An Exploration Into Automated Clinical Drug Classification

By

John C. Koerner, B.S., B.M.

CAPSTONE PROJECT

Presented to the Department of Medical Informatics & Clinical Epidemiology
and the Oregon Health & Science University School of Medicine
in partial fulfillment of the requirements for the degree of
Master of Biomedical Informatics

August 2009

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

John C. Koerner

"An Exploration Into Automated Clinical Drug Classification"

Has been approved

Aaron M. Cohen, M.D., M.S.

TABLE OF CONTENTS

Abstract

Electronic health record (EHR) systems provide a means of tracking a broad range of patient health information including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports[1]. To the author's best knowledge, there exists no universal standard in medication list formatting. As a result, one can imagine the ability for a clinician to quickly extract critical details regarding a patient's medication profile may be hindered. Indeed, a recent Healthcare Information and Management Systems Society (HIMMS) publication[2] reported that approximately 25 percent of medication errors in the 2006 Pharmacopeia MEDMARX involved computer technology as a contributing cause and cited several studies documenting instances of 'terminology confusion' as a significant source of issues. We therefore believe the availability of a means of categorizing clinical drugs should therefore serve to promote greater expediency as well as realization of best treatments in the delivery of patient care.

Here, we assess the feasibility of utilizing several complimentary machine learning techniques to extract categorical information for eventual use in creating a comprehensive pharmaceutical drug/category ontology. We obtain a list of generic and proprietary drug names from the RxNorm database while using the web-based encyclopedia, Wikipedia, as our primary data set from which to extract semantic knowledge of the drugs. Support vector machine (SVM) algorithms are utilized on a pared-down, manually-curated test set in attempts to develop a robust classifier to distinguish drug class from non-drug class entries with the intent of identifying valid medication categories and subsequently using them to group drugs. We evaluate classifier performance and suggest additional approaches that may prove more effective.

Introduction

Electronic health record (EHR) systems provide a means of tracking a broad range of patient health information including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports[1] . As such, they provide clinicians an unprecedented wealth of readily-accessible knowledge from which to make critical decisions regarding patient treatment when necessary. With President Obama's recent pledge to commit approximately $20 billion in funding over the next five years to promote widespread adoption of electronic health information systems[3], the need for maximizing utility of these systems is clear.

The medication field of an EHR contains comprehensive list(s) of patient medications representing both proprietary and generic drug names. To the author's best knowledge, there exists no universal standard in the formatting of these medication lists. This is suggestive of the potential for variation in drug names representing identical, redundant or interacting chemical compounds to obfuscate the understanding of a patient's medication profile by clinicians and other caregivers and therefore hinder their ability to expediently assess best treatments and avoid potentially dangerous drug interactions. Indeed, a recent Healthcare Information and Management Systems Society (HIMMS) publication[3] reported that approximately 25 percent of medication errors in the 2006 Pharmacopeia MEDMARX involved computer technology as a contributing cause and cited several studies documenting instances of 'terminology confusion' as a significant source of issues. Thus, the availability of a manner of classifying entries in these medication lists into higher level drug categories should prove beneficial to the delivery of patient care.

Wikipedia is a freely available encyclopedia representing a knowledge base developed and maintained by a large community of users. Currently, the English version contains over 2.8 million articles and is constantly expanding[4]. As such, it represents a depth of information and coverage of topics that attracts researchers from many fields. Here, we explore the possibility of creating automated processes to extract semantic knowledge pertaining to pharmaceutical classifications from Wikipedia encyclopedia entries. In doing so, our intent is to create a comprehensive ontology of all common clinical drugs present in the RxNorm database, classifying each into one or more broader, clinically useful drug categories. The categories themselves will be determined from an analysis of page content and titled using the associated Wikipedia page titles. Though Wikipedia's "semi-structured design and idiosyncratic markup"[5] complicates the data mining process, we believe its general characteristics as a knowledge base (namely topic/concept page titles, embedded topic relation data and thorough coverage) will provide an excellent source from which to extract these broad drug categories.

Our research efforts focus on the utilization of various supervised machine learning techniques in the development of classifiers used to identify Wikipedia pages representing clinical drug categories. Upon identifying a valid category page, we hope to delineate specific drug-to-category associations based on article content (drug names) and page title (categories). In essence, we intend to extract semantic knowledge of pharmaceutical groups as well as the specific drugs they encompass in an automated fashion.

Following the work of [Gleim, Mehler, Dehmer][6], who successfully utilized both clustering and support vector machine (SVM) algorithms to categorize Wikipedia documents, we choose to explore several SVM-based approaches as they've demonstrated a combination of "high performance and efficiency with theoretic understanding and improved robustness"[7] (over other techniques) in the realm of text classification. We believe the proven success of SVMs in the case of sparse, high dimensional and noisy feature vectors[7] validates their utilization for our purposes. SVMs create a decision surface (hyperplane) based on vector inputs in an n-dimensional space representing positive and negative example classes. The associated algorithms optimize this decision hyperplane to maximize the margin (separation between closest class example and decision surface) between classes. The following figure provides a visual representation of this concept:

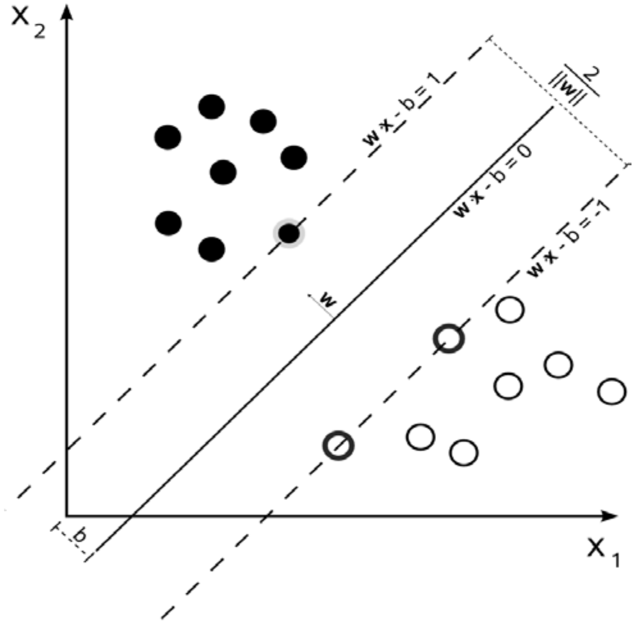

Figure 1: (SVM trained from 2 class sample, solid line represents optimal hyperplane. Dotted lines represent "support vectors") [8]

Methods

As a means of compiling a “master list” of clinical drug names, we use the March 2009 RxTerms release. RxTerms is a “drug interface terminology derived from RxNorm for prescription writing or medication history recording”[9]. As such it provides us with a reasonably comprehensive coverage (~99%)[9] of proprietary and generic names for U.S. prescribable drugs . The drug name fields of the RxTerms text are parsed and normalized by ignoring dosage and route information (e.g. oral pill). This process results in a list of approximately 12,000 unique drug name tokens. We also track generic/proprietary associations for possible later use.

Our chosen text corpus is comprised of a full article dump of the English language Wikipedia database as of March 2009. Encompassing approximately 35 gigabytes of raw data, the dump represents over 2.8 million text articles with associated XML formatting. We initially pare-down the data to a more relevant subset by including only those articles containing the term "drug" or "drugs". This, more manageable subset can be further reduced by excluding entries not containing at least one occurrence of a word token from our master drug list. For the scope of this research effort, we remain focused solely on drug categorization and are therefore not interested in making any sort of disease/medication associations. We therefore compile a list of disease categories as classified by the Medical Subject Heading (MeSH)[10] database and exclude dataset entries representing these categories (as indicated by page titles). The resultant filtered full dataset is comprised of ~34,000 articles, or an approximate 99% reduction in article count from the original.

