Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature:

_____ _____

Walter Scott Askew IV Date

Predicting Disease Comorbidity by Mining Large Text Corpora

by

Walter Scott Askew IV

Adviser Eugene Agichtein

Department of Mathematics and Computer Science

_____

Eugene Agichtein
Adviser

_____

James Lu
Committee Member

_____

Susan Tamasi
Committee Member

_____

Date

Predicting Disease Comorbidity by Mining Large Text Corpora

By

Walter Scott Askew IV

Adviser Eugene Agichtein

An abstract of
A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Mathematics and Computer Science

2009

Predicting Disease Comorbidity by Mining Large Text Corpora

By Walter Scott Askew IV

Natural language processing techniques have a variety of applications in the public health field. This paper discusses a method for predicting whether two diseases are frequently comorbid. A system is presented which applies previous work into using textual information to compute similarity between words to predict disease comorbidity. The work is based on the assumption that the rate of comorbidity between two diseases should be reflected by linguistic similarity of their cooccurrences. Perhaps most excitingly, the paper demonstrates that corpora such as web forums provide useful data for training the system. The ability to mine web based sources for new medical information has many exciting implications in public health. The web could be used to monitor disease trends and epidemic outbreaks, and to uncover new medical knowledge directly from disease suffers. The evaluation of this system shows that it performs above baseline levels in predicting frequency of comorbidity between diseases.

Predicting Disease Comorbidity by Mining Large Text Corpora

By

Walter Scott Askew IV

Adviser Eugene Agichtein

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Mathematics and Computer Science

2009

Contents

List of Tables

List of Figures

1 Introduction

Comorbidity is an important concept in public health. Two diseases are said to be comorbid if they exist in the same patient simultaneously. In individual cases, the interactions between numerous diseases in the same patient must be considered by the patient's doctor. On a public health scale, statistically significant occurrences of a comorbidity indicate a relation between the diseases. This paper summarizes an attempt to calculate the strength of the comorbidity between two diseases using textual data from various sources.

The paper's most important contribution is its use of web forums as a source of medical knowledge. By demonstrating that web forums can be mined for useful medical information, this paper suggests that the Internet could be an import public health tool.

1.1 Rationale

The guiding hypothesis of this work is that frequency of comorbidity between diseases should be reflected by linguistic similarity. This paper follows the work of researchers such as Ido Dagan [1] [2] in using textual information to compute similarity between words. Dagan has shown that similarity between two words can be computed by comparing the context in which the two words appear. For example, two words such as 'cat' and 'dog' might be computed to be similar because they frequently appear in proximity to many of the same words, such as 'pet' or 'collar.' What is original in this paper is the application of such techniques in measuring frequency of comorbidity between diseases.

The work presented here is based on the hypothesis that diseases which are highly comorbid will have similar sets of cooccurrences. That is, if the same words appear with statistically significant frequencies in discussions of two different diseases, then this should indicate some level of relatedness between the two diseases.

2 Methodology

To test the hypothesis, cooccurrence data was collected from different data sources and used to measure disease comorbidity. What follows is a description of the system designed to measure disease relatedness.

The system uses four computational stages in order to make experimentation more efficient. New experiments which only require changes to the

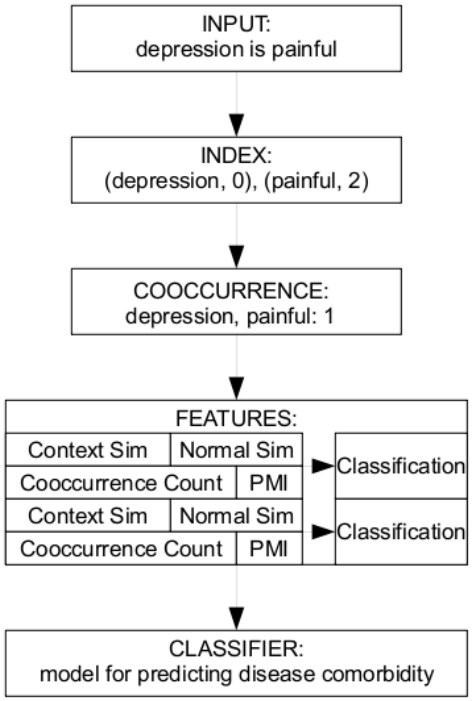

Figure 1: overall depiction of system

third stage of the system need not involve computation in the first. The first stage processes and indexes a corpus so that the corpus may be efficiently used in future stages. The second stage gathers cooccurrence counts from the index produced in the first. The third stage uses the gathered coccurrence counts to calculate disease relationships. Finally, the calculated disease relationship values are used as features to train a classifier which predicts disease comorbidity.

