# BUILDING A PREDICTIVE MODELING SYSTEM FOR SENTENCE CLASSIFICATION: A CASE STUDY USING TARDIVE DYSKINESIA

by

Xia Bi

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics and Computational Biology

Spring 2012

**BUILDING A PREDICTIVE MODELING SYSTEM FOR SENTENCE CLASSIFICATION: A CASE STUDY USING TARDIVE DYSKINESIA**

by

Xia Bi

Approved: _____
Cathy H. Wu, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____
Errol Lloyd, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Interim Dean, College of Engineering

Approved: _____
Charles G. Riordan, Ph.D.
Vice Provost for Graduate and Professional Education

## ACKNOWLEDGMENTS

I wish to thank my adviser, Cathy; and my committee members Peter, Hongzhan, Sherri, and Hagit for their continuous advice, guidance, and academic support during the past two years. I must also thank my professional friends and colleagues, who have supported and helped me throughout my graduate education.

This manuscript is dedicated to my parents, Ming Bi, Thomas Bonnes and Jiuhong Bonnes for their unconditional love.

.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ABSTRACT

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 are 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. Several text mining tools have been developed to facilitate this purpose. They perform adequately well in identifying facts and entities using on-the-fly search of scientific articles from many different databases; however, they cannot analyze the type of relationship that exist between the objects identified. In this thesis, we propose a novel method to analyze drug-disease relationships using a combination of in-house and open-source tools that exploit the Multinomial Naïve Bayes (MNB) modeling technique.

The main motivation behind this thesis work is to assist researchers to quickly identify disease-drug relationships from the biomedical literature using the case study of tardive dyskinesia (TD) and to classify those relationships into specific categories to enable better understanding of various drug effects. We have manually developed and annotated a biomedical training corpus for TD via sentence classification. Using the MNB modeling technique, we generated a learning model and built a predictive

classifier system using data preprocessing and filtering algorithms. To assess whether the model would generalize to an independent dataset, we applied the 10-fold cross-validation method to evaluate the model using precision, recall, F-measure, and ROC area. The precision, recall, and F-measure were approximately 88%, and ROC area was over 97%.

One particular challenge in sentence classification is the co-existence of contrasting biological observations that cause confusion to the classification model. To address this ambiguity issue, we passed the output data to Metamap to identify and separate distinct biological observations in biomedical text. By further discerning the semantic meaning of biological observations, we classified biomedical sentences into more refined categories, which helped to elucidate various drug effects and proved to be an initial effort toward the sophisticated task of disease-drug relationship extraction.

# Chapter 1

# 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 are 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. The main motivation behind this thesis work is to assist researchers to quickly identify disease-drug relationships from the biomedical literature using the case study of tardive dyskinesia (TD) and to classify those relationships into specific categories to enable better understanding of various drug effects.

TD is a serious, irreversible neurological disorder characterized by repetitive, involuntary, purposeless movements of various body parts. The most typical sign of TD is orofacial dyskinesia (i.e. chewing movements and tongue protrusions), but the body trunk and extremities may also be affected [1]. It is frequently associated with long-term or high-dose use of dopaminergic antagonists, usually antipsychotic medications such as haloperidol. Although the prevalence rates are difficult to estimate and have reportedly differed between studies, a meta-analysis including

39,187 subjects from 76 studies found an overall prevalence of 24.2% [2]. The underlying neurological mechanisms of TD are not yet completely understood. 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 work 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 this thesis work, we first set out to manually develop and annotate a biomedical training corpus for TD via sentence classification. To extract meaning from terms and sentences, we employed supervised machine learning and word context methods to generate a learning model and build a predictive classifier system. This was accomplished with the Multinomial Naïve Bayes modeling technique using data preprocessing and filtering algorithms in WEKA (Waikato Environment for Knowledge Analysis) [4]. To assess whether the model would generalize to an independent dataset, we used the 10-fold cross-validation method to evaluate the model using precision, recall, F-measure, and ROC area. Our precision, recall, and F-measure were approximately 88%, and ROC area was over 97%.

Finally, we looked into more sophisticated semantic processing to handle complex event descriptions in the sentences. One particular challenge encountered previously during sentence classification was the co-existence of contrasting biological observations that caused confusion to the classification model. To address

this ambiguity issue, we passed the output data to Metamap to identify and separate distinct biological observations in biomedical text. By further discerning the semantic meaning of biological observations, we classified biomedical sentences into more refined categories, which helped to elucidate various drug effects and proved to be an initial effort toward the sophisticated task of disease-drug relationship extraction. Our model may be extended to other biological diseases and can be used to mine relationships in aspects other than 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 [24] [25]. A biological process or pathway may also be associated with certain genes or proteins to understand the molecular mechanisms that underlie a disease [26] [27].

We organized 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.

## Chapter 2

## 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 to apply translational bioinformatics to integrate 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. PolySearch [5] is a web-based text mining system that identifies tagged terms and their relationships, then organizes this information in a formatted and structured database. The tagged terms include human diseases, genes, mutations, drugs and metabolites. Its strength lies in on-the-fly search of scientific articles from many different databases including DrugBank [6], SwissProt [7], HGMD [8], OMIM [9], etc. However, as a text mining tool, PolySearch uses a relatively simple dictionary approach to identify biological or biomedical associations, which means PolySearch cannot identify novel or newly named diseases, genes, cell types, drugs or metabolites [5]. Another limitation is that PolySearch does not utilize artificial intelligence (AI), word context or machine learning (ML) methods in its current term identification term, hence it is unable to extract context or meaning from tagged terms or sentences [5].

MedMiner [10] is a keyword-based system that requires the user or programmer to supply the drug and gene names. EDGAR [11], 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. It uses a part-of-speech tagger and is able to generate relational assertions with correct arguments from syntactically complex sentences, but the system is still in development and its performance has not been quantified. Its accuracy is best characterized as moderate and it cannot analyze the type of relationship that exist between the objects identified [11].