Due to the widely varying page content of the full filtered dataset, we choose to employ a sequential, two-step binary classification routine: first to distinguish drug related from non-drug related pages, then to discover true drug category entries from the previously identified drug related pages. We therefore segregate a training set from the aforementioned test set by randomly selecting articles and manually assigning each to one of three categories:

1. Non-drug related
2. Drug related/non-category
3. Drug related/drug category.

We use this curated data to develop SVMs for each of the two classification tasks (stage 1: category 1 vs. 2 or 3, stage 2: category 2 vs. 3).

As SVM is a supervised machine learning technique, the development of our SVM-based classifier required a reasonably large set of training data from which to train the algorithm. To serve this purpose, 1000 articles were randomly selected from the full filtered dataset and manually curated into various classification categories (i.e. we read all articles individually and chose the appropriate category for each); this represents our stage 1 training set.

After the SVMs demonstrate satisfactory discriminative ability (as indicated by cross-validation results) for each classification task on the training set, we can apply it to the full filtered test set and begin to evaluate its efficacy in identifying a variety of valid drug categories. The classifier training and evaluation process will likely prove more convoluted than one may expect in a typical classification task, as we are imposing no restraints on any given category's scale. That is, the degree of granularity represented by an identified drug category is free to vary to each extreme, both fine and coarse (though one would assume more coarse, broad-based groups to be more prevalent given the classifier construct). This dynamic brings rise the question of category criteria for manual curation of our training set. For this purpose, we employ several category distinctions representing both broad and narrowly focused groupings determined by general chemical structure and/or function:

- By chemical functional group (sulfoxide, hydrazone, etc.)
- By pharmacological / biological function

Though categories comprised of basic functional groupings provide minimal utility as clinical drug classes, we include these pages for their general characteristics as categorical page entries and common usage by pharmacists and physicians. Allowing this sort of freedom with respect to the scope of permitted categories implies an additional source of complexity in the evaluation process as individual drugs will inevitably be associated with several, perhaps many respective categories. Nevertheless, we hope to achieve some insight into real performance (and thus, feasibility) of our automated drug classification approach by comparing category/drug results with physician-verified test data. Specifically, we've obtained exhaustive “gold-standard” medication lists for two universally-recognized drug categories: proton pump inhibitors (PPIs) and non-steroidal anti-inflammatory drugs (NSAIDs). A summary of the overall work-flow is illustrated by the following (Figure 2):

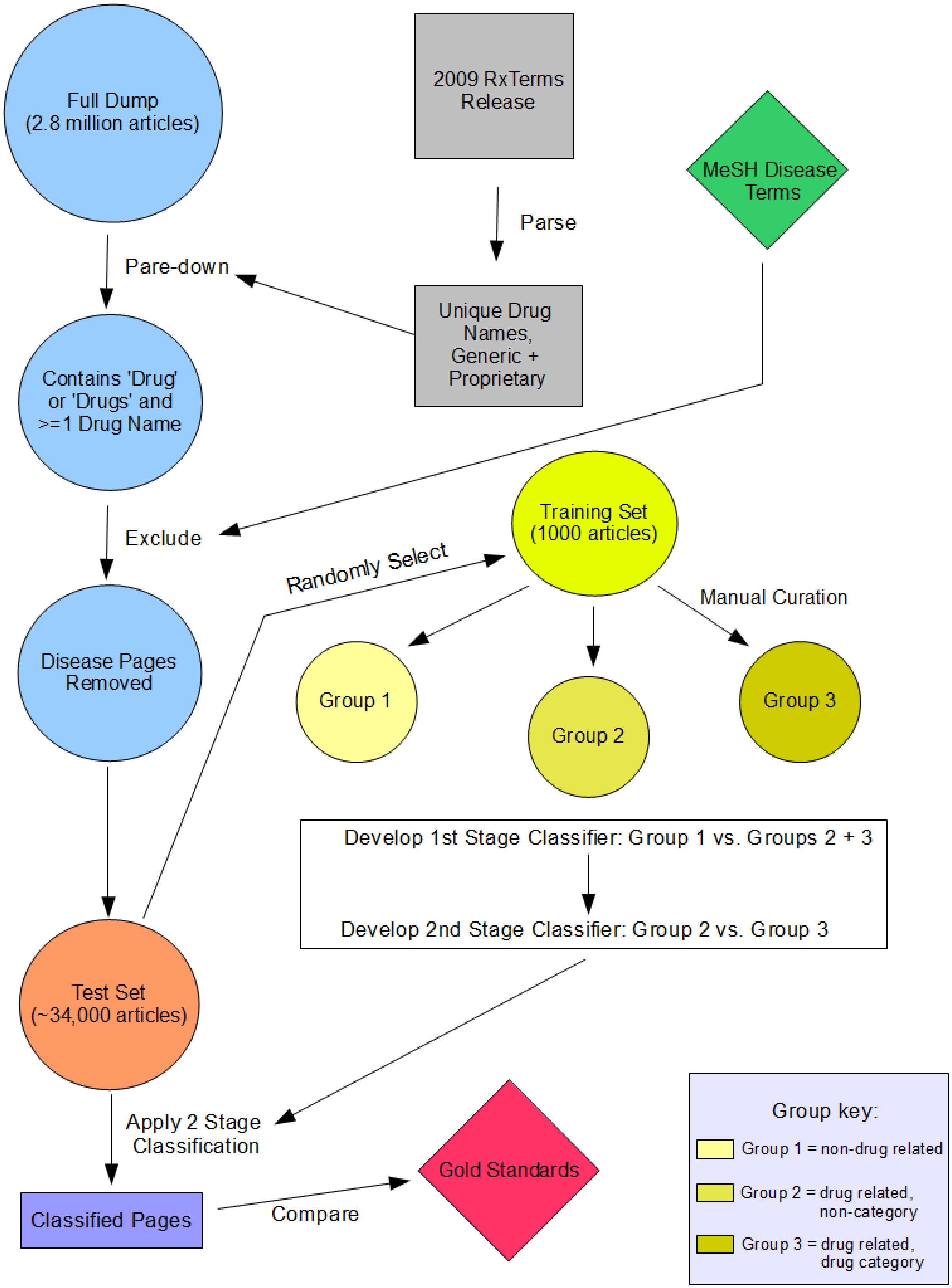
Full Dump
(2.8 million articles)
2009 RxTerms Release
Parse
MeSH Disease Terms
Pare-down
Unique Drug Names, Generic + Proprietary
Contains 'Drug' or 'Drugs' and >=1 Drug Name
Exclude
Training Set (1000 articles)
Randomly Select
Manual Curation
Disease Pages Removed
Group 1
Group 2
Group 3
Develop 1st Stage Classifier: Group 1 vs. Groups 2 + 3
Develop 2nd Stage Classifier: Group 2 vs. Group 3
Test Set (~34,000 articles)
Apply 2 Stage Classification
Gold Standards
Classified Pages
Compare
Group key:
Group 1 = non-drug related
Group 2 = drug related, non-category
Group 3 = drug related, drug category

Figure 2: Summary of work-flow

To perform the first stage binary classification task, we first parsed our filtered full data set to compile an exhaustive list of all words present in this minimized corpus. Due to the XML formatting, some pre-processing had to be performed to extract the full set of word tokens. Simple regular expressions were used to remove all text markup, punctuation, and capitalization. The resultant list represents the normalized set of all unique word tokens present in our corpus. We then compiled article abstractions consisting of the list of unique unigram (1 word) occurrences for each entry. These article-specific unigram lists serve as uniform-weighted (binary article occurrence) example feature vector inputs into our SVM classifier ($\mathrm{SVM}^{\mathrm{light}}$)[11]. That is, each article either contains a specified word token or does not. We obtain reasonable approximations of classifier performance (in terms of accuracy, precision and recall) by using $\mathrm{SVM}^{\mathrm{light}}$ in leave-one-out cross validation mode[7].

