining systems include MedGene [8], LitMiner [9], iHOP [10], and ALIBABA [11]. However, these text mining tools were designed only to identify and extract relevant terms without further analysis on the specific relationships between

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

biological entities and facts. As such, researchers using these systems in an attempt to identify adverse drug effects in a specific disease may obtain much data that may contain many false positives. For example, a search for TD and its associated drugs using Textpresso for mouse yielded 859 matches in 115 documents. Given the large number of matches returned, it would be a very time-consuming task for a researcher to analyze the type of relationships that exist between the objects identified and to understand specific drug effects for this particular disease.

III. METHODS

A. Overview of the Pipeline

Fig. 1 shows an overview of our pipeline for document retrieval and sentence classification. It combines publicly available open-source components such as Genia Sentence Splitter [12] and Weka [4] with Perl scripts for data processing that we have written for this purpose. TD-related abstracts are retrieved from the PubMed database, fed into the Genia Sentence Splitter, tagged for drug name mentions, then manually categorized. Next, the text is tokenized into individual words and passed to Weka to build a sentence classifier. The classifier is compared against manual annotation and evaluated.

B. Document Retrieval and Sentence Classification

We first retrieved a set of abstracts that are related to TD from PubMed. According to the Unified Medical Language System (UMLS) Metathesaurus [13], which is a large (more than 620,000 concepts) compilation of several controlled vocabularies in the biomedical domain, TD has several textual variants that should be used when retrieving relevant biomedical text. A search using *tardive dyskinesia* and its related synonyms in either the title or abstract was performed. A total of 2783 PubMed abstracts were retrieved, of which 1734 were published between 1/1/1990-12/12/2011. We omit abstracts published earlier than 1/1/1990 because they do not contain the most up-to-date information about the disease in which we are interested. Fig. 2 shows the number of PubMed articles pertaining to TD, as a function of publication year for all articles and those that have an actual abstract stored.

The abstracts were passed to the GENIA Sentence Splitter [12], whose performance is reported to be an F-score of 99.7% [12]. A total of 16468 sentences were correctly obtained from 1734 PubMed abstracts, giving an average of 9 sentences per abstract, with a maximum of 38 sentences and a minimum of 2 sentences as shown in Fig. 3.

The sentences were then passed to a Perl script that looks for specific drug mentions. The drug ontology that we start with has 1494 drug names and synonyms from DrugBank's list of FDA-approved drugs. An additional 337 small molecules and 1138 drug classes from PharmGKB [14] were subsequently added. The resulting drug ontology was then manually curated by the first author, altogether consisting of 2968 drugs, small molecules, and drug classes. This drug ontology may be used to mine drug name and class mentions in relation to other diseases in the future. By including drug classes, our system is able to correctly identify 12.90% more TD-specific sentences compared to only having drug names.

Out of a total of 16468 sentences, 3993 (24.25%) sentences were found to contain one or more drug names. Those were parsed from 1734 PubMed abstracts, which gave an average of 2-3 drug-related sentences per abstract. We used a random set of 607 drug-containing sentences, and examined the number of drug mentions per sentence. When there are multiple drug name mentions in a sentence, we use the first drug because this is typically the focus of the sentence, except when a connector such as *and* or *or* was used. This was found to be true 95.72% of the time (581 sentences). Following this pattern, we associated the disease with the first drug mentioned and found 574 (94.56%) sentences to contain one drug name, and 33 (5.44%) sentences to contain two or more drug names.

Extensive manual classification of the 607 sentences was carried out by three annotators to ensure consistency. Two of these annotators are authors of this paper (which may constitute a limitation in classification results); one is independent from the paper and is blinded to the design and development of our system. All three annotators are experienced biologists holding at least an MSc degree, are familiar with pharmacology and have several years of professional annotation experience.

The sentences were classified into one of three categories: sentences that demonstrate benefit of a drug in relation to a disease were assigned to the Positive category, i.e. the drug is used to treat the disease. Sentences that involve negative effects between a drug and a disease were assigned to the Negative category, i.e. the drug induces the disease or is associated with progression of the disease. Sentences that belong to neither the positive nor the negative effect category were assigned to the Neither category. This occurs when the drug has no relation to the biological disease or when the sentence is inconclusive or exploratory in nature.

The agreement rate between at least two of the annotators

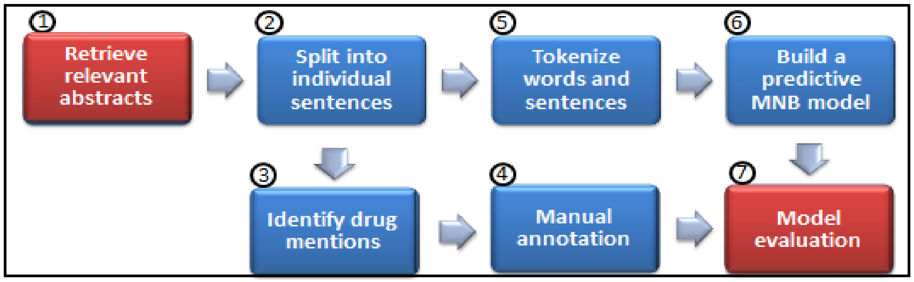
1
Retrieve relevant abstracts
2
Split into individual sentences
5
Tokenize words and sentences
6
Build a predictive MNB model
3
Identify drug mentions
4
Manual annotation
7
Model evaluation