Textpresso [12] supports full text literature searches of categories of terms pertaining to several model organisms including Caenorhabditis elegans and Mouse. Adapted from Textpresso, Pharmspresso [13] uses a dictionary-based approach to find references to human genes, polymorphisms, drugs, diseases, and their relationships from full text articles. It allows the user to query for a specific pattern such as "{drug} {association} {gene}" within a given sentence. For example, specific instantiations of the gene"BRCA2" can be sought by querying for the keyword "BRCA2" with the categories {drug} and {association}. Because Pharmspresso explores full text articles which contain richer information than literature abstracts, it requires availability of full text articles and only work on pre-defined corpus of relevant literature. Since this work is so labor-intensive, only 1025 articles have been included in Pharmspresso thus far [13].

Some other biomedical text mining systems include MedGene [14], LitMiner [15], iHOP [16], ALIBABA [17] and EBIMed [18]. However, these text mining tools were designed to only identify and extract relevant terms without further analysis on the specific relationships between biological entities and facts. Researchers will be bombarded with all the data presented that virtually contain many false positives. For example, a search for TD and its associated drugs using Textpresso for mouse yielded 859 matches in 115 documents is shown in Figure 2.1. 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.

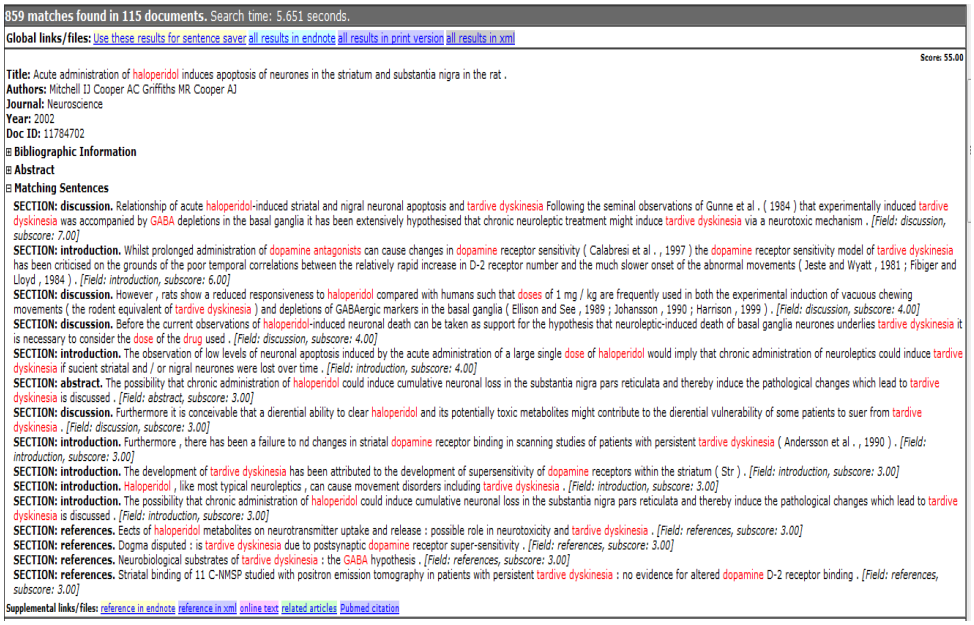


Figure 2.1 Results of Textpresso Mouse. Snapshot of Textpresso Mouse results page. User is searching for tardive dyskinesia and a member of the {drug} category.

# Chapter 3

# METHODS

## 3.1 Overview of the Pipeline

Figure 3.1 shows an overview of the pipeline for document retrieval and sentence classification. We combined publicly available open-source tools such as Genia Sentence Splitter [19] and Weka with data processing Perl scripts we had written for this purpose.

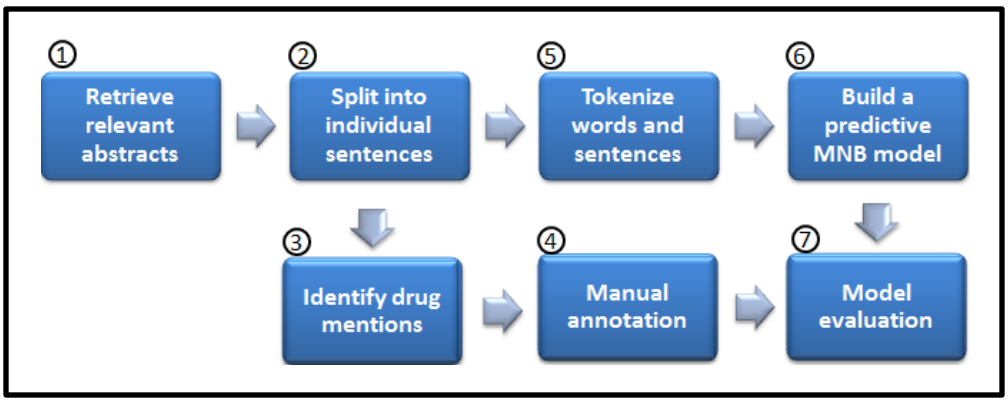


Figure 3.1 Document retrieval and sentence classification pipeline overview. TD-related abstracts are retrieved from PubMed, fed into the Genia Sentence Splitter, tagged for drug name mentions, then manually categorized. Next, the text is tokenized into individual words and passed to Weka to build a predictive modeling system for sentence classification. The model is compared against manual annotation and evaluated using ROC measures.

## 3.2 Retrieving Relevant PubMed Abstracts

We first set out to retrieve a set of abstracts that are related to TD from PubMed. According to the Unified Medical Language System ® (UMLS®) Metathesaurus ® [20], 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 taken into consideration when retrieving relevant biomedical text. A search using the following keywords in either the title or abstract was performed: "tardive dyskinesia; dyskinesia tardive; drug-induced tardive dyskinesia; oral-facial dyskinesia; tardive dystonia; tardive oral dyskinesia." There are a total of 2783 PubMed abstracts, of which 1734 are published in the time frame of 1/1/1990-12/12/2011. We decide to omit abstracts prior to 1/1/1990 because the format and organization of those are vastly different from abstracts published in recent decades and do not contain the most up-to-date information in the disease we are interested in. Figure 3.2 shows the number of PubMed articles by publication year for all articles and those that contain an abstract.