After achieving satisfactory performance with the non-relevant vs. relevant classification, we then employed a variety of approaches for the second, more difficult task of discriminating between articles representing drug categories vs. pages that are simply drug related. Initially, we attempted the full article unigram, binary occurrence (non-weighted) vector approach as above. The process was repeated using simple feature weighting for the vector inputs by number of token occurrences per article. We also utilized a 2x2 chi-square test to assess feature significance as a means of feature selection for input into the full article unigram SVM. The binary occurrence full article SVM run was repeated using only those tokens deemed significant at the 95% confidence level. An alternative feature selection process was also attempted, this time using only clinical drug occurrences (from the compiled RxTerms list) as article features, both in a binary and occurrence-weighted fashion. This should evaluate the idea of the drug names themselves being the most important features in assessing category pages.

As a result of being limited to a small training set with low positive example prevalence (~6% for stage two classification), we chose to dedicate a great deal of effort in optimizing SVM^{light} parameters using leave-one-out estimates as our performance metric. That is, the high degree of imbalance between numbers of positive and negative examples suggested the use of cost factor parameter optimization as detailed in [K. Morik, P. Brockhausen, T. Joachims][12] . The trade-off between training error and margin (-c switch in SVM^{light}) as well as the cost-factor for training errors on positive vs. negative examples (-j switch) were optimized separately, then together in an iterative fashion until an optimal tradeoff between precision/recall estimates (as determined by f-measure) and error rate was achieved. This process was performed for both the full article and RxTerm-only runs as detailed above.

Once an optimally-performing second stage classifier was decided upon, the two SVMs were run sequentially on the full test set. All drug name tokens (from RxTerm compiled list) occurring in discovered category pages were grouped and assigned to the category represented by each respective page title. Category results and individual drug groupings were compared with the physician-generated "gold standard" data sets for PPIs and NSAIDs. First and foremost, determination should be made as to whether the classifier correctly identified Wikipedia entries for PPIs and NSAIDs as category pages. If so, we can then assess the coverage and error rate of the associated drug names assigned to each. Additionally, within-group drug co-occurrences from the physician-verified categories can be compared to groupings of drug names our models assign to discovered categories as an admittedly somewhat crude means of assessing the validity of our results for non-standard categories. E.g. the list of gold standard PPIs (esomeprazole, omeprazole, etc.) is compared to drug lists compiled from discovered categories to assess drug name groupings. This should reveal any existence of true gold standard category identifications in the event that the associated category titles differ from the norm.

Results

Manual curation of the 1000 randomly selected samples from our reduced corpus (stage 1 training set) yielded 717 unrelated and 283 clinical drug related articles. Using full text unigram occurrence as uniformly weighted example features for the first stage classifier (non-related vs. drug related), LOOCV mode of SVM^{light} estimated an error rate of 5.16% with associated 88.61% precision ([true positives] / [true positives + false positives]) and 90.52% recall ([true positives] / [true positives + false negatives]); this corresponds to an F-measure of .8955. Brief excerpts from representative examples of articles representing each class are found in Table 1 (below).

Positive Page Title: "Phenylephrine"	"Phenylephrine or Neo-Synephrine is an α1-adrenergic receptor agonist used primarily as a decongestant, as an agent to dilate the pupil, and to increase blood pressure. Phenylephrine has recently been marketed as a substitute for pseudoephedrine (e.g., Pfizer's Sudafed (Original Formulation)), but there are recent claims that oral phenylephrine may be no more effective as a decongestant than a placebo."
Negative Page Title: "Johnny Cash"	"Johnny Cash (February 26, 1932–September 12, 2003), born J. R. Cash, was an American singer-songwriter and one of the most influential musicians of the 20th century. Primarily a country music artist, his songs and sound spanned many other genres including rockabilly and rock and roll (especially early in his career), as well as blues, folk and gospel." "The officers suspected that he was smuggling heroin from Mexico, but it was prescription narcotics and amphetamines that the singer had hidden inside his guitar case. Because they were prescription drugs rather than illegal narcotics, he received a suspended sentence."

Table 1: Positive (Clinical drug related) and Negative (non-related) example excerpts

As a means of examining the most informative features for this classification task, we identified those deemed significant at the 95% confidence level by a token/page occurrence chi-square test as detailed in [Cohen/ Bhupatiraju / Hersh][13]. This process revealed a total of 6603 significant word tokens representing a preponderance of negative predictive value features (as indicated by an odds ratio <1.0), many of which represent pronouns ('he', 'his', 'who', 'was'). Positive predictive value features (as indicated by an odds ratio >1.0), while less prevalent, included terms such as 'clinical', 'protein', and 'patients'. A table summarizing token/page occurrence chi-square and odds ratio for the top 60 features sorted by chi-square value is found below.

Feature	X2	OR	Feature	X2	OR	Feature	X2	OR
he	279.72	0.038	et	112.12	6.769	they	93.45	0.226
his	273.81	0.057	her	109.74	0.092	year	93.26	0.168
who	204.79	0.104	inhibitor	107.35	46.272	treatment	90.88	4.505
was	204.31	0.082	disease	103.82	5.718	tissue	90.61	13.264
clinical	199.72	14.673	liver	103.27	12.754	receptors	90.04	30.388
had	181.77	0.110	proteins	102.26	28.560	all	89.75	0.230
protein	180.93	34.915	later	101.21	0.178	first	89.12	0.237
him	165.32	0.035	new	100.04	0.213	home	88.14	0.102
patients	163.71	10.443	were	99.62	0.216	binding	87.76	11.662
receptor	151.44	36.513	city	98.67	0.065	about	87.75	0.229
acid	134.79	11.377	doses	97.85	20.972	them	87.63	0.190
enzyme	129.58	43.737	inhibition	97.49	41.977	med	87.53	12.888
time	124.5	0.169	molecular	96.95	12.657	compounds	87.53	12.888
at	123.41	0.154	after	96.93	0.222	became	87.25	0.142
effects	122.69	6.215	over	96.35	0.210	vitro	86.81	29.347
symptoms	122.41	13.608	from	96.03	0.166	molecule	85.53	20.900
1016/j	119.57	40.242	former	96.01	0.073	sci	84.79	16.692
cells	118.52	9.408	cell	94.92	6.548	took	84.53	0.069
people	118.35	0.149	she	94.23	0.074	effect	84.49	4.646
out	112.25	0.166	up	94.07	0.214	inhibitors	84.48	36.447

Table 2: Stage 1, chi-square and odds-ratio for 60 most significant features (by chi-square)

The 283 drug related articles manually identified from the 1000 article sampling (positives from stage 1: group 2 + group 3) were found to represent only 17 drug category entries, leaving 266 clinical drug related pages (stage 2 training set); this corresponds to an estimated overall drug category prevalence of 1.7% in our reduced corpus. For the stage 2 classification, all 5 runs using the default cost parameters in SVM^{light} exhibited no discriminative ability; that is, all examples were classified as negative leading to recall/precision values of zero. After optimizing the cost parameters (-c and -j switches), classifier efficacy was greatly improved. Specifically, precision/recall/f-measure estimates increased from zero for all three metrics to 0.500/0.2931/0.3704, respectively for our full text binary occurrence run while estimates for the RxTerm-only run exhibited a less dramatic increase from zero to 0.0333/0.1765/0.2308. Table 3 summarizes the cross-validation results for all stage two classifiers.

