Automated Extraction and Classification of Drug-Drug Interactions from Text

Tasnia Tahsin
Department of Biomedical Informatics
Arizona State University
13212 East Shea Boulevard
Scottsdale, AZ 85289
ttahsin@asu.edu

Ehsan Emadzadeh
Department of Biomedical Informatics
Arizona State University
13212 East Shea Boulevard
Scottsdale, AZ 85289
eemadzad@asu.edu

Graciela Gonzalez
Department of Biomedical Informatics
Arizona State University
13212 East Shea Boulevard
Scottsdale, AZ 85289
graciela.gonzalez@asu.edu

ABSTRACT

Drug-drug interactions (DDIs) account for 3 to 5% of all in-hospital medication errors, and are also an important cause of patient visits to emergency departments. In order to accelerate the discovery of DDIs, it is essential to be able to build upon already published interactions. However, the time taken to manually review all information available on this subject severely limits its practical application. Automated extraction of drug-drug interaction from text can therefore prove to be of great importance for biomedical researchers. The SemEval 2013 DDI extraction task was organized recently to further promote research in this area. Participating teams were asked to develop systems for automated classification of drug pair mentions in text into one of the following categories: no interaction, advice, effect, mechanism and generic interaction. In this paper we describe a data mining approach for implementing this classification task. Using a model built from the Weka SMO classifier, we were able to obtain an f-measure of 0.84 on five-fold cross validation of our training dataset.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining.*

I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation, feature evaluation and selection.*

I.5.4 [**Pattern Recognition**]: Applications – *text processing.*

General Terms

Algorithms, Performance, Design.

Keywords

Drug-drug interaction; text mining; relationship extraction.

KDD-DMH'13, August 11, 2013, Chicago, Illinois, USA.
Copyright © 2013 ACM 978-1-4503-2174-7/13/08...$15.00.

1. INTRODUCTION

Adverse drug reaction (ADR) is one of the leading causes of mortality and morbidity in the United States, accounting for over 700,000 emergency department visits and 120,000 hospitalizations every year. It is estimated to impose an additional annual cost of $3.5 billion on the healthcare system and is a significant public health problem that can potentially be avoided. One of the primary causes of ADRs is unforeseen drug-drug interactions in which the effect of one drug is modified by the presence of another. In a system analysis of ADRs, drug-drug interactions were found to be responsible for 3%-5% of all in-hospital medication errors and as new drugs continue to be introduced into the market, this figure is likely to increase [11,12,14].

The task of discovering novel drug-drug interactions has therefore attracted a significant amount of attention within the biomedical research community, leading to the identification of a large number of potential DDIs. However, the high volume of documents containing information on this topic makes manual extraction of DDI-related information a slow and cumbersome process. An effective mechanism for automated extraction of DDI from text is therefore needed to make this process more efficient.

The SemEval 2013 DDI extraction task was organized with the purpose of stimulating research in automated DDI extraction from biomedical text. A large annotated training corpus containing biomedical documents from DrugBank and Medline on the subject of DDI was provided. The corpus annotated mentions four different types of drug-related entities and identified four different types of interactions between them. Task participants were asked to build a system using this corpus that can extract all possible pairs of interacting drugs and classify the type of each DDI within a sentence given the drug entity annotations [16]. By applying data mining techniques on the training corpus provided for this challenge, we developed support vector machine, ensemble, and probabilistic models which demonstrated promising results for implementing this task.