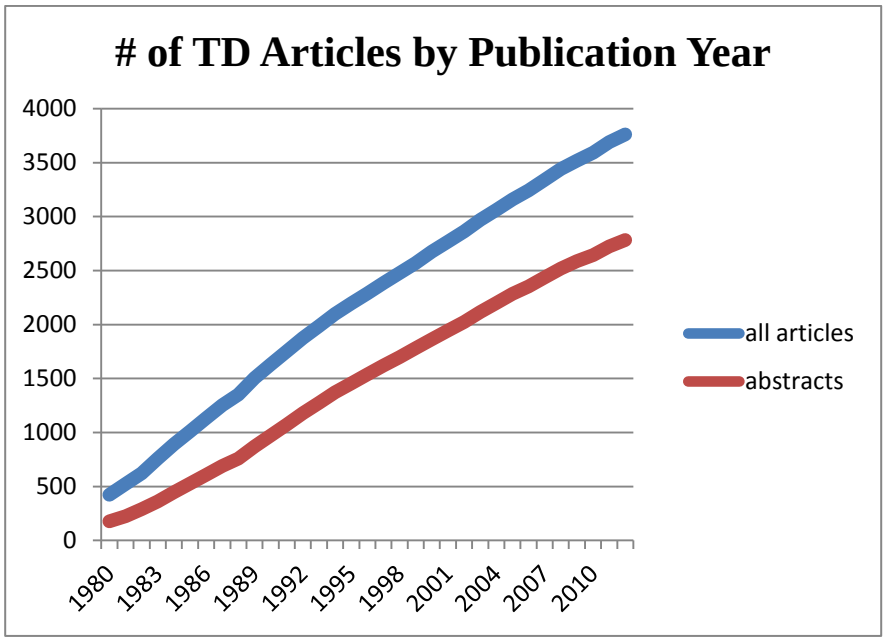


Figure 3.2 Number of TD articles by publication year. Plots the number of TD-related articles by publication year for all articles and those that contain an abstract. Abstracts published before 1/1/1990 have been omitted.

### 3.3 Splitting Sentences

The abstracts were passed to the Genia Sentence Splitter [19], which has been optimized for biomedical texts. The splitter employs a classification model based on supervised learning method using maximum entropy modeling, and has obtained an F-score of 99.7 on 200 unseen GENIA abstracts [19]. A total number of 16468 sentences were correctly split from 1734 PubMed abstracts, giving an average of 9-10 sentences per abstract, with maximum 38 sentences and minimum 2 sentences as shown in Figure 3.3.

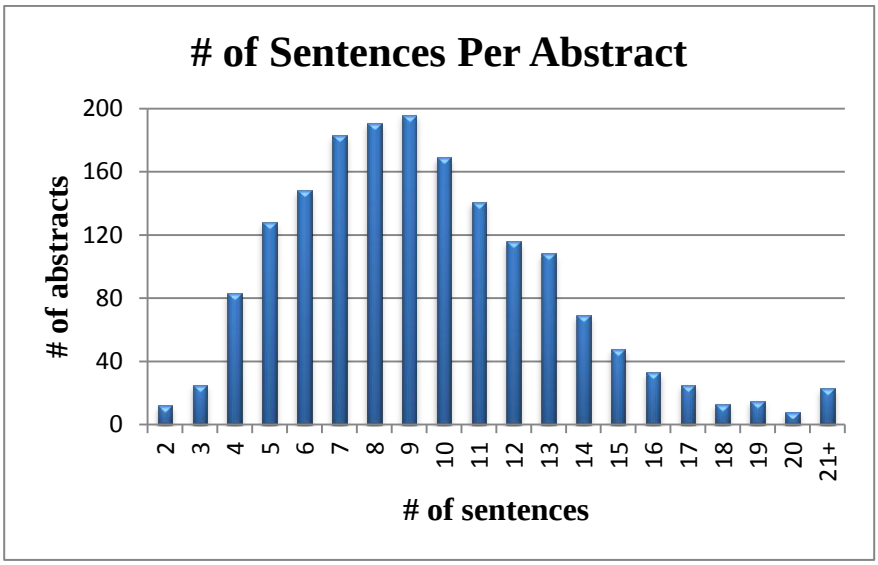


Figure 3.3 Number of sentences per abstract. Plots the number of sentences per abstract and the corresponding number of abstracts that contain those sentences. Total = 16468 sentences, 1734 abstracts, average = 9.497 sentences/abstract, max = 38, min = 2.

### 3.4 Identifying Drug Mentions

The sentences were then passed to a Perl script that looks for specific drug mentions. The drug ontology consists of 1494 drug names and synonyms from DrugBank's list of FDA-approved drugs. An additional 337 small molecules and 1138 drug classes from PharmGKB [21] were subsequently added. The resulting drug ontology was then manually curated by the primary author, Xia Bi, altogether leading to 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. Current text mining systems only look for specific drugs and omit those that mention category

of drugs. Our system ensures retrieval of relevant sentences that mention only the parent drug class, i.e. first-generation antipsychotic drugs. By including drug categories, our system is able to correctly identify 12.90% more TD-specific sentences compared to only having drug names.