		Error	Precision	Recall	F-Measure
Stage 1	Full text (binary occurrence)	0.0516	0.8861	0.9052	0.8955
Stage 2 (Default Parameters)	Full text (binary occurrence)	0.0601	0.0000	0.0000	NA
	Full text (counted occurrences)	0.0601	0.0000	0.0000	NA
	Significant features only (binary occur.)	0.0601	0.0000	0.0000	NA
	RxTerms only (binary occurrence)	0.0601	0.0000	0.0000	NA
	RxTerms only (counted occurrences)	0.0601	0.0000	0.0000	NA
Stage 2 (Parameters Optimized)	Full text (binary occurrence, J=10, C=.01)	0.0733	0.5000	0.2941	0.3704
	RxTerms only (binary occurrence, J=30, C=.02)	0.0862	0.0333	0.1765	0.2308

Table 3: Cross-validation estimates for all training runs

Output from the parameter optimization runs for full article and RxTerm-only classifiers can be found below in tables 4 and 5, respectively. Maximum performance as indicated by error rate and f-measure are achieved with J=10, C=.01 for the full article classifier and J=30, C=.02 for the RxTerm-only SVM. Once parameters were set, we identified the optimized full article SVM as our best performing classifier with a 7.33% error rate, precision of 0.500, recall of 0.2941, f-measure of 0.3704 and utilized its modest discriminative ability for the stage two classification.

J=5	C=.001	C=.01	C=.02	C=.03	C=.04	C=.05	C=.1
error	7.76	7.33	7.33	7.33	7.33	7.33	7.33
recall	17.65	29.41	29.41	29.41	29.41	29.41	29.41
precision	42.86	50	50	50	50	50	50
F-measure	0.250	0.3704	0.370	0.370	0.370	0.370	0.370
J=10	C=.001	C=.01	C=.02	C=.03	C=.04	C=.05	C=.1
error	12.07	7.33	7.33	7.33	7.33	7.33	7.33
recall	23.53	29.41	29.41	29.41	29.41	29.41	29.41
precision	21.05	50	50	50	50	50	50
F-measure	0.222	0.370	0.370	0.370	0.370	0.370	0.370
J=15	C=.001	C=.01	C=.02	C=.03	C=.04	C=.05	C=.1
error	12.93	7.33	7.33	7.33	7.33	7.33	7.33
recall	47.06	29.41	29.41	29.41	29.41	29.41	29.41
precision	27.59	50	50	50	50	50	50
F-measure	0.348	0.370	0.370	0.370	0.370	0.370	0.370
J=20	C=.001	C=.01	C=.02	C=.03	C=.04	C=.05	C=.1
error	12.93	7.33	7.33	7.33	7.33	7.33	7.33
recall	47.06	29.41	29.41	29.41	29.41	29.41	29.41
precision	27.59	50	50	50	50	50	50
F-measure	0.348	0.370	0.370	0.370	0.370	0.370	0.370
J=30	C=.001	C=.01	C=.02	C=.03	C=.04	C=.05	C=.1
error	12.93	7.33	7.33	7.33	7.33	7.33	7.33
recall	47.06	29.41	29.41	29.41	29.41	29.41	29.41
precision	27.59	50	50	50	50	50	50
F-measure	0.348	0.370	0.370	0.370	0.370	0.370	0.370
J=250							
	C=.001	C=.01	C=.02	C=.03	C=.04	C=.05	C=.1
error	12.93	7.33	7.33	7.33	7.33	7.33	7.33
recall	47.06	29.41	29.41	29.41	29.41	29.41	29.41
precision	27.59	50	50	50	50	50	50
F-measure	0.348	0.370	0.370	0.370	0.370	0.370	0.370

Table 4: Leave-one-out performance estimates for binary full article optimization runs

J=5	**C=.001**	**C=.01**	**C=.02**	**C=.03**	**C=.04**	**C=.05**	**C=.1**	**C=.2**	**C=.3**	**C=.4**
error	NA	NA	8.1900	7.7600	7.7600	8.1900	8.1900	8.1900	8.1900	7.7600
recall	NA	NA	0.0000	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	NA	NA	0.0000	33.3330	33.3330	25.0000	25.0000	25.0000	25.0000	33.3330
F-meas				0.1000	0.1000	0.0952	0.0952	0.0952	0.0952	0.1000
J=10										
error	NA	7.3300	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	NA	11.7600	11.7600	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	NA	50.0000	28.5700	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas		0.1904	0.1666	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=15										
error	92.6700	9.0500	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	11.7600	11.7600	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	25.0000	28.5700	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1600	0.1666	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=20										
error	92.6700	13.3600	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	17.6500	11.7600	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	15.0000	28.5700	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1622	0.1666	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=30										
error	92.6700	34.0500	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	41.1800	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	9.2100	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3330
F-meas	0.1366	0.1505	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=40										
error	92.6700	47.8400	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	52.9000	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	8.0400	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1396	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=50										
error	92.6700	52.5900	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	52.9400	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	7.3200	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1286	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=75										
error	92.6700	52.5900	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	52.9400	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	7.3200	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1286	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=100										
error	92.6700	52.5900	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	52.9400	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	7.3200	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3330
F-meas	0.1366	0.1286	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=150										
error	92.6700	52.5900	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	52.9400	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	7.3200	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1286	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000
J=200										
error	92.6700	52.5900	8.6200	8.6200	9.0500	9.0500	8.1900	8.1900	8.1900	7.7600
recall	100.0000	52.9400	17.6500	11.7600	5.8800	5.8800	5.8800	5.8800	5.8800	5.8800
precision	7.3300	7.3200	33.3300	28.5700	16.6700	16.6700	25.0000	25.0000	25.0000	33.3300
F-meas	0.1366	0.1286	0.2308	0.1666	0.0869	0.0869	0.0952	0.0952	0.0952	0.1000

Table 5: Leave-one-out performance estimates for binary RxTerm-only optimization runs

Excerpts of representative examples for both a drug category page (group 3, positive) and non-category page (group 2, negative) are shown in Table 6 (below). Token/page occurrence chi-square test revealed 5896 significant features for this stage 2 task, representing an overwhelming preponderance of positive predictive value features (as indicated by OR > 1.0). Table 7 shows leave-one-out estimates for token/page occurrence chi-square and odds ratio for the top 60 features sorted by chi-square value.

Positive Page Title: “Category:Leukotriene antagonists”	“Inhibitors of leukotriene action in asthma. Main members are montelukast and zafirlukast.” “The following 5 pages are in this category, out of 5 total. This list may not reflect recent changes (learn more). A * Ablukast * Amlexanox M * Montelukast P * Pranlukast Z * Zafirlukast “
Positive Page Title: “Aminoglycoside”	“An aminoglycoside is a molecule composed of a sugar group and an amino group. Several aminoglycosides function as antibiotics that are effective against certain types of bacteria. They include amikacin, arbekacin, gentamicin, kanamycin, neomycin, netilmicin, paromomycin, rhodostreptomycin[2], streptomycin, tobramycin, and apramycin. Anthracyclines are another group of aminoglycosides. These compounds are used in chemotherapy.”
Non-Category Example Page Title: “Nicotine Gum”	“Nicotine gum is a type of chewing gum that delivers nicotine to the body. It is used as an aid in smoking cessation and in quitting smokeless tobacco. The nicotine is delivered to the bloodstream via absorption by the tissues of the mouth. It is currently available over-the-counter in Europe, the US and elsewhere. The pieces are usually available in individual foil packages and come in various flavors including orange, and mint. Each piece typically contains 2 or 4 mg of nicotine, roughly the nicotine content of 1 or 2 cigarettes, with the appropriate dosage depending on the smoking habits of the user. Popular brands include Nicoderm/Nicorette and Nicotinell.”