2. RELATED WORK

Extraction of drug-drug interaction from text is a relatively new research area which has recently spurred a high level of interest within the text mining field. Most of the work in this area was performed as a result of the 2011 DDI extraction task. This task required each competing team to design a system for sentence level DDI extraction i.e. for every pair of drugs in each sentence within the corpus provided, the system had to determine whether or not an interaction was being described. 10 different teams participated in the challenge, each applying a different approach to the problem. The best performing system used a majority voting scheme with two kernels (all paths graph and shallow linguistic) and a case based reasoning classifier. It had a precision, recall and f-measure of 0.6054, 0.7192 and 0.6574 respectively on the test dataset. Using ten-fold cross validation on the training set it had a precision, recall and f-measure of 0.59, 0.63 and 0.606 respectively. Other kernels found to perform well in this task include mildly extended dependency tree (MEDT) kernel, phrase structure tree (PST) kernel and global context kernel. Feature based methods were also used by several systems, often along with kernel based methods. Classifiers used for this task include SVM, RandomForest, AdaBoost, NaiveBayes and RLS [1, 3, 4, 13, 15, 18].

The SemEval 2013 DDI Extraction task required participants to not only identify mentions of drug interactions in text, but also classify the type of each interaction. A test set was provided to the participating teams for system evaluation. It consisted of 161 Drugbank documents and 34 Medline abstracts. Systems were ranked based on their macro-averaged f-measure which was computed by first calculating precision and recall for each type of interaction and then taking an average of them all. The highest ranked system had an f-measure, recall, and precision of 0.648, .685, and .615 for the test set, respectively [5].

3. METHODS

3.1 Description of the Training Corpus

The training corpus provided by the organizers of the SemEval 2013 DDI Extraction task contained 571 documents from the Drugbank database and 142 Medline abstracts on the subject of DDIs, all annotated in xml format. Each xml document was annotated with the following information:

- the ID of the document
- the text and ID of each sentence in the document
- the ID, type and name of each entity mentioned in each sentence in the document
- the ID and type of each pair of entities mentioned within each sentence in the document along with the IDs of the two entities involved in the pair.

Tables 1 and 2 describe the different types of drug-related entities and entity pairs annotated in the documents. Further details about the corpus can be found in [16].

3.2 Feature Extraction

The first and most crucial step in our project involved the selection and extraction of relevant features from sentences within the training corpus for each drug entity pair mentioned in them. We selected nine different types of features for our system and the novelty of our approach lies in the specific set of attributes chosen. The following is a list of the features extracted, the last one being unique to our system to the best of our knowledge:

- Type of Entity 1 (E1) and Entity 2 (E2)
- Number of tokens between E1 and E2
- List of tokens between E1 and E2
- Three windows of tokens before and after E1
- Three windows of tokens before and after E2
- Parts of speech (POS) tags between E1 and E2
- Tokens directly linked to E1 in Stanford typed dependency representation
- Tokens directly linked to E2 in Stanford typed dependency representation
- Tokens in the Djikstra Shortest Path between E1 and E2 in graphical representation of Stanford dependencies
- Semantic Similarity between E1 and E2

The type of E1 and E2 was extracted directly from the xml annotations. To compute the semantic similarity between the two entities we used the freely available package SemanticVectors [17]. Stanford parser [5] was used for tokenizing, POS tagging, and typed dependency generation of the sentences. The Djikstra Shortest Path between E1 and E2 in graphical representation of Stanford dependencies was found using the JGraphT [8] package. The absence or presence of each token was represented using binary attributes. This feature extraction process produced 15,364 distinct features.

Table 1. Description of each type of entity annotated in the training corpus

Entity Type	Description
Drug	Any chemical agent used in the treatment, cure, prevention or diagnosis of diseases which have been approved for human use
Brand	Any drug that was first developed by a pharmaceutical company
Group	Any term in text designating a chemical or pharmacologic relationship between a group of drugs
Non-Human	Any chemical agent that affects living organisms but has not been approved for use in humans for medicinal purpose

Table 2. Description of each type of entity pair annotated in the training corpus

Entity Pair Type	Description
No interaction	No interaction is described between the two entities
Advice	An advice regarding the concomitant use of the two entities is described
Effect	An effect of an interaction between the two entities is described
Mechanism	The pharmacokinetic mechanism of an interaction between the two entities is described
Int	An interaction is described between the two entities without specific details

3.3 Pre-processing

The initial dataset generated from the training corpus contained 21,881 instances, 83% of which belonged to class type ‘no interaction’. To facilitate the development of a well-performing model, we improved the class distribution of the training set through selective down-sampling. Since instances next to each other in the training set were most likely to hold class and feature information about pairs of entities from the same document, and often the same sentence, it was assumed that less useful information would be lost if a subsample was generated by sequentially going through the dataset instead of taking a random subsample. Sequential sampling was thus performed to select every 5th entry of class type ‘no interaction’. The final training set contained 7300 instances with a more balanced class distribution as shown in Figure 1.