Out of 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. The number of drug mentions per sentence was also examined. Upon manual curation, it was observed that the first drug mentioned in the sentence is almost always the focus of the sentence, and whenever two or more drugs are the subject of a sentence, a connector word such as "and" or "or" is used to link multiple drugs. For example,

> "Olanzapine has demonstrated efficacy in maintenance treatment as well as a reduced risk of tardive dyskinesia compared with haloperidol." (PMID:9847048)

Both olanzapine and haloperidol are mentioned in this sentence, but olanzapine is the drug of interest, so a positive relationship between olanzapine and TD is established. Following this pattern, we eliminated drug mentions that are inconsequential, and obtained the number of drug mentions per sentence as shown in Table 3.1 and Figure 3.4. Since approximately 95% of all sentences focus on one specific drug, our model is applicable for the majority of sentences parsed from biomedical text. It is important to note, however, that the small percentage of sentences that contain multiple drug mentions have practical implications in sentence classification. Because observations can refer to different drugs, there may be ambiguity and misclassification of sentences

Table 3.1 Number and percentage of drug name mentions per sentence.

| # of drug names/sentence | # of sentences | % of all sentences |
|---|---|---|
| 1 | 574 | 94.56% |
| 2 | 31 | 5.11% |
| 3 | 1 | 0.16% |
| 4 | 1 | 0.16% |

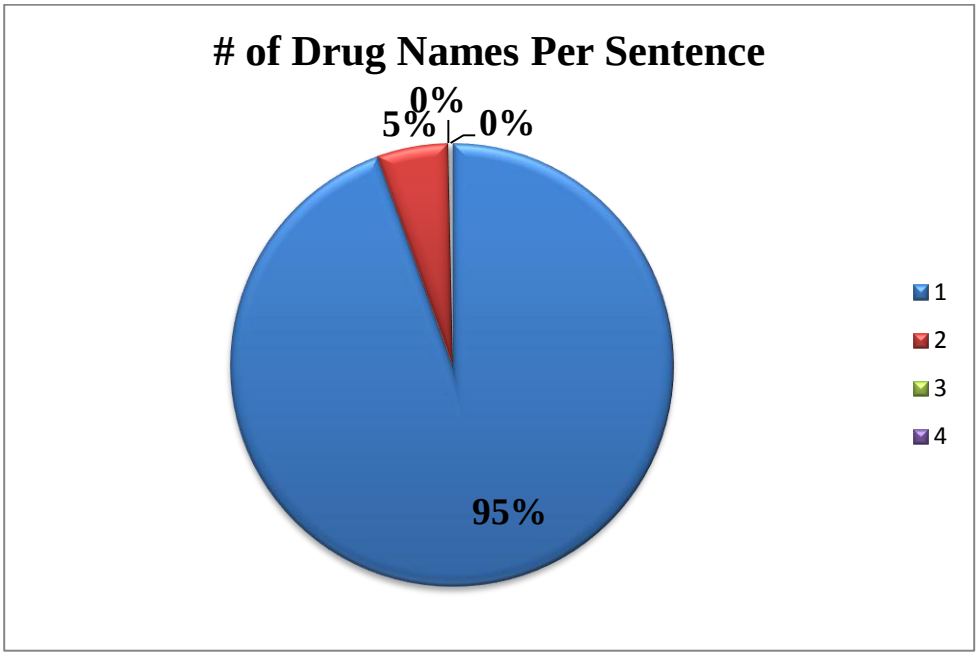


Figure 3.4 Percentage of drug name mentions per sentence.

in constructing the sentence classifier. Hence, we looked into more sophisticated semantic processing to handle complex event descriptions in the sentences. An attempt to separate distinct observations and to increase the accuracy of our model was carried out using Metamap [22], which proves to be an initial effort toward the sophisticated task of disease-drug relationship extraction.

### 3.5 Annotating Sentences

Extensive manual classification of 607 drug-containing sentences was carried out by three annotators to ensure consistency. All three annotaters are experienced biologists who are familiar with pharmacology. The agreement rate between the three annotaters was 81.25%. The sentences were classified into one of three categories: sentences that contain a positive relationship between the drug and 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 or groups of drugs and a disease were assigned to the negative category, i.e. the drug induces the disease or is associated with progression of the disease. It is crucial to take context into consideration while categorizing sentences. Some sentences indicate that a drug has a less severe risk of inducing the disease compared to other neuroleptics. For example,

> "However, if cases do develop, the risk of tardive dyskinesia is likely to be less with clozapine than with typical neuroleptics." (PMID:8104929)

Since we wish to capture the superiority of clozapine, we have assigned this sentence to the positive category even though clozapine may induce the disease.

Sentences that belong to neither the positive or negative effect category were assigned to the third category. This occurs when the drug has no relation to the biological disease or when the sentence is inconclusive or exploratory in nature. An example of a sentence for each respective category is shown in Table 3.2.

Out of 607 drug-containing annotated sentences, 191 were classified to the first category, 161 were classified to the second category, and 252 sentences were

Table 3.2 Example for each of the three categories in manual sentence classification.

| **Pub Med ID** | **Title** | **Sentence** | **Drug** | **Category** |
|---|---|---|---|---|
| 9466234 | Risperidone in children and adolescents with pervasive developmental disorder: pilot trial and follow-up. | Overall, 5 of the 6 patients derived significant clinical benefits from risperidone. | Risperidone | 1 |
| 21112461 | Duloxetine-related tardive dystonia and tardive dyskinesia: a case report. | Even though this association has been rarely reported, duloxetine may pose a potential risk of inducing tardive syndrome. | Duloxetine | 2 |
| 10775299 | Risperidone implicated in the onset of tardive dyskinesia in a young woman. | She had received small dosages of typical antipsychotics before and during receiving risperidone for short periods. | Typical antipsychotics | 3 |

classified to the third category, the last 3 sentences were classified as unsure as shown in Table 3.3. For predictive modeling purposes, we omit the unsure sentences, and build a classification scheme using the remaining 604 sentences.