Table 6: Stage 2 classification, positive and negative examples

Feature	X2	OR	Feature	X2	OR	Feature	X2	OR
classes	24.82	15.14	mimetic	24.64	40.5	ligand	16.74	11.31
molpharm	24.64	40.5	peptidase	24.64	40.5	carbonyl	16.74	11.31
theamino	24.64	40.5	is	24.36	0.08	transmembrane	16.74	11.31
phase-iii	24.64	40.5	moiety	23.49	19.02	alkaloids	16.74	11.31
bioavailable	24.64	40.5	alkyl	23.49	19.02	pronounced	16.74	11.31
takeda	24.64	40.5	bonding	23.49	19.02	antagonists	15.39	8.25
stabilized	24.64	40.5	block	23.14	9.97	the	15.39	0.12
structure-activity	24.64	40.5	actions	23.14	9.97	analogs	14.63	13.38
strengthened	24.64	40.5	residues	21.81	12.56	scopolamine	14.63	13.38
pharmacophore	24.64	40.5	bond	20.11	9.55	contractility	14.63	13.38
umi	24.64	40.5	thechemical	19.68	14.2	belladonna	14.63	13.38
non-competitive	24.64	40.5	merck	19.68	14.2	prostate	14.63	13.38
pyrimidinedione	24.64	40.5	table	19.68	14.2	effector	14.63	13.38
thetakeda	24.64	40.5	optimization	18.62	20.16	novartis	14.63	13.38
pyrimidine	24.64	40.5	aryl	18.62	20.16	blood-brain	14.63	13.38
sir2	24.64	40.5	substituents	18.62	20.16	fused	14.63	13.38
uracil	24.64	40.5	templates	18.62	20.16	interacts	14.63	13.38
theenzyme	24.64	40.5	sciencedirect	18.62	20.16	blockers	14.4	9.38
acidgroups	24.64	40.5	represented	18.62	20.16	reversible	14.4	9.38
proquest	24.64	40.5	catalytic	17.19	9.33	inhibitors	14.37	5.85

Table 7: Stage 2, chi-square and odds-ratio for 60 most significant features (by chi-square)

A summary of all datasets used including category information (where appropriate) can be found in Table 8. Results from the two stage classification on the full filtered test set are seen in Table 9. Manual review of the second stage SVM-identified category pages revealed a total of 68 unique pages. Excluding proprietary drug names for the sake of clarity, the simplistic method of assigning all drug terms present in identified articles to each page's respective title yielded 919 total drug-to-category associations as shown in Table 10.

Data Set Name	Total Pages	Positive Pages	Negative Pages
Full Dump	2800000	NA	NA
'Drug' or 'Drugs'	89524	NA	NA
Disease Pages Removed	89072	NA	NA
Full Filtered Test Set	33734	8054	25680
Stage 2 Test Set	8054	68	7986
Stage 1 Training Set	1000	283	717
Stage 2 Training Set	283	17	266

Table 8: Summary details of all data sets

Discovered Categories
1,2,3-triazole
3-quinuclidinyl benzilate
acetylcholine
aldosterone antagonist
alpha-2 blocker
amine
analog (chemistry)
atropine
azimilide
beta blocker
diabetic neuropathy
diaminopyrimidine
diastolic dysfunction
digitalis purpurea
dimethylheptylpyran
dipeptidyl peptidase-4 inhibitor
ditran
diuretic
ductus arteriosus
eukaryotic initiation factor
exercise and stimulants
extrapyramidal system
federal analog act
glossary of diabetes
harmine
homatropine
inotrope
ion channel
isomer
isothiazole
leukotriene antagonist
list of biochemistry topics
list of biology topics
mast cell
microbial toxins
microglia
monoamine transporter
montelukast
muscarinic acetylcholine receptor
natural product
neuromuscular-blocking drug
noncovalent bonding
norepinephrine
norepinephrine reuptake inhibitor
obidoxime
pde3 inhibitor
piperidinedione
potassium-sparing diuretic
ppar modulator
propanolamine
psychedelic drug
quaternary ammonium muscle relaxants
serotonin uptake inhibitor
stimulant
stomachic
substituted amphetamines
supramolecular chemistry
template:anticholinergics
thiazide
thioxanthene
trimetaphan camsilate
tropane alkaloid
type ii topoisomerase
vascular smooth muscle
vasoconstriction
vasodilation
vasospasm
vitamin b12