3.4 Model Building in Weka

We used Weka [7] as our data mining tool for this classification task. Three different classifiers were applied on the training dataset with various parameters to find the best performing model. The classifiers used include SMO classifier [10], Naïve Bayes [9] and RandomForest [2]. All three of them have been previously found to perform well in relationship extraction tasks. We applied the linear kernel for the SMO classifier with c values of 0.1, 1 and 10. Best performance was achieved with c value of 1. For RandomForest, we varied the number of trees from 10 to 30 in incremental steps of 5 but used the default value of 14 for the number of random features considered at every split when constructing these trees. Best performance was achieved using 30 trees. Each of the classifiers was tested using 5-fold cross validation on our training dataset. We were unable to use the test dataset provided for participants of the SemEval 2013 DDI extraction task as it was not publicly available at the time of the experiment.

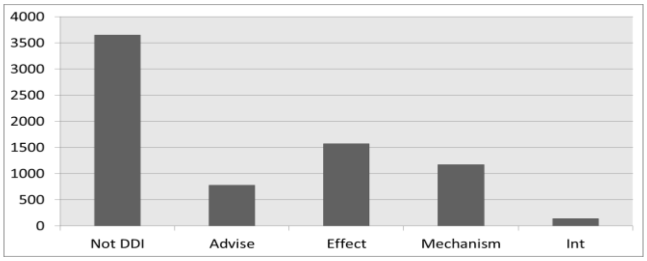

Figure 1. Class Distribution in the Final Training Set.

4. RESULTS

The best performing model was built using the SMO classifier. Table 4 shows the performance metrics for this classifier using 5-fold cross validation. The model performed best for classification of ‘no interactions’ and worst for those of type ‘mechanism’. The overall f-measure was 0.84. Time taken to build the model was 192.15 seconds.

Table 3. Five-fold cross validation performance metrics of SMO model with linear kernel and c=1.0

Class	Precision	Recall	F-measure
No interaction	.88	.87	.88
Advice	.80	.83	.82
Effect	.81	.83	.82
Mechanism	.78	.76	.77
Int	.78	.81	.80
Weighted Average	**.84**	**.84**	**.84**

Table 4. Five-fold cross validation performance metrics of RandomForest model with 30 trees and 14 random features

Class	Precision	Recall	F-measure
No interaction	.83	.92	.87
Advice	.83	.74	.78
Effect	.82	.76	.79
Mechanism	.81	.70	.75
Int	.83	.66	.74
Weighted Average	**.82**	**.82**	**.82**

Table 5. Five-fold cross validation performance metrics of NaiveBayes

Class	Precision	Recall	F-measure
No interaction	.79	.78	.79
Advice	.78	.66	.71
Effect	.75	.75	.75
Mechanism	.62	.79	.69
Int	0	0	0
Weighted Average	**.74**	**.75**	**.74**

Although RandomForest produced comparable results with an overall f-measure of 0.82, it took 397.41 seconds to build the model. Table 2 shows the performance metrics for this classifier. The f-measure for each of the classes was lower for RandomForest than for the SMO classifier. For class type ‘no interaction’, RandomForest had higher recall than SMO but lower precision. For all other classes, it had higher precision but lower recall.

NaïveBayes had the worst performance among the three with an overall f-measure of 0.74 but it was the fastest classifier, taking only 18.15 seconds to build the model. It had a precision and recall of 0 for class type ‘int’.