Table 3.3 Overview of manual classification results

| | |
|---|---|
| # articles | 607 |
| Positive (1) | 191 |
| Negative (2) | 161 |
| Neither (3) | 252 |
| Unsure | 3 |

The top ten most-mentioned drugs are shown in Table 3.4. Drugs that are frequently mentioned tend to be the most relevant or controversial with respect to the symptom or disease. Haloperidol, which is the most mentioned drug out of all 607 manually annotated sentences, is used to induce TD in the animal model. The drug classes “Atypical antipsychotics” and “Typical antipsychotics” are two other frequently mentioned drug terms from the input dataset. Certain drugs may belong to different classes of compounds depending on the date of publication and the author, but identification of drug classes serves as an additional piece of information that the user can choose to keep or ignore.

Table 3.4 Top ten most-mentioned drugs from 607 manually annotated sentences.

| **Drug** | **Count** |
|---|---|
| Haloperidol | 89 |
| Clozapine | 68 |
| Risperidone | 58 |
| Atypical antipsychotics | 47 |
| Olanzapine | 35 |
| Vitamin E | 29 |
| Metoclopramide | 28 |
| Typical antipsychotics | 20 |
| Aripiprazole | 17 |
| Reserpine | 13 |

### 3.5 Building a Multinomial Naïve Bayes Classifier

Following manual annotation, the next step entails generating a learning model and building a predictive classifier system using data preprocessing and classification techniques. The steps are summarized below:

1. Convert the sentences into datasets suitable for processing by Weka
2. Import the dataset and preprocess to select important features using filtering algorithms
3. Apply Multinomial Naïve Bayes (MNB) algorithm to train the model

The software requirement for building the predictive model is Weka [4], a free software originally developed at the University of Waikato, New Zealand. Weka contains a collection of visualization tools and machine learning algorithms for data analysis and predictive modeling, and provides a simple graphical user interface for ease of use. The current version is written in the Java programming language, which can be run on almost any computing platform. It supports several standard tasks to solve real-word data mining problems, such as data preprocessing, classification, regression, and feature selection [4]. Because of these advantages, Weka has been used in many different application areas to analyze large datasets for educational purposes and research.

Weka contains three user interfaces: Explorer, Knowledge Flow, and Experimenter. Each interface has several panels that feature different functionalities of the workbench. In the Explorer interface, the Preprocess panel facilitates data imported from a database or a text file, and pre-processes this data using several

filtering algorithms such as stemming and tokenizing. Stemming reduces inflected words (i.e. inflicted, inflicting) to their stem, base or root form (i.e. inflict). Tokenization breaks up a stream of text into words, phrases, symbols, or other meaningful elements called tokens, which are important for assigning a probability and are then used as input for further processing steps in classification. The Classify panel in the Explorer interface applies classification and regression algorithms to the input dataset, builds a predictive model that best approximates the input data, and assesses the accuracy of the model using various measures. The Associate panel enables the user to identify important interrelationships between attributes in the data. Others include the Cluster, Select, and the Visualize panel, which provide access to different clustering techniques and visualization tools in Weka.

The MNB modeling technique was implemented to build a statistical model for sentence classification. This technique is computationally efficient and has relatively good predictive performance [28] [29]. The MNB model views each sentence as a collection of words and each word is independent from each other given the class variable. The probability that the $n$-th word of a given document occurs in a class value $C$ can be represented by $p\ (w_n|C)$.

The procedures carried out to build a MNB model is the following: Perl script is used to convert the sentences into a collection of datasets suitable for processing by Weka. The MNB algorithm in Weka opens each dataset, and breaks sentences into individual features (words) based on blank spaces and punctuation using the built-in tokenizer. Data filtering was applied to eliminate features that occur less than 3 times.

The algorithm then counts the frequency of each feature and creates a string-to-word vector. In this vector, the rows correspond to sentences and columns correspond to each feature. Thus, each element in the vector is typically the number of occurrences of the feature in a given sentence. For example, if the word *efficient* appears in a sentence for 3 times, then the feature vector of the word *efficient* in that particular sentence is 3. This arrangement is used to represent the sentence class by counting the frequency of semantically significant features. This is the process by which Weka preprocesses the input data to select important features using filtering algorithms. Altogether, 577 words (features) are filtered out as shown in Figure 3.5.

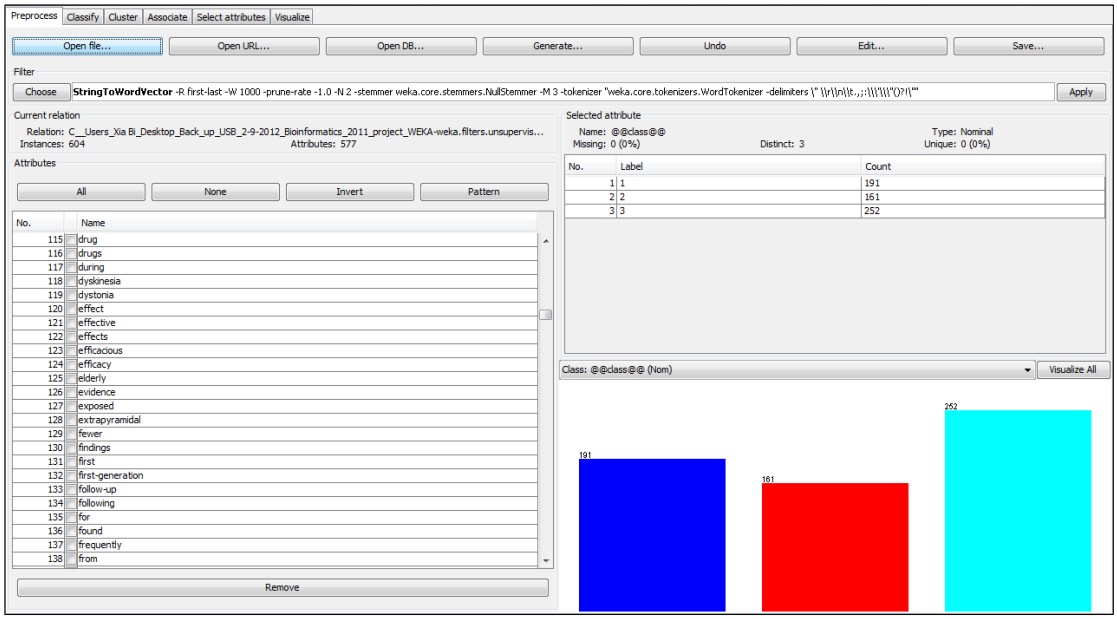


Figure 3.5 Snapshot of Weka Preprocess page. String-to-word vector was applied to the input dataset to select important features using filtering algorithms.