2.1 Indexing

In order to create an index of a corpora, the corpora is first broken down into a sequence of documents. Each document is further broken down into a title, metadata and content. All document text is normalized to ensure easy processing at later stages of the system. The text is converted to lower case, and punctuation and formatting characters such as newlines, carriage returns and tabs are removed. Stop words such as 'the' and 'yet' which do not provide interesting cooccurrence information are also removed from the

text. The normalized content text is then stored as an indexed series of (word, word position) pairs associated with the title and metadata of the document in which it appears.

2.2 Counting Cooccurrence

Once an index is created from a corpora, cooccurrence data is collected from the index. The system is provided with a list of target diseases about which to collect cooccurrence data. The system is then used to read through the index and maintain counts of how often different cooccurrence pairs occur. A cooccurrence pair is defined as an appearance of two words in a text within N words of each other. The word position values are used to determine whether a word lies within the specified cooccurrence window. Note that the word position values are necessary to compute word distance, because stop words have been removed, making word distances otherwise impossible to compute accurately. Words occurring in the title of a text are automatically counted as cooccurrences with any disease occurring in the content of a text.

The cooccurrence counts are stored in a Berkeley DB. The Berkeley DB format is an efficient format for storing and retrieving data values which are associated with a given key. In this case, cooccurrence pairs are associated with a count representing how many times the cooccurrence was discovered. Once these cooccurrence counts have been stored in a Berkley DB, the system uses these counts to calculate comorbidity between the target diseases.

2.3 Similarity Calculation

The cooccurrence counts are used to calculate point-wise mutual information values (PMI's) between the cooccurrence pairs. PMI is a measure of word association [3]. The value represents the change in the probability of x appearing when y is present. That is, it represents how much the appearance of x depends upon the appearance of y. It is calculated as

$$PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{1}$$

where P(x) and P(y) are the probabilities of occurrence of the individual words, and P(x, y) is the probability of their cooccurrence.

Once PMI's have been calculated between each cooccurrence pair, they are used to measure comorbidity between diseases. Related diseases should share similar PMI values between the words with which they cooccur. That

is, for similar diseases, PMI(disease_1, word) and PMI(disease_2, word) should be similar values. Thus, PMI values between two different diseases and a shared cooccurrence were used to compute similarity scores between words. The system calculates two different values from these PMI values.

Dagan proposes a context similarity metric that measures the similarity of words based on their PMI's between a third word [2]. The metric is useful for calculating the likelihood of a cooccurrence (w1, w2) even if the cooccurrence is not actually observed in the corpus. For example, to estimate the likelihood of an unobserved cooccurrence such as 'breakfast beer', cooccuurences which do appear in the corpus such as 'light breakfast' and 'light beer' are used. Thus, the similarity of the context surrounding two words w1 and w2 may be calculated using shared a cooccurrence w. Context similarity is calculated as

$$sim(w_1, w_2, w) = \frac{min(PMI(w, w1), PMI(w, w2))}{max(PMI(w, w1), PMI(w, w2))} \tag{2}$$

As suggested by Dagan, these values are weighted by max(PMI(w, w1), PMI(w, w2)) [2]. The use of weights is nescesary because small PMI values usually indicate less important cooccurrences which are more vulnerable to noisy data.

The normal is another value proposed by Dagan used to measure word similarity given two words and a common cooccurrence [1]. It measures how related two words are given how similar the cooccurrence patterns are. The normal measures word similarity by calculating the similarity of two words' cooccurrence patterns. Words w1 and w2 are assumed to be similair if they share approximate PMI values with a shared cooccuurrence w. For example, words such as 'drink' and 'sip' might be considered similar because they have nearly approximate PMI values with a shared cooccurrence such as 'tea.' The normal is calculated as

$$normal(w_1, w_2, w) = |PMI(w, w1) - PMI(w, w2)| \tag{3}$$

Once again, these values were weighted by max(PMI(w, w1), PMI(w, w2)).