Table 9: Discovered drug categories

analog (chemistry)	**dimethylheptylpyran**	**piperidinedione**	**digitalis purpurea**	**noncovalent bonding**	**1,2,3-triazole**	**stomachic**
vitamin b	choline	methyprylon	calcium	air	air	histamine
oseltamivir	tetrahydrocannabinol	nitrogen	urea		nitrogen	
	lithium	glutethimide	digitoxin			
	water		potassium			
	acetate		digoxin			
harmine	**isothiazole**	**ductus arteriosus**	**eukaryotic initiation factor**	**federal analog act**	**substituted amphetamines**	**alpha-2 blocker**
dopamine	nitrogen	oxygen	methionine	cocaine	cocaine	dopamine
phenelzine	sulfur	indomethacin	iron	air	amphetamine	atipamezole
melatonin	ziprasidone	ibuprofen		glutamate	iodine	piperazine
iron						
trimetaphan camsilate	**azimilide**	**diaminopyrimidine**	**potassium-sparing diuretic**	**leukotriene antagonist**	**microbial toxins**	**diastolic dysfunction**
acetylcholine	acetylcholine	folate	amiloride	oxygen	zinc	calcium
choline	choline	trimethoprim	triamterene	montelukast	calcium	oxygen
trimethaphan	piperazine	trimetrexate	spironolactone	zileuton	choline	air
	sotalol	pyrimethamine	iron	zafirlukast	escherichia coli	enalapril
	potassium		potassium			ramipril
			eplerenone			
pde3 inhibitor	**ditran**	**ppar modulator**	**supramolecular chemistry**	**aldosterone antagonist**	**thioxanthene**	**dipeptidyl peptidase-4 inhibitor**
adenosine	choline	cholesterol	biotin	spironolactone	chlorprothixene	glucose
nitrogen	trifluoperazine	gemfibrozil	air	iron	oxygen	sitagliptin
theophylline	scopolamine	glucose	urea	potassium	nitrogen	glucagon
cilostazol	ketamine	air	silver	water	sulfur	
adenosine monophosphate	glycolate	fenofibrate	hemin	eplerenone	thiothixene	
milrinone	amphetamine	clofibrate	iron			
papaverine	acetate	ibuprofen	water			
vascular smooth muscle	**monoamine transporter**	**homatropine**	**psychedelic drug**	**exercise and stimulants**	**montelukast**	**obidoxime**
epinephrine	cocaine	acetylcholine	choline	calcium	acetic acid	acetylcholine
oxygen	epinephrine	choline	tetrahydrocannabinol	cocaine	oxygen	pralidoxime
norepinephrine	norepinephrine	homatropine	piperazine	epinephrine	air	choline
doxazosin	dopamine	hydrocodone	ketamine	norepinephrine	montelukast	nitrogen
helium	methamphetamine	papaverine	morphine	dopamine	loratadine	histidine
prazosin	fluoxetine	acetate	amphetamine	ephedrine	theophylline	atropine
	bupropion	atropine	atropine	methamphetamine	zileuton	phosphorus
	amphetamine			caffeine	histamine	
	glutamate			amphetamine	zafirlukast	
tropane alkaloid	**type ii topoisomerase**	**mast cell**	**isomer**	**norepinephrine reuptake inhibitor**	**propanolamine**	**ion channel**
cocaine	novobiocin	calcium	oxygen	epinephrine	betaxolol	calcium
choline	air	air	urea	norepinephrine	penbutolol	acetylcholine
hyoscyamine	adenosine	heparin	methamphetamine	dopamine	alcohols	choline
scopolamine	teniposide	montelukast	caffeine	desipramine	atenolol	oxygen
belladonna alkaloids	magnesium	helium	theophylline	nortriptyline	metoprolol	urea
atropine	tyrosine	silver	acetylene	atomoxetine	pindolol	adenosine
	etoposide	iron	isopropyl alcohol	maprotiline	bisoprolol	magnesium
	isoleucine	histamine	amphetamine	bupropion	nadolol	helium
		nedocromil	phentermine	venlafaxine	propranolol	potassium
		zafirlukast		mazindol	phenylpropanolamine	water
				duloxetine	ritodrine	lidocaine
					timolol	glutamate
					acebutolol	
neuromuscular-blocking drug	**atropine**	**microglia**	**3-quinuclidinyl benzilate**	**serotonin uptake inhibitor**	**vasospasm**	**thiazide**
acetylcholine	acetylcholine	air	propylene glycol	citalopram	cholesterol	cholesterol
choline	pralidoxime	dopamine	acetylcholine	trazodone	calcium	calcium
succinylcholine	choline	secretin	choline	epinephrine	oxygen	amiloride
doxacurium	oxygen	nitric oxide	hyoscyamine	fluvoxamine	nifedipine	epinephrine
pancuronium	hyoscyamine	tyrosine	tetrahydrocannabinol	norepinephrine	verapamil	glucose
rapacuronium	phenylephrine	tetracycline	air	escitalopram	isosorbide	norepinephrine
vecuronium	nitrogen	minocycline	scopolamine	fluoxetine	sildenafil	air
pipecuronium	diphenhydramine	iron	pyridostigmine	sertraline	helium	sodium chloride
mivacurium	tropicamide	potassium	ketamine	paroxetine	nitric oxide	metolazone
tubocurarine	opium	lovastatin	iron	venlafaxine	nitroglycerin	chlorothiazide
histamine	benztropine	acetate	fentanyl	amoxapine	propranolol	hydrochlorothiazide
potassium	water	glutamate	glycolate	clomipramine	isosorbide dinitrate	potassium chloride
atracurium	pilocarpine		water	duloxetine		folic acid
cisatracurium	sulfur		physostigmine			cysteine
gallamine	physostigmine		acetate			potassium
rocuronium	atropine		dimethyl sulfoxide			
			atropine			

Table 10: Discovered drug-to-category associations

extrapyramidal system	list of biology topics	stimulant	inotrope	natural product	quaternary ammonium - muscle relaxants	vasoconstriction
chlorpromazine	ethanol	cocaine	calcium	cholesterol	acetylcholine	calcium
quetiapine	citric acid	epinephrine	procainamide	ethanol	choline	boron
choline	glucose	methylphenidate	quinidine	cocaine	succinylcholine	cocaine
metoclopramide	air	norepinephrine	epinephrine	choline	oxygen	acetylcholine
risperidone	colchicine	dopamine	norepinephrine	vincristine	doxacurium	choline
dopamine	urea	ephedrine	verapamil	captopril	pancuronium	epinephrine
aripiprazole	adenosine	modafinil	dopamine	emetine	rapacuronium	methylphenidate
diphenhydramine	starch	piperazine	diltiazem	reserpine	nitrogen	norepinephrine
haloperidol	nitrogen	aspirin	metoprolol	ipecac	vecuronium	phenylephrine
clozapine	chlorophyll	ketamine	glucagon	quinine	pipecuronium	ephedrine
benztropine	streptomycin	methamphetamine	flecainide	opium	opium	adenosine
olanzapine	iron	pseudoephedrine	carvedilol	tetracycline	mivacurium	inositol
promazine	adenosine monophosphate	caffeine	bisoprolol	morphine	tubocurarine	glycerol
amoxapine	water	nicotine	theophylline	nicotine	iron	nitric oxide
trihexyphenidyl	lactic acid	bupropion	digoxin	digitoxin	histamine	pseudoephedrine
ziprasidone	cellulose	fentanyl	milrinone	tubocurarine	water	oxymetazoline
		amphetamine	dobutamine	iron	atracurium	iron
		mazindol	inamrinone	water	cisatracurium	histamine
		dextroamphetamine	disopyramide	lovastatin	gallamine	adenosine monophosphate
		phentermine	isoproterenol	chloramphenicol	rocuronium	amphetamine
				atropine		tetrahydrozoline
				paclitaxel		arginine
				pectin		

muscarinic acetylcholine-receptor	diabetic neuropathy	norepinephrine	diuretic	vasodilation	amine	template:anticholinergics
calcium	citalopram	cocaine	calcium	calcium	zinc	ethanol
acetylcholine	epinephrine	acetylcholine	ethanol	ethanol	ethanol	acetylcholine
choline	oxygen	choline	boron	boron	chlorpromazine	pralidoxime
epinephrine	glucose	epinephrine	amiloride	epinephrine	epinephrine	choline
norepinephrine	norepinephrine	glucose	glucose	oxygen	oxygen	succinylcholine
air	imipramine	methylphenidate	bumetanide	glucose	norepinephrine	doxacurium
carbachol	carbamazepine	norepinephrine	dorzolamide	norepinephrine	alcohols	pancuronium
adenosine	air	alcohols	lithium	tetrahydrocannabinol	imipramine	hyoscyamine
methacholine	desipramine	guanethidine	dopamine	adenosine	phenylephrine	orphenadrine
inositol	pregabalin	dopamine	aspirin	isosorbide	chlorpheniramine	biperiden
scopolamine	nortriptyline	phenoxybenzamine	sodium chloride	sildenafil	air	metocurine
helium	helium	tetrabenazine	triamterene	helium	lithium	scopolamine
opium	oxcarbazepine	reserpine	furosemide	amyl nitrite	dopamine	glycopyrrolate
nitric oxide	glycerol	desipramine	torsemide	opium	desipramine	mecamylamine
oxybutynin	nitric oxide	adenosine	chlorothiazide	nitric oxide	ephedrine	homatropine
nicotine	topiramate	nitrogen	silver	nitroglycerin	nitrogen	vecuronium
ipratropium	fluoxetine	atomoxetine	hydrochlorothiazide	niacin	nortriptyline	diphenhydramine
tiotropium	sertraline	tyrosine	spironolactone	pentaerythritol tetranitrate	acetone	tropicamide
potassium	amitriptyline	amino acids	caffeine	carbon dioxide	formaldehyde	pipecuronium
pilocarpine	sorbitol	fluoxetine	theophylline	vardenafil	methamphetamine	opium
tolterodine	paroxetine	histamine	iron	iron	phenol	oxybutynin
bethanechol	water	tryptophan	flumethiazide	histamine	isopropanol	mivacurium
gallamine	fructose	amphetamine	potassium	pentaerythritol	copper	cyclopentolate
atropine	gabapentin	levodopa	water	potassium	amitriptyline	nicotine
	duloxetine	alanine	indapamide	lactic acid	carbon dioxide	tubocurarine
	pectin	phenylalanine	amphotericin b	nitroprusside	hydrochloric acid	ipratropium
		dextroamphetamine	bendroflumethiazide	papaverine	iron	histamine
			mannitol	isosorbide mononitrate	histamine	tiotropium
			lithium citrate	arginine	promazine	atracurium
			arginine	isosorbide dinitrate	water	cisatracurium
			acetazolamide	tadalafil	amphetamine	tolterodine
					amoxapine	gallamine
					clomipramine	trihexyphenidyl
					hydroiodic acid	atropine
					sulfur	rocuronium
					pheniramine	
					dimethyl sulfoxide	