Figure 1. Document retrieval and sentence classification pipeline

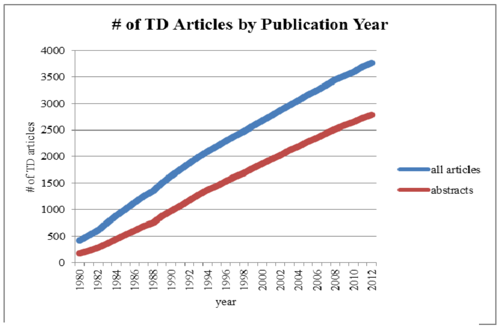

Figure 2. Number of TD articles by publication year

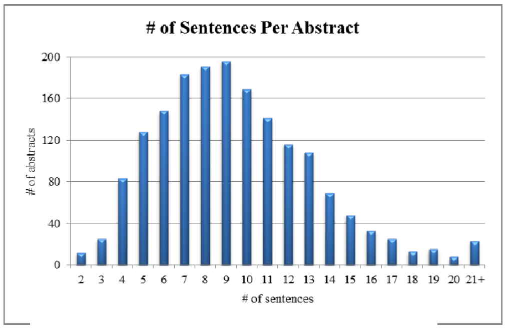

Figure 3. Number of sentences per abstract

was 88.6%, and among all three annotators was 81.25%. The disagreement was primarily between the Negative and Neither classes. In case of disagreement, categories were assigned by majority votes among all three annotators. We excluded sentences that had three-way disagreement. The number of annotated sentences that belong to each category is shown in Table I. For training and testing a classifier, we use the 604 sentences for which the classification was determined into one of the three main classes.

C. *Building a Naïve Bayes Classifier*

Following manual annotation, the next step entails generating a classifier using the annotated dataset for training and testing. For preprocessing the data and training/testing the naïve Bayes classifier, we use the Weka [4] tools, which have been used in a variety of other applications. We use the naïve Bayes classifier as it is simple and computationally efficient, and has relatively good predictive performance [15, 16].

The procedure carried out to train our naïve Bayes classifier is the following: We used programs in Weka to break sentences into individual words based on blank spaces and punctuation marks, and eliminate words whose term frequencies were fewer than 3 times. We experimented with several thresholds on term frequency, and 3 was established to be the best cutoff point to build an effective classifier. Our baseline measurement against which we compare our results obtained using the naïve Bayes classifier for the Positive, Negative, and Neither classes were 0.32, 0.27, and 0.41, respectively.

TABLE I. MANUAL CLASSIFICATION RESULTS

# articles	607
Positive (1)	191
Negative (2)	161
Neither (3)	252
Excluded	3

Some examples of words along with their corresponding conditional probabilities as calculated by training are shown in Table II. As expected, words associated with positive outcomes such as *therapeutic* and *improvement* have a higher probability to occur in the Positive class; whereas words with negative outcomes such as *vacuous* (as in *vacuous chewing movement*) and *neurotoxic* have a higher probability to occur in the Negative class. It is important to note that differences in word probability across the classes are quite small, so that the words are not necessarily informative, which might be due to the small dataset. Further research in feature selection is needed to aid the explicit choice of distinguishing terms as was done in other area [17].

IV. RESULTS AND DISCUSSION

Cross validation is commonly used to evaluate performance of predictive modeling techniques such as naïve Bayes [18]. We applied 10-fold cross-validation to train and test the classifier, where each observation is used for testing exactly once. Results from the 10-fold cross-validation were measured in terms of precision, recall, F-measure, and area under the ROC curve, as shown in Table III. Precision, recall, and F-measure were calculated according to conventional definitions.

The classifier achieved a fairly good precision, recall, and area under the ROC curve in classifying sentences retrieved from abstracts associated with TD. We were able to obtain a precision, recall, and area under the ROC curve of approximately 66%, 66%, and 83%, respectively. As a baseline of comparison, if we were to assign all sentences to the majority class – the *Neither* class, we would obtain precision and recall of only 41.68%.

Some factors account for misclassification. Sentences containing multiple drug names may be associated with both positive and negative words. Analyzing the output of the classifier, we noted that 28% of the sentences (20 sentences) that were misclassified contained an ambiguous statement of the form: “Drug A is used to treat Disease B, but causes C as a side effect.” To address this issue of contrasting biological observations that contributed to classification error, we looked into the possibility of using syntactic and semantic processing to identify multiple event descriptions in the sentences by passing the output data to MetaMap [19] (results not shown).