These values were calculated between all discovered cooccurrence pairs, and a weighted average of each of the two similarity metric were taken as the final measures of similarity:

$$weight(w_1, w_2, w) = max(PMI(w, w_1), PMI(w, w_2)) \tag{4}$$

$$similarity_1(w_1, w_2) = \frac{\sum_{w \in lexicon} sim(w_1, w_2, w) * weight(w_1, w_2, w)}{\sum_{w \in lexicon} weight(w_1, w_2, w)} \tag{5}$$

$$similarity_2(w_1, w_2) = \frac{\sum_{w \in lexicon} normal(w_1, w_2, w) * weight(w_1, w_2, w)}{\sum_{w \in lexicon} weight(w_1, w_2, w)} \tag{6}$$

2.4 Tuning

Several variables were tuned in the approach outlined in section 2.

ψ is the cooccurrence window. A word must be within ψ words of the disease name in either direction in order to be considered a cooccurrence.

ρ is the PMI threshold. PMI's are only calculated between cooccurrence pairs if the pair has appeared at least ρ times. For example, if ρ is 20, then the PMI between 'diabetes' and 'weight' would only be calculated if the words cooccur at least twenty times. If a cooccurrence pair occurs less than ρ times, then the cooccurrence pair is disregarded. This value restricts PMI's calculated from infrequent cooccurrences from affecting the final similarity measure.

τ is the similarity threshold. Similarity values are only calculated between diseases which share at least τ cooccurrence words. For example, if the relationship between 'diabetes' and 'obesity' is to be calculated, they must both cooccur with at least τ of the same words. This value restricts similarity values calculated from small numbers of share cooccurrences from being calculated. τ is a confidence threshold which influences the number of predictions the system will make. If two diseases do not share enough cooccurrence words, then no prediction is made regarding their comorbidity.

Improvements in accuracy across ρ and ψ are presented in figure 2. The effects of varying τ are presented in section 4.

2.5 Machine Learning

The system uses machine learning techniques to detect disease relatedness. Machine learning focuses on the induction of models from relevant features. Machine learning classifiers are provided with data which is relevant to the given classification task (for example, the similarity metrics described above) and the correct classification information (in this case, whether two diseases are significantly comorbid or not.) After being supplied with enough classification data and examples of correct classifications, the classifier produces a model which can be used for future classification tasks.

Many different classifiers are available for machine learning which create models using different strategies. Three different classifiers, Naive Bayes, J48 decision tree and SMO support vector machine, were used in the experiments

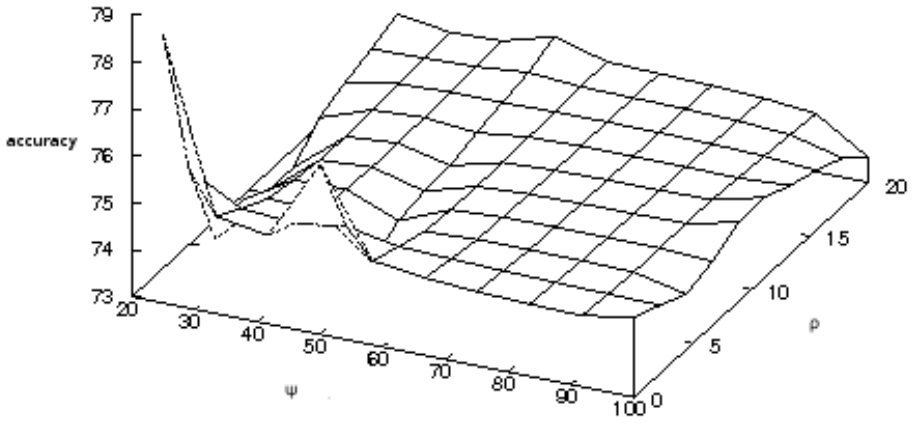

Figure 2: tuning results

presented in this paper,

Naive Bayes A naive Bayes classifier assumes that each feature is unrelated to the presence of each other feature. For example, a person may be considered a bachelor if he is both male and unmarried. Even though correct classification requires that the two variable be considered together, as both features must be present to classify a person as a bachelor, a naive Bayes classifier considers each feature to contribute independently to the probability that a given person is a bachelor.

SMO The SMO classifier is from the Support vector machine (SVM) family of classifiers. The classifiers treat input data as two sets of vectors in an n-dimensional space. An SVM will attempt to create a hyperplane which separates the vectors and maximizes the margin between the two data sets.

A good separation between the data is achieved by finding a hyperplane which divides the data sets furthest from one another. The SMO classifier essentially attempts to find boundaries between classes by exploring which features contribute clearly separate values for different classes.