Table 10 (cont.): Discovered drug-to-category associations

vitamin b12	**beta blocker**	**glossary of diabetes**	**acetylcholine**	**list of biochemistry topics**
liver extract	betaxolol	cholesterol	calcium	calcium
pantoprazole	metipranolol	calcium	acetylcholine	ethanol
cholesterol	cocaine	capsaicin	pralidoxime	acetylcholine
calcium	epinephrine	glimepiride	choline	choline
ethanol	oxygen	epinephrine	succinylcholine	calcitriol
folate	esmolol	oxygen	epinephrine	epinephrine
oxygen	penbutolol	glucose	acetic acid	acetic acid
primidone	glucose	alcohols	oxygen	oxygen
famotidine	norepinephrine	chlorpropamide	chloride ion	vitamin d
omeprazole	air	fluorescein	doxacurium	thrombin
phenytoin	atenolol	air	norepinephrine	citric acid
phenobarbital	metoprolol	urea	pancuronium	glucose
air	labetalol	acetohexamide	trimethaphan	glutamine
neomycin	pindolol	inositol	carbachol	air
colchicine	furosemide	starch	edrophonium	dopamine
adenosine	glucagon	nitrogen	metocurine	secretin
intrinsic factor	carvedilol	acetone	scopolamine	colchicine
pyridoxine	silver	glucagon	pyridostigmine	urea
esomeprazole	levobunolol	glipizide	nitrogen	adenosine
chlorophyll	bisoprolol	glyburide	mecamylamine	thyroxine
methionine	nadolol	glycerol	tacrine	starch
potassium citrate	melatonin	amino acids	vecuronium	nitrogen
potassium chloride	nitroglycerin	metformin	rivastigmine	glucagon
metformin	carteolol	xylitol	opium	threonine
cimetidine	propranolol	tolazamide	neostigmine	formaldehyde
nitrous oxide	phentolamine	carbon dioxide	mivacurium	chlorophyll
nizatidine	sotalol	tolbutamide	nicotine	methionine
nicotine	potassium	iron	tubocurarine	corticotropin
folic acid	water	histamine	ipratropium	sincalide
rabeprazole	amphetamine	sorbitol	tiotropium	tyrosine
cysteine	timolol	potassium	potassium	phenol
iron	nebivolol	water	pilocarpine	acetylcysteine
histamine	acebutolol	lactic acid	atracurium	nitroglycerin
metronidazole		fructose	echothiophate	progesterone
potassium		cellulose	cisatracurium	hemin
lansoprazole		lactase	malathion	oxytocin
water		lactose	physostigmine	chorionic gonadotropin
zidovudine			donepezil	histidine
aminosalicylic acid			acetate	cysteine
charcoal			cevimeline	factor viii
vitamin a			bethanechol	thyrotropin-releasing hormone
cobalamins			galantamine	iron
sodium thiosulfate			atropine	histamine
pentagastrin			rocuronium	dactinomycin
cholestyramine				interferon type ii
vitamin b				isoleucine
hydroxocobalamin				adenosine monophosphate
colestipol				potassium
chloramphenicol				tryptophan
activated charcoal				water
ranitidine				octreotide
salicylic acid				lactic acid
				cyclosporine
				lysine
				corticotropin-releasing hormone
				somatropin
				alanine
				triiodothyronine
				sulfur
				arginine
				glycine
				somatotropin
				gonadorelin
				polymyxin b
				phenylalanine
				estradiol
				cellulose
				glutamate
				phosphorus

Table 10 (cont.): Discovered drug-to-category associations

Manual review of the positive drug category pages identified by the SVM shows that the “gold standard” categories (PPIs and NSAIDs) were not discovered by the stage 2 classifier. Closer examination of the stage 2 test set revealed that both PPIs and NSAIDs were identified by the stage 1 classifier as being drug related but the stage 2 classifier was unable to identify them as true drug categories. The second stage classifier was, however, able to identify several within-group associations with gold standard drugs; that is, we see several instances of gold standard drug co-occurrence being replicated in discovered category pages. Examining these drug co-occurrences and corresponding Wikipedia pages, we determine conclusively that the discovered categories 'ductus arteriosus' and 'ppar modulator' do not represent NSAIDs while the category 'vitamin b12' is not analogous to the PPIs gold standard category. It is interesting to note that in each case, the gold standard categories were explicitly mentioned by correct name on discovered category pages almost immediately preceding the occurrence of the drug terms. E.g., from the 'ductus artertiosus' entry: “Closure may be induced with a drug class known as NSAIDs such as indomethacin or ibuprofen”. This characteristic could certainly be leveraged in developing a more sophisticated means of parsing discovered category pages for drug-to-category associations. Table 11 provides a summary of these results.

Gold Standard Associations

Gld Std.	Discovered	Gld Std.	Discovered		
NSAIDs	**ductus arteriosus**	**PPIs**	**vitamin b12**		
Bromfenac	Ibuprofen	Esomeprazole	activated charcoal	lansoprazole	vitamin a
Celecoxib	Indomethacin	Lansoprazole	adenosine	liver extract	vitamin b
Diclofenac	oxygen	Misoprostol	air	metformin	w ater
Etodolac		Omeprazole	aminosalicylic acid	methionine	zidovudine
Fenoprofen	**ppar modulator**	Pantoprazole	calcium	metronidazole	
Flurbiprofen		Rabeprazole	charcoal	neomycin	
Ibuprofen	air		chloramphenicol	nicotine	
Ibuprofen-Diphenhydramine	cholesterol		chlorophyll	nitrous oxide	
Indomethacin	clofibrate		cholesterol	nizatidine	
Ketoprofen	fenofibrate		cholestyramine	omeprazole	
Ketorolac Tromethamine	gemfibrozil		cimetidine	oxygen	
Lansoprazole-Naproxen	glucose		cobalamins	pantoprazole	
Meclofenamate	Ibuprofen		colchicine	pentagastrin	
Mefenamic Acid			colestipol	phenobarbital	
Meloxicam			cysteine	phenytoin	
Nabumetone			esomeprazole	potassium	
Naproxen			ethanol	potassium chloride	
Naproxen Sodium			famotidine	potassium citrate	
Oxaprozin			folate	primidone	
Piroxicam			folic acid	pyridoxine	
Rofecoxib			histamine	rabeprazole	
Sulindac			hydroxocobalamin	ranitidine	
Tolmetin			intrinsic factor	salicylic acid	
Valdecoxib			iron	sodium thiosulfate	

Table 11: Gold standard drug associations

Discussion

The robust discriminative ability of the first stage binary classifier was somewhat surprising considering the SVM example vectors included nearly 57,000 features. Our results seem to support Thorsten Joachims' assertion that “in text categorization there are only very few irrelevant features”[14] and affirm the relative resiliency to overfitting exhibited by support vector machine-based classifiers. Examination of the most informative features for this task reveals an overwhelming prevalence of those providing negative predictive value. We see that pronouns are strong predictive features which, upon brief perusal of the training set, makes intuitive sense as a large proportion of the non-drug relevant articles represent notable individuals or socio-political groups and their respective drug activities. Unfortunately, it provides minimal insight into potential approaches for the second stage classification task.