The MNB algorithm then calculates the probability that a given sentence *S* belongs to a given class *C* according to the following equation:

$$p(C|S) = \frac{p(S \cap C)}{p(S)} = \frac{p(C)}{p(S)} p(S|C)$$

where *p(S|C)* is the probability that a given sentence *S* contains all of the words $w_n$, given a class *C*. We apply the conditional independence assumption:

$$p(S|C) = p(w_1, w_2, \dots, w_n|C)$$

$$= p(w_1|C)p(w_2|C, w_1)p(w_3|C, w_1, w_2) \dots p(w_n|C, w_1, w_2, \dots, w_{n-1})$$

$$= \prod_{w \in S} p(w|C)^{n_{wS}}$$

and obtain the probability of class *C* given sentence *S* as:

$$p(C|S) = \frac{p(C)}{p(S)} \prod_{w \in S} p\,(w|C)^{n_{wS}}$$

for the probability model of the MNB classifier, where *p(C)* is the prior probability of class *C* and is estimated by the proportion of training documents pertaining to each class. This is also the baseline measurement (shown in Figure 3.6a) against which we would compare our results using the MNB classifier system. *p(S)* is a constant that makes the probability of different classes sum to 1, and $n_{wS}$ is the number of times the word *w* occurs in sentence *S*.

a

**The prior probability of a class *p(C)***

| | |
|---|---|
| 1 | 0.316310 |
| 2 | 0.266886 |
| 3 | 0.416804 |

b

**The probability of a word given the class**

| | 1 | 2 | 3 |
|---|---|---|---|
| useful | 5.50130E-4 | 1.509206E-4 | 1.16009E-4 |
| well-tolerated | 2.75065E-4 | 1.509206E-4 | 1.16009E-4 |
| therapeutic | 0.001100 | 1.509206E-4 | 4.64037E-4 |
| tetrabenazine | 9.62728E-4 | 3.018412E-4 | 3.48027E-4 |
| induced | 4.12598E-4 | 0.001207 | 3.48028E-4 |
| toxicity | 1.37532E-4 | 6.036824E-4 | 1.16009E-4 |
| vacuous | 2.75065E-4 | 0.001660 | 1.16009E-4 |

Figure 3.6 Baseline measurement and word probability. 3.6a shows the prior probability of each class *C* and 3.6b shows some examples words with their corresponding *p(w|C)*.

The predictive performance of the MNB model can be improved by appropriate data transformations. We applied data filtering to the input data to eliminate features that occur less than 3 times. This was found to have better precision, recall, and ROC area than including all features (2886). A tokenizing algorithm was employed to break sentences into individual words based on blank spaces and punctuation. Using the baseline measurement, we obtain an accuracy of 41.68% if we always pick the third class which has the highest probability. However, using the MNB classifier system, we are able to obtain a precision and recall of approximately

88%, which gives much higher confidence. Some word examples with their corresponding *p(w|C)* is shown in Figure 3.6b. As expected, words with positive outcomes such as “useful”, “well-tolerated”, and “tetrabenazine” have a higher probability for the first class; whereas words with negative outcomes such as “induced”, “toxicity”, and “vacuous” (as in “vacuous chewing movement”) have a higher probability for the second class.

Finally, we construct a MNB classifier from the probability model using maximum likelihood estimation, which estimates the parameters of our statistical model and selects the class that is the most probable given the model as defined below:

$$classify(S) = \arg\max_{c} p(C = c) \prod_{i=1} p\,(w = w_i | C = c)^{n_{wiS}}$$

# Chapter 4

# RESULTS

## 4.1 Model Evaluation

We would like to assess whether the MNB classifier will generalize to an independent dataset using the cross validation method, and evaluate the classifier using several measures including precision, recall, F-measure, and receiver operating characteristic (ROC) area.

Cross validation has been used to evaluate performance of predictive modeling techniques such as naïve bayes and support vector machine [30]. To evaluate performance of the MNB classifier, we applied the 10-fold cross-validation to assess whether the statistical model would generalize to an independent data set that has never been seen. The validation method randomly partitioned the sample into 10 complementary subsets. 9 subsets were used as training data for training the model and the remaining subset was retained as the validation data for testing the model. The cross-validation process was repeated for 10 times, with each of the 10 subsets used exactly nine times as the training data and once as the validation data. Results from all 10 cross validations were computed to produce a single estimation. This method used all observations for both training and validation, and each observation was used for validation exactly once, which reduced bias from random sub-sampling.

Results from the 10-fold cross-validation test were measured in terms of accuracy, recall, and ROC curve as shown in Table 4.1. Precision, recall, and F-measure were calculated according to the conventional definition as shown below, where TP stands for true positive, FP stands for false positive, and FN stands for false negative.

$$\text{Precision (P)} = TP / (TP+FP)$$
$$\text{Recall (R)} = TP / (TP + FN)$$
$$\text{F-measure}=(2 \times P \times R)/(P+R)$$

Table 4.1 Detailed accuracy by class.

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0.859 | 0.044 | 0.901 | 0.859 | 0.879 | 0.972 | 1 |
| | 0.857 | 0.05 | 0.863 | 0.857 | 0.86 | 0.973 | 2 |
| | 0.921 | 0.085 | 0.885 | 0.921 | 0.903 | 0.973 | 3 |
| **Weighted Avg.** | 0.884 | 0.063 | 0.884 | 0.884 | 0.884 | 0.973 | |

The ROC curve plots the true positive versus false positive rate for the MNB classifier system. The ROC area is the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. For three classes, ROC area of 33% is considered random guessing. The ROC area for all three classes was approximately 97%, much higher than random guessing.