Misclassification also occurred in sentences that have multiple drug mentions, where we associated an effect with the drug mentioned first, as that is typically the focus of the sentence. However, exception to this rule occurs when

TABLE II. SOME EXAMPLES OF WORD PROBABILITY

The probability of a word given the class			
	Positive	***Negative***	***Neither***
therapeutic	0.0022	3.228E-4	9.164E-4
improvement	0.0055	6.456E-4	4.582E-4
vacuous	5.490E-4	0.0036	2.291E-4
neurotoxic	2.745E-4	0.0013	2.291E-4

TABLE III DETAILED ACCURACY BY CLASS

Class	Precision	Recall	F-measure	ROC area
1	0.645	0.686	0.665	0.827
2	0.627	0.615	0.621	0.837
3	0.691	0.667	0.679	0.824
Weighted Avg.	0.66	0.659	0.659	0.828

biological effects discussed do not pertain to the first drug in the sentence. Better method may be developed to accurately analyze drug mentions and reported effects.

This work can be further extended to other biological diseases and can be used to mine relationships other than those between diseases and drugs. For instance, gene name mentions may be identified and associated with drug mentions to examine the role of genetic variants in individual drug response [20, 21]. Biological processes or pathways may also be associated with certain genes or proteins to understand the molecular mechanisms that underlie a disease [22, 23].

V. CONCLUSION

We have manually developed and annotated a large biomedical training corpus for tardive dyskinesia by manually classifying sentences into one of three classes. We used the annotated data to train and test a naïve Bayes classifier, employing 10-fold cross-validation. Our precision, recall, and F-measure were 66%, and area under the ROC curve was over 80%. We also looked into the possibility of using syntactic and semantic processing to identify multiple event descriptions in the sentences by passing the output data to MetaMap.

The work includes several components that are not found in many of the text mining systems that extract relationships between diseases and drugs. These include: (1) a comprehensive drug ontology that consists of 2968 drugs, small molecules, and drug classes; (2) biomedical training corpus on tardive dyskinesia which has been consistently and extensively annotated for classification purposes; and (3) identification of distinct biological observations for disease-drug relationships found in biomedical text using software tools that are open-source and readily available.

ACKNOWLEDGMENT

This work was supported in part by NIH/NLM 5G08LM010720-02 and NIH/NIGMS 8P20GM103446-12 from the National Institutes of Health.

REFERENCES

[1] Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics. 2010 Oct;11(10):1467-89.

[2] van Harten PN, Tenback DE. Tardive dyskinesia: clinical presentation and treatment. Int Rev Neurobiol. 2011;98:187-210.

[3] Hoerger, Michael. The primacy of neuroleptic-induced D2 receptor hypersensitivity in tardive dyskinesia. 2007. Psychiatry Online (Psychiatry Online) vol.13 (no.12): 18–26.

[4] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update. 2009. SIGKDD Explorations, Volume 11, Issue 1.

[5] Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques. 1999;27:1210–1217.

[6] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput. 2000:517–528.

[7] Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2004 Nov;2(11):e309. Epub 2004 Sep 21.

[8] Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J. Analysis of genomic and proteomic data using advanced literature mining. J. Proteome Res. 2003;2:405–412.

[9] Maier H, Dohr S, Grote K, O'Keeffe S, Werner T, Hrabe de Angelis M, Schneider R. LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts.Nucleic Acids Res. 2005;33(Webserver issue):W779–W782.

[10] Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature.Bioinformatics. 2005;21(Suppl. 2):ii252–ii258.

[11] Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. Alibaba: PubMed as a graph.Bioinformatics. 2006;22:2444–2445.

[12] Kim J.D., Ohta T., Tateishi Y., and Tsujii J., GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl. 1):180–i182, 2003.

[13] Humphreys, B.L., D.A.B.Lindberg, H.M.Schoolman, and G.O.Barnett. The Unified Medical Language System: An informatics research collaboration. Journal of the American Medical Informatics Association. 1998. 5(1): 1-13.

[14] E.M. McDonagh, M. Whirl-Carrillo, Y. Garten, R.B. Altman and T.E. Klein, "From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource."Biomarkers in Medicine (2011) Dec; 5(6):795-806.

[15] Cheng BY, Carbonell JG, Klein-Seetharaman J. Protein classification based on text document classification techniques. Proteins. 2005 Mar 1;58(4):955-70.

[16] Eibe Frank, Remco R. Bouckaert. Naive bayes for text classification with unbalanced classes. Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases. 2006. Springer-Verlag Berlin, Heidelberg.

[17] Brady S, Shatkay H. EpiLoc: a (working) text-based system for predicting protein subcellular location. Pac Symp Biocomput. 2008: 604-15.

[18] Mitchell, Tom. Machine Leaning. McGraw Hill, 1997. ISBN 0-07-042807-7.

[19] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

[20] Limdi NA, Veenstra DL. Expectations, validity, and reality in pharmacogenetics. J Clin Epidemiol. 2010 Sep;63(9):960-9.

[21] Kalow W. Human pharmacogenomics: the development of a science. Hum Genomics. 2004 Aug;1(5):375-80.

[22] van der Helm-van Mil AH, Wesoly JZ, Huizinga TW. Understanding the genetic contribution to rheumatoid arthritis. Curr Opin Rheumatol. 2005 May;17(3):299-304.

[23] Herwig R, Lehrach H. Expression profiling of drug response--from genes to pathways. Dialogues Clin Neurosci. 2006;8(3):283-93.