5. DISCUSSION

As expected, all three classifiers were best at classifying ‘no interacton’ – the class with the highest number of instances in the training set. The SMO and RandomForest classifier had high f-measure values for all of the other classes as well, despite their relatively low frequency. However, with NaiveBayes, interactions of type ‘int’ were not classfied at all during cross-validation since it had the smallest representive sample in the training dataset. Using a training set with a more balanced class distribution may help improve its performance. Although SMO proved to be the best performer, overall, for this classification task it is possible that with a greater number of trees, the performance of RandomForest could be further increased.

6. CONCLUSION

The results from the performed cross-validations suggest that our chosen set of features can be used to classify drug pair mentions in text with reasonably high precision and recall given an adequately large dataset to train on. SMO classifier produced the highest performance level with an overall f-measure of 0.84. Some of the limitations of our applied methods include under-sampling of the majority class in our training dataset and the use of cross-validation rather than a held-out test set for performance evaluation. Future work will involve evaluation of our system on the test set provided for the SemEval 2013 DDI extraction task, selection and evaluation of different sets of features using our training dataset to find the optimal feature set for this task and evaluation of the feature set described in this paper for other relationship extraction tasks in the biomedical domain such as extraction of protein-protein interactions.

7. ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number P30AG019610. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8. REFERENCES

[1] Björne, Jari, et al. "Drug-drug interaction extraction from biomedical texts with svm and rls classifiers." *Proceedings of DDIExtraction-2011 challenge task*(2011): 35-42.

[2] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[3] Chowdhury, Faisal Mahbub, et al. "Two different machine learning techniques for drug-drug interaction extraction." *Challenge Task on Drug-Drug Interaction Extraction* (2011): 19-26.

[4] Chowdhury, Faisal Mahbub, and Alberto Lavelli. "Drug-drug interaction extraction using composite kernels." *Challenge Task on Drug-Drug Interaction Extraction* (2011): 27-33.

[5] De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. "Generating typed dependency parses from phrase structure parses." *Proceedings of LREC*. Vol. 6. 2006.

[6] "Evaluation". Extraction of Drug-Drug Interaction from Biomedical Texts. web 15 May 2013. http://www.cs.york.ac.uk/semeval-2013/task9/index.php?id=evaluation.

[7] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD Explorations Newsletter* 11.1 (2009): 10-18.

[8] JGraphT. web 10 March 2013. http://jgrapht.org/.

[9] John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995.

[10] Keerthi, S. Sathiya, et al. "Improvements to Platt's SMO algorithm for SVM classifier design." *Neural Computation* 13.3 (2001): 637-649.

[11] Magro, Lara, Ugo Moretti, and Roberto Leone. "Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions." *Expert opinion on drug safety* 11.1 (2012): 83-94.

[12] "Medication Safety Basics", CDC. web 26 Apr 2013. http://www.cdc.gov/medicationsafety/basics.html.

[13] Minard, Anne-Lyse, et al. "Feature Selection for Drug-Drug Interaction Detection Using Machine-Learning Based Approaches." *Challenge Task on Drug-Drug Interaction Extraction* (2011): 43-50.

[14] Mirošević Skvrce, Nikica, et al. "Adverse drug reactions caused by drug-drug interactions reported to Croatian Agency for Medicinal Products and Medical Devices: a retrospective observational study." *Croatian medical journal* 52.5 (2011): 604-614.

[15] Segura-Bedmar, Isabel, Paloma Martınez, and Daniel Sánchez-Cisneros. "The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts." *Challenge Task on Drug-Drug Interaction Extraction* 2011 (2011): 1-9.

[16] Segura-Bedmar, I., Martínez, P, Herrero-Zazo, M. "SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts". *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013).* 2013.

[17] SemanticVectors. web 10 March 2013. https://code.google.com/p/semanticvectors/

[18] Thomas, Philippe, et al. "Relation extraction for drug-drug interactions using ensemble learning." *Training* 4.2,402 (2011): 21-425.