We applied data filtering to the input data to eliminate features that occur only a few times to improve the measurement outcome. Several minimum term frequencies were attempted, and 3 was established to be the ideal cutoff point to significantly improve model accuracy, yet still retain sufficient features (577) to build a valid classification model.

Table 4.2 shows the confusion matrix, which tabulates the number of correctly and incorrectly classified sentences for all three classes. For all 191 sentences categorized to the first class by manual annotation, the MNB classifier system correctly classified 164 as a (class 1), incorrectly classified 10 as b (class 2), and 17 as c (class 3). The same principle applies to 161 sentences in the second and 252 sentences in the third class.

Table 4.2 Confusion Matrix.

| **=== Confusion Matrix ===** | | | |
|---|---|---|---|
| a | b | c | <-- classified as |
| 164 | 10 | 17 | a=1 |
| 10 | 138 | 13 | b=2 |
| 8 | 12 | 232 | c=3 |

## 4.2 Text Annotation Using Metamap

Due to the complex interplay of biological diseases and pharmaceutical substances, biomedical texts usually contain sentences that involve one or more

clinical observations that are difficult to classify into a single category. Sentences that contain two or more drug name mentions may associate each drug with a contrasting biological effect. Even sentences that focus on only one drug entity may exert different physiological effects that cause ambiguity to the classification model using a simple bag-of-words approach. This could account for classification errors in our current model. Therefore, in addition to the naïve bayes classification technique, we attempt to employ more sophisticated semantic processing to handle complex event descriptions in the sentences and to separate multiple observations to increase the accuracy of our model. Also, by further discerning the semantic meaning of biological observations, we classify biomedical text into more refined categories, which paints a clearer picture of different drug effects and proves to be an initial effort toward the sophisticated task of disease-drug relationship extraction.

To attain this goal, we used the Semantic Knowledge Represention (SKR) Project [31], which was initiated at the National Library of Medicine to provide usable semantic representation of biomedical free text. Using the batch mode, we passed sentences of the first and second classes (i.e. contain either a positive or negative relationship between the drug and disease) into Metamap, which amounted to 352 sentences that contain a definitive relationship.

Output from Metamap placed entities into distinctive categories, i.e. “clozapine” as Pharmacologic Substance, “reduction” as Qualitative Concept, “acute schizophrenia” as Mental or Behavioral Dysfunction. Upon close inspection, we identified 6 classes of observations (Disease or Syndrome; Mental or Behavioral

Dysfunction; Sign or Symptom; Pathologic Function; Cell or Molecular Dysfunction; Organ or Tissue Function), and manually added multiple entries to complement missing terms in the sentences. Here we define an observation as an abnormal condition affecting the body of an organism. We did not consider the terms that qualify as action or relationship words, since these classes were observed to encompass non-action words as well, i.e. “reduction” is a Qualitative Concept, but so is “possibly”. The added entries included “vacuous chewing movements,” “tongue protrusions,” “motor function,” “DNA methylation,” “neuronal toxicity,” etc.

We first used Metamap to identify all biomedical entities and observations in the sentences, then filtered out irrelevant classes and retained those with the above-mentioned biological observations. This output has the format of the drug-containing sentence followed by each biological observation and their respective class. For example,

> "Metoclopramide, the only drug approved by the FDA for treatment of diabetic gastroparesis, but used off-label for a variety of other gastrointestinal indications, has many potentially troublesome adverse neurologic effects, particularly movement disorders. "
>
> of diabetic gastroparesis, ==> Disease or Syndrome
> particularly movement disorders. ==> Disease or Syndrome

This sentence contains two biological observations: “diabetic gastroparesis” and “movement disorders,” where the drug Metoclopramide is used to treat one and induces the other. Using Metamap, we are able to separate the observations and associate each with a different word probability for the classification model. We may further employ sentence simplification to associate the drug entity to each observation

as formatted input for the MNB classification model. This method is expected to improve the current sentence classification system.

According to our definition of an observation, some sentences lack a biological effect. For example, the sentence “Furthermore, the toxic effect of chronic haloperidol on NOS system selectivity takes place in the neostriatum” (PMID: 14573391) does not contain a biological observation. One might argue that “haloperidol… takes place in the neostriatum” constitutes an observation; however, the word “takes place” is highly ambiguous and it is difficult to judge the biological effect based solely on this word. 28 such sentences are found to contain no observations. Altogether, 465 biological observations are identified from 324 sentences. This approach proves that Metamap serves as a useful tool in identifying multiple observations in biomedical text. Future work entails employing sentence simplification to associate the drug and each of its biological observations to improve the accuracy of sentence classification modeling technique.

# Chapter 5

## DISCUSSION AND FUTURE DIRECTIONS

The MNB classifier model achieves a fairly high precision, recall, and ROC area in classifying un-seen sentences retrieved from abstracts associated with tardive dyskinesia. Using the MNB classifier system, we were able to obtain a precision and recall of approximately 88%, which gives high confidence. In addition, ROC curve of 33% is considered random guessing, but we obtained a 97% ROC curve, which indicates our MNB classifier system performs significantly better than random guessing and that sentence classification is not a trivial task. To assess the ease and accuracy of classification, we use the prior probability of each class as a baseline of comparison, and obtain an accuracy of only 41.68% if we always pick the third class which has the highest probability.