J48 The J48 classifier is a decision tree classifier. A decision tree is a data model that encodes the distribution of the class label in terms of

feature attributes. Features are represented as nodes, and the possible value of a feature as an arc to a child node. A leaf in this tree represents a possible value of a feature given the value of the features along the path from the root node. Thus, the classifier learns which paths along the decision tree through various features should lead to which classification labels.

2.6 Classifier Training

In order to predict comorbidity using the calculated similarity values, the WEKA machine learning suite was used to train a classifier which detects comorbidity [4]. Four features were used to train the classifier.

Feature Name	Description
context similarity	a measure of how related two words are, given a set of shared cooccurrences; useful when dealing with sparse data; see Eq. (5)
normal similarity	a second measure of how related two words are, given a set of shared cooccurrences; see Eq. (6)
cooccurrence count	number of times the two diseases cooccurred with one another
PMI	PMI calculated between the two diseases; see Eq. (1)

Table 1: Features used to train the classifiers

The first two features measure the similarity of the cooccurrences between two diseases, while the second two features measure how frequently the diseases appear in proximity to one another.

3 Evaluation

The evaluation process used the trained classifiers to guess whether or not two diseases are significantly comorbid. The output from the classifiers was compared to actual medical data in order to test the accuracy of the trained models

3.1 Evaluation Metrics

Precision, accuracy, F-measure and recall are used to evaluate the system's performance.

Precision is a measure of how likely a classification of significant comorbidity is a correct classification.

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{7}$$

Recall is a measure of how many disease pairs were classified as significantly comorbid versus how many disease pairs should have been classified as significantly comorbid. It punishes classifiers which rarely classify diseases as significantly comorbid, whether due to the inaccuracy of a classifier or its reluctance to make predictions.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{8}$$

F-measure is a combination of recall and precision scores. It provides a good overall measure of a classifier's effectiveness.

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{9}$$

Accuracy is a measure of how many correct guesses the classifier makes. It is a measure of how accurately the classifier is able to classify both strong comorbidity and lack of comorbidity.

$$accuracy = \frac{correct\ classifications}{all\ classifications} \tag{10}$$

3.2 Corpora

Cooccurrence data was gathered from two different types of sources.

Corpus	Number of Words	Number of Posts/Documents
Medline	1840736537	15223668
Psych Forums	22199263	210022

Table 2: Corpora used to gather cooccurrence data

The first source of cooccurrence data comes from medline. Medline[1] is a bibliographic database containing more than 16 million references to journal

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

articles in life sciences with a concentration on biomedicine. The version of Medline used in the described experiments contains articles from 1949 to 2006. For older articles, generally only the abstract and other bibliographical information is available, while the full text is available for more recent articles.

The second source of cooccurrence data comes from an internet message board. `psychforums.com`[2] is a discussion board for sufferers of various mental illnesses. It is a moderately sized forum, containing more than 200,000 posts and more than 17,000 registered members.

A larger set of cooccurrence data was mined from Medline. This is due both to disparities in size and the lack of explicit use of medical disease names in the psychforums corpus in favor of acronyms, shorthands and colloquialisms.

3.3 Ground Truth

The truth data for evaluation comes from two sources. Both data sources report the Pearson Correlation values between various pairs of diseases. Pearson correlation values range from -1 to 1. Values close to 1 indicate a strongly positive relationship, while values closer to -1 indicate a strongly negative relationship. Diseases are considered to be significantly comorbid if they have a Pearson Correlation value greater than .1 [5].

The first set of truth data comes from the National Comorbidity Study Replication (NCSR) data [5]. This data provides Pearson correlation values between many different mental health diseases and symptoms. The Pearson values were calculated based on surveys of 10,000 respondents. The evaluation on this set of data was performed on twenty diseases.

The second set of truth data comes from the California Health Interview Survey (CHIS) [6] [7] [8]. The data consists of Pearson correlation values between various diseases, conditions and risk factors. The survey, carried out in 2001, 2003 and 2005, interviewed between 40 and 50 thousand adults each year. A smaller number of children and adolescents were also interviewed, but only the data from the adult interviews was used in experimentation, due to its sample size being three to nine times larger. The experiments described in this paper use only the most recent surveys from 2005. The evaluation on this set of data was performed on fifteen diseases.