We were disappointed, though not surprised with the modest discriminative ability exhibited by our various SVMs for the second stage classification. Following the full text binary occurrence run, both feature set pruning approaches (significant tokens only, RxTerms only) yielded no benefit. Given the sparsity of positive example data we had to work with, it stands to reason that maximum classifier robustness was achieved when all document features were included. The high degree of imbalance between numbers of positive and negative examples suggested the use of cost factor parameter optimization as detailed in [K. Morik, P. Brockhausen, T. Joachims][12]. Specifically, we took advantage of SVM^{light}'s ability to adjust cost factoring in its weighting of false negatives vs. false positives ('j' switch) in addition to trade-off between training error and margin ('c' switch). Our optimization results as indicated by cross-validation output suggest a modest degree of sensitivity to both parameters; this is especially evident in the binary RxTerm-only runs. As a result, the optimization process involved some element of subjectivity in deciding on final values to use on the filtered full data set. Every effort was taken to achieve an optimal balance between precision/recall as evident by f-measure while minimizing error rate.

Though significantly lacking with regards to medication class coverage, the categories identified by the SVMs are generally within reason, given our defined category criteria. Notable exceptions to this include 'vasospasm', 'diabetic neuropathy' and 'diastolic dysfunction' as well as 'federal analog act' and the glossary/list pages. Despite filtering out all pages representing exact MeSH disease categories from our data set, 'vasospasm', 'diabetic neuropathy' and 'diastolic dysfunction' disorders remained. More thorough disease/disorder lists would alleviate this source of noise in the data. While a brief examination of article features for the 'federal analog act' reveals probable explanation for its false positive classification(presence of drug class tokens 'stimulant', 'depressant' in addition to several of the most significant positive predictive features: 'actions' 'represented') no clear means of avoiding these erroneous results is immediately apparent. Similarly, the glossary/list pages contain a preponderance of individual drug and drug class terms in addition to high significance positive predictive value tokens 'classes', 'actions', 'represented', to name a few. Perhaps a larger training set would alleviate these and similar issues.

Generally speaking, our various SVM-based approaches for the stage two classifier have proven somewhat ineffective given the available training data. Examining the discovered category results, we see overall classifier sensitivity was notably low given the apparent lack of drug category coverage. One could suggest that the unigram word occurrences alone, across full articles may not be sufficient to provide substantive traction in SVM-based discrimination between drug related and drug-category entries. That is not to say we believe the classification task to be intractable, however. Unfortunately, manual curation of training examples proved rather expensive in terms of man hours. This led to our efforts being somewhat stymied by the relatively small number of known positives in the training set. With only 17 positive training examples, the opportunity we afforded for within-class similarities to emerge was undoubtedly insufficient, especially considering the drastically varying page structure/content we observe (see Table 2, “Category: Leukotriene vs. “Nicotine Gum”).

Beyond simple data sparsity issues, we suggest several possibilities for improvement within the domain of SVMs in addition to other machine learning techniques. Though not described here, exploratory efforts taken to create normalized article location-specific features showed significant potential. Simply making the distinction between drug tokens located in the head/intro paragraph vs. the remainder of the page provided a modest though surprising degree of discriminative power (F-measure of ~.15 using only drug token occurrence in intro vs. remainder) . Similarly, we believe the opportunity for substantial utility exists in developing an effective means of identifying article similarities based on congruous page structure. Evidence for this assertion resides in our observation that entries representing related concepts often exhibit many similar if not identical section headings (albeit sometimes reordered). As we've elucidated with the Table 2 category excerpts, there is a great deal of variation between articles belonging to the drug category class, as currently defined. Given this general characteristic disparity between drug category pages, one could argue this strict two class framework for the stage two classification is unnecessarily prohibitive and suggest a multi-class approach as a more appropriate model construct. Clustering algorithms (k-means etc.) could be used to create initial page groupings by structural similarity from which to randomly select and curate training examples. A binary SVM could then be implemented for each article cluster to identify drug category pages. Presumably, the initial page clustering would render this second stage classification task far-more tractable.

Conclusion

Until some viable means of EHR medication list normalization can be implemented on a broad scale, the potential for computer technology-related medication errors as a result terminology confusion[3] will persist. We have provided thorough analysis of a two stage support vector machine-based approach to automated drug category extraction from Wikipedia pages. Though our inability to robustly distinguish true drug category classes from drug related pages has prevented us from creating a comprehensive drug-to-category ontology, we have gained significant insight into the nature of the task and identified specific areas upon which future research may be improved.

References

(1) HIMSS – Electronic Health Record. [Online]. Available from http://www.himss.org/ASP/topics_ehr.asp. Cited September 2009.

(2) HIMSS EHR Usability Task Force. (2009) Defining and Testing EMR Usability: Principles and Proposed Methods of EMR usability Evaluation and Rating. Available from http://www.himss.org/content/files/HIMSS_DefiningandTestingEMRUsability.pdf.

(3) Recovery.gov. [Online]. Available from http://www.recovery.gov/. Cited Septermber 2009.

(4) Wikipedia, Size of Wikipedia. [Online]. Available from http://en.wikipedia.org/wiki/Size_of_Wikipedia. Cited March 2009.

(5) Jamie Taylor, Colin Evans, Toby Segaran. (2008) Machine Learning for Knowledge Extraction from Wikipedia & Other Semantically Weak Sources. In *Proceedings of the O'Reilly Open Source Convention (OSCON) 2008.*

(6) Rudiger Gleime , Alexander Mehler, Matthias Dehmer. (2007). Web Corpus Mining by instance of Wikipedia. In *Proc. 2nd Web as Corpus Workshop at EACL 2006.*

(7) Joachims, T., Learning to Classify Text Using Support Vector Machines. [Online]. Dissertation, Kluwer, 2002. Available from http://textclassification.joachims.org/. Cited 2009 March 8.

(8) Wikipedia, Support Vector Machine. [Online]. Available from http://en.wikipedia.org/wiki/Support_vector_machine. Cited March 2009.

(9) Fung, Kin Wah. RxTerms: the interface terminology to RxNorm. [Online]. Available from http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/data/RxTerms_demo_AMIA2008.pdf

(10) United States National Library of Medicine, National Institutes of Health, Medical Subject Headings. Available from http://www.nlm.nih.gov/mesh/ . Cited March 2009.

(11) Joachims, T., SVM-Light Support Vector Machine [homepage on the Internet]. Available from http://svmlight.joachims.org/. Cited February 2009.

(12) K. Morik, P.Brockhausen, T. Joachims. (1999) Combining statistical learning with a knowledge-based approach – A case study in intensive care monitoring. In *Proc. 16th Int'l Conf. On Machine Learning* (ICML-99), 1999.

(13) A.M. Cohen, R.T. Bhupatiraju, W.R. Hersh. (2004) Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In *Proceedings of the Text Retrieval Conference (TREC) 2004*.

(14) Joachims, T., Text categorization with Support Vector Machines: Learning with many relevant features. Machine Learning: ECML-98 1998:137-142.