Out of a total of 604 annotated sentences, 534 were classified correctly, and 70 were classified incorrectly using the MNB classifier model. More detailed statistics are included in Table 5.1. There exist some factors that may account for misclassification. First, sentences containing multiple drug names may be associated with both positive and negative words. Since different observations can refer to different drugs, there may be ambiguity and misclassification of sentences in constructing the sentence classifier. Hence, we attempted to separate distinct

Table 5.1 Detailed statistics applying the MNB classifier to un-seen sentences.

| === Stratified cross-validation === | |
|---|---|
| === Summary === | |
| Time taken to build model | 0.02 seconds |
| Correctly Classified Instances | 534 (88.4106 %) |
| Incorrectly Classified Instances | 70 (11.5894 %) |
| Kappa statistic | 0.8226 |
| Mean absolute error | 0.0961 |
| Root mean squared error | 0.2406 |
| Relative absolute error | 22.0127 % |
| Root relative squared error | 51.4952 % |
| Total Number of Instances | 604 |

observations and to increase the accuracy of our model by relaying our output data to Metamap, which proved to be an initial effort toward the sophisticated task of disease-drug relationship extraction.

Second, we have made the assumption that the first drug mentioned is always the focus of the sentence unless linked by other connector words such as “and” or “or.” However, there exist some exceptions to this rule, in which the first drug mentioned is not the subject of the sentence, so that the class of the sentence refers to some other drug that is not associated with the biological effect. A more advanced method may be developed to accurately identify the subject of the sentence. Such a method will require the use of a sophisticated syntactic parser to interpret and analyze parts of a sentence to determine the drug of interest. Complex, convoluted sentences with many subjects and verbs may also lead to classification error as the model will often find co-occurrence of contrasting terms a confusing task. Sentence simplification

may be employed here to automatically reduce the complexity of sentences in biomedical abstracts in order to improve the performance of syntactic parser and relationship extraction on the processed sentences.

Lastly, the tokenizer employed in Weka breaks sentences down to individual words based on blank spaces and punctuation. The predictive performance can be further improved by tokenizing sentences into short phrases so that words that frequently appear together (i.e. vacuous chewing movements) have a single word probability given the class and can be consistently classified to the same category.

This work may be extended to other biological diseases and can be used to mine relationships in aspects other than 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 [24] [25]. A biological process or pathway may also be associated with certain genes or proteins to understand the molecular mechanisms that underlie a disease [26] [27].

The current approach may also be improved with some future work. Extracting biological observations using Metamap is not a comprehensive method to capture all possible instantiation of observations from biomedical sentences. Some sentences simply do not contain an observation, while others may have an observation that does not belong to any specific category in Metamap. In the future, more comprehensive methods may be developed to identify additional biological observations from biomedical text. Such a method may also take the location and proximity of certain words into consideration in a ranking scheme, so that disease and drug that are

mentioned within n-grams of each other are deemed to be more significant than if they are mentioned within the entire abstract.

In addition, relationships described over several sentences using pronoun references are not captured in this work. This is the case of the sentence “This drug reduces purposeless limb movements.” In the future, we can consider using sophisticated algorithms to explore the vicinity of the sentence to find the explicit name that the reference points to, a process called anaphora resolution [23].

# Chapter 6

# CONCLUSION

The MNB classifier model achieves a fairly In this thesis work, we have manually developed and annotated a large biomedical training corpus for tardive dyskinesia by manually classifying sentences into one of three classes: the first class denoting a positive relationship between drug and disease, the second class denoting a negative relationship between drug and disease, and the third class denoting neither. Using the Multinomial Naïve Bayes modeling technique, we have generated a learning model and built a predictive classifier system using data preprocessing and filtering algorithms in Weka. To assess whether the model will generalize to an independent dataset, we used the 10-fold cross-validation method to evaluate the model using precision, recall, F-measure, and ROC area. Our precision, recall, and F-measure were approximately 88%, and ROC area was over 97%.

Finally, we looked into more sophisticated semantic processing to handle complex event descriptions in the sentences. One particular challenge in sentence classification is the co-existence of contrasting biological observations that cause confusion to the classification model. To address this ambiguity issue, we passed the output data to Metamap to identify and separate distinct biological observations in biomedical text. By further discerning the semantic meaning of biological

observations, we classified biomedical sentences into more refined categories, which conveyed a clearer picture of various drug effects and proved to be an initial effort toward the sophisticated task of disease-drug relationship extraction.

The thesis work includes various 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 includes 2968 drugs, small molecules, and drug classes; (2) biomedical training corpus on tardive dyskinesia which has been consistently and extensively annotated for classification purposes; (3) use of various pre-processing and filtering algorithms to build a MNB training model to classify un-seen sentences; and (4) separate distinct biological observations found in biomedical text using software tools that are open-source and readily available (Metamap). This thesis work will be presented as a paper submission to the 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), which will be held in Philadelphia, PA from October 4-7, 2012.

## REFERENCES


[1] Yassa, R., Jeste, DV. Gender differences in tardive dyskinesia: a critical review of the literature. 1992. Schizophr. Bull. 18, 701-715.

[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] Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res. 2008;36:W399–W405.

[6] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res.2006;34(Database issue):D668–D672.

[7] Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biological knowledge via a protein database. Curr. Issues Mol. Biol. 2001;3:47–55.

[8] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. Human gene mutation database (HGMD®): 2003 update. Hum. Mutat. 2003;21:577–581.

[9] Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res.2002;30:52–55.

[10] 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.

[11] 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.

[12] 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.

[13] Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 2009;10(Suppl 2):S6.

[14] 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.

[15] 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.

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

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

[18] Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P. EBIMed—text crunching to gather facts for proteins from Medline. Bioinformatics. 2007;23:e237–e244.

[19] 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.

[20] 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.

[21] 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.

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

[23] Segura-Bedmar I, Crespo M, de Pablo-Sánchez C, Martínez P. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinformatics.2010;11(Suppl 2):S1.

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

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

[26] 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.

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

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

[29] 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.

[30] Kohavi, Ron (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143.

[31] Rindflesch, Thomas C., Marcelo Fiszman, Halil Kilicoglu, Bisharah Libbus. Semantic Knowledge Representation Project; A report to the Board of Scientific Counselors. 2003. http://skr.nlm.nih.gov/