[2] http://www.psychforums.com

4 Results

The following results measure the accuracy of the trained classifiers against the ground truth data. Precision, recall and F-measure scores are provided for the positive correlation field, and accuracy scores are provided which combine the positive and negative fields. The ψ, ρ, and τ values represent the thresholds used to create the classifiers in the following tables. The baseline scores are the results of using a classifier which always guesses the most frequently occurring class.

Cross-validation means is the process of using the truth data on which the classifier was trained to measure its effectiveness. That is, a classifier trained on the NCSR truth data would in turn be asked to make predictions on the same data upon which it was trained.

4.1 Medline Corpus Cross-Validation

The following results were produced by classifiers trained on features from medline and truth data from the NCSR study.

J48 Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
25	20	4000	.815	.239	.370	80.488
25	20	1000	.772	.663	.713	74.757
25	20	500	.765	.565	.650	69.725
25	20	0	.761	.864	.809	73.387
baseline			.645	1	.784	64.458

Table 3: J48 classifier trained on Medline corpus and cross validated on NCSR data

SMO Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
25	20	4000	.792	.235	.362	73.171
25	20	1000	.730	.802	.764	72.812
25	20	500	.648	.840	.732	65.138
25	20	0	.655	.963	.780	64.516
baseline			.645	1	.784	64.458

Table 4: SMO classifier trained on Medline corpus and cross validated on NCSR data

Naive Bayes Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
25	20	4000	.882	.185	.306	70.732
25	20	1000	.919	.420	.577	63.107
25	20	500	.850	.209	.337	49.541
25	20	0	.833	.370	.513	54.032
baseline			.645	1	.784	64.458

Table 5: Naive Bayes classifier trained on Medline corpus and cross validated on NCSR data

The two sets of truth data were not equally useful in classifier training. The results from cross-validation on CHIS data are not reported, because they are essentially baseline in performance. The data from the CHIS did not provide enough examples of significant comorbidity to train an effective classifier. Because the survey is a general survey of various health issues and most diseases are not significantly comorbid this is not surprising. The data from the NCSR, however, provided enough examples of comorbidity to train effective classifiers.

4.2 Psych Forums Corpus Cross-Validation

The following results were produced by classifiers trained on features from psychforums and truth data from the NCSR study.

Naive Bayes Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
100	0	1000	.923	.112	.200	92.857
100	0	0	.846	.103	.184	73.684
baseline			.629	1	.722	62.857

Table 6: Naive Bayes classifier trained on psychforums corpus and cross validated on NCSR data

The success of the Naive Bayes classifier seems exaggerated due to the smaller number of disease pairs judged in this task. Due to factors discussed more thoroughly in section 5, the results presented above are based on one quarter of the number of classifications made in the cross validation from Medline data. In this task, most of the classifiers discover models that choose the positive class with high frequency. The Naive Bayes classifier chooses the negative class only slightly more accurately than the other classifiers, but

SMO Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
100	0	1000	.857	.112	.198	85.714
100	0	0	.737	.131	.222	73.684
baseline			.629	1	.722	62.857

Table 7: SMO classifier trained on psychforums corpus and cross validated on NCSR data

J48 Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
100	0	1000	.857	.112	.198	85.714
100	0	0	.846	.103	.184	73.684
baseline			.629	1	.722	62.857

Table 8: J48 trained on psychforums corpus and cross validated on NCSR data

even moderate success at predicting the negative class is greatly rewarded in this task, due to the smaller number of classifications made.

4.3 Classifiers Trained On NCSR Truth Data and Validated on CHIS 2005 Truth Data

The following results were produced by classifiers trained on features from both corpora and truth data from the NCSR study, and validated on CHIS truth data.

Naive Bayes Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
25	20	4000	.254	1.000	.405	31.250
25	20	1000	.245	.800	.375	37.500
25	20	500	.370	.667	.476	65.625
25	20	0	.357	.667	.465	64.062
baseline			0.00	0.00	0.00	76.563

Table 9: Naive Bayes classifier trained on Medline corpus and NCSR truth data, validated on CHIS 2005 data

The classifiers generated by the J48 and SMO classifiers when trained on Medline almost always guessed no correlation on the CHIS data, resulting in

Naive Bayes Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
100	0	1000	.210	.867	.338	20.313
100	0	0	.257	.600	.360	50.000
baseline			0.00	0.00	0.00	76.563

Table 10: Naive Bayes classifier trained on psychforums corpus and NCSR truth data, validated on CHIS 2005 data

J48 Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
100	0	1000	.538	.467	.500	78.125
100	0	0	.273	1.00	.429	37.500
baseline			0.00	0.00	0.00	76.563

Table 11: J48 classifier trained on psychforums corpus and NCSR truth data, validated on CHIS 2005 data

SMO Classifier						
ψ	ρ	τ	precision	recall	F-measure	accuracy
100	0	1000	.234	1.00	.380	23.438
100	0	0	.234	1.00	.380	23.438
baseline			0.00	0.00	0.00	76.563

Table 12: SMO classifier trained on psychforums corpus and NCSR truth data, validated on CHIS 2005 data

near baseline behavior. Thus, only the results from J48 and SMO classifiers trained using psychforums cooccurrence data are presented in the CHIS classification task. When trained using psychforums cooccurrences, the J48 classifier in particular was much more successful at classifying CHIS diseases than when trained using medline cooccurrences.

5 Discussion

The relationship between precision, recall and τ is mostly unsurprising. The cross-validation tasks consistently show that as the confidence value *gammma* increases, thus restricting the number of classification made to only those which the classifier is most confident of, precision increases while recall decreases. The classifier is correct with greater frequency at higher

confidence values, but when forced to make more classifications at lower confidence values the classifier makes a higher number of correct classifications.

This relationship holds less true on the CHIS classification tasks. The results from the Naive Bayes classifier in particular seem to have a relationship opposite to the one described in the cross validation task. The figures seem to indicate that classifiers trained at lower τ values guess positive correlation with a high frequency. This indicates that when constrained to high τ values, the classifiers see a disproportionate number of positive correlations. This indicates that the methods proposed in this paper are more successful at identifying positive correlation than negative, because these are the values about which the classifier is most confident.

The results from the classifiers trained on web forum cooccurrence data are somewhat problematic because they judge many fewer disease pairs. In gathering cooccurrence data, words were counted as cooccurrences if they occurred in proximity to an explicit disease term, such as 'agoraphobia' or 'conduct disorder.' Because there are many fewer explicit uses of the explicit disease names in web forum discussions, only one quarter of the cooccurrence disease pairs found in the Medline corpus are found in the psychforums corpus.

Interestingly, the classifiers trained on the psychforums corpus all were biased towards predicting positive correlation. More than seventy percent of the disease pairs discovered in the forum data are significantly comorbid, which indicates that a failure to find a disease pair in this corpus should itself indicate a level of negative correlation. However, even given the much smaller amount of disease pairs for training, the classifiers trained on the psychforums corpus performed much better than those trained on Medline in the CHIS classification task.

6 Conclusion

Overall, the system presented is useful at finding evidence that two diseases are related, but has trouble deducing that two diseases are unrelated. Rather than using lack of evidence for significant comorbidity as an indicator of negative correlation, an improved system would utilize features specifically designed to detect negative correlation.

On each set of tasks, at least one of the classifiers was able to perform at above baseline levels. This seems to affirm the stated hypothesis of this paper, that frequency of comorbidity between diseases should be reflected by linguistic similarity. Given the success of the classifiers trained on the

psychforums corpora, this paper further indicates that the internet could be a beneficial tool in the public health sphere. Although the experiments described in this paper are concerned with affirming already known disease relations, future systems have the potential to chart new disease relations which the medical community is unaware of.

In general, this paper has shown that mining data from textual sources, including web forums, can produce relevant and accurate medical information. The use of web forums is particularly exciting, because it contains information produced by the sufferers of diseases themselves. By using more advanced techniques, it might be possible to use internet sources to chart disease distributions and trends across populations, monitor breaking epidemics, or produce new bodies of medical knowledge.

References

[1] Lillian Lee Ido Dagan and Fernando C. N. Pereira. Similarity-based models of word cooccurence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.

[2] Shaul Marcus Ido Dagan and Shaul Markovitch. Contextual word similarity and estimation from sparse data. *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 164–171, 1993.

[3] K.W. Church and P. Hanks. Word association norms, mutual informations, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[4] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[5] Ronald C. Kessler Margarita Alegria, James S. Jackson and David Takeuch. Collaborative psychiatric epidemiology surveys (cpes), 2001 - 2003.

[6] California Health Interview Survey. CHIS 2001 Adult Public Use File, Release 4 [computer readable file], 2008.

[7] California Health Interview Survey. CHIS 2003 Adult Public Use File, Release 2 [computer file], October 2005.

[8] California Health Interview Survey. CHIS 2005 Adult Survey, February 2008.