Is Drug Drug Interaction a Pure NLP Task?

Chinmay Swami[†] **Jungyeul Park**[‡]
[†]Department of Computer Science and Engineering
[‡]Department of Linguistics
The State University of New York at Buffalo
{chinmayp,jungyeul}@buffalo.edu

Abstract

Drugs play an important role in treating diseases ranging from those that are innocuous to those that are extremely noxious.However, administering the right drug is pivotal to the revitalization of the patients health. Administering the wrong combination of drugs has ramifications that can be harmless or can also be life threatening. Such types of adverse effects is found to be 8th leading cause of death is the US (Goldstein et al., 2005). Hence we will be exploring options to detect whether, given a pair of drugs, would there be any interaction among them and if yes what type of interaction would that be. In this paper we will be leveraging machine learning techniques to create a model which will learn from the sentences that describe various kinds of drug drug interactions and would predict the type of interaction that would happen between the two drugs when administered together.Also along the way we will be exploring whether we can rely solely on NLP techniques for the task of predicting drug-drug interaction with the help of multiple experiments. We will also be leveraging domain knowledge, for example the information pertaining to molecular structure (Asada et al., 2018) of the drug to enhance the predicting capabilities of the model.

1 Introduction

The definition of drug-drug interaction (DDI) is broadly described as a change in the effects of one drug by the presence of another drug (e.g. Acetazolamide reduces urinary excretion of quinidine and may enhance its effect) (Segura-Bedmar et al., 2014). Drug-Drug Interaction is a task that involves predicting how the pair of drugs when administered together would react with each other and whether that reaction if present would cause any harm to the patient and also the type of interaction that happens.However, identifying whether a pair of drugs may have adverse effect on one's health or not completely depends upon the data set you are using and also on many other external factors which can be used to augment the performance of a model which purely rely on the data set given.

In 2013 SemEval concocted a task called "Extraction of Drug-Drug Interactions from BioMedical Texts" where they provided a Drug-Drug Interaction (DDI) corpus (Herrero-Zazo et al., 2013; Isabel Segura Bedmar, 2013) which consisted of XML files of drugs from DrugBank database and Medline abstracts on Drug-Drug interactions.The data set consists of XML files for each drugs.The training set consists of xml files which are created from two different sources namely DrugBank (Wishart, 2006) and MedLine. Now, one can cogently say that what if there is one such drug that hasn't been seen by the model during training or what if there is some drug-drug pair that the model didn't encounter during training but at the time of test or in production environment came across it or it can be the case that when the model was trained and developed a particular drug didn't exist but later on came into market after passing all the required approvals. In these cases the model is bound to fail if it purely relies on the data sets provided. These were the questions faced by us during our implementation and motivated us to wonder whether Drug-Drug interaction was purely a NLP task or whether it relied heavily on getting expert knowledge from the field of medicine. For exploring this we carried out various experiments that involved a variety of feature set designed to exploit the linguistics features made available by the DDI corpus. One such example of this is a paper published on the same task which used a combination of convolutional neural network and Graph convolutional network. The GCNs were used to predict interaction between two drugs us-

No of	DrugBank	MedLine
Drug Files	573	142
Sentences	5675	1301
Entities	12929	1836
Average Entities	2.2782	1.4112
Pairs	26005	1787
Average Pairs	4.5823	1.3735

Table 1: Statistics of data from DrugBank and MedLine

Instances	Number
True	4020
False	23772

Table 2: Imbalance pertaining to true and false instances in dataset.

ing their molecular structures.Hence to explore this facet of this task we created multiple models which were based on different techniques used to represent the data about the drug drug pairs and their interactions which will be discussed in coming sections (Asada et al., 2018).

2 A Glimpse into Data

The data set consisted of separate XML files for each drugs. The XML files were generated based on drug-drug interactions documented in the DrugBank database (Wishart, 2006) and in Medline abstracts which were based on the drug drug interactions.The information obtained from these files can be distinguished with the help of the document ID.Table 1, Table 2 and Table 3 provides a glimpse into the statistics about the data set. Figure 1 shows a part of XML file which is present in the MedLine dataset[Training] in DDI corpus.

The XML file contains `<document>` tag which has an attribute called ID which can be used to uniquely identify all the attributes of the drugs from this particular document.The document tag contains one or more sentence tag which has an attribute id used to uniquely identify the sentence in the corpus and an attribute called text which has the actual sentence itself which may or may not have any interaction information.The sentence tag is made up of entity tag and pair tag.

- **`<entity>`** Entity tag contained all the words from the sentence that are identified as drug names along with some extra details.

Instances	Number
False	1463
Mechanism	561
Int	77
Effect	264
Advice	198

Table 3: Data set stats for which we found SMILE files.

All the identified drug names were given ID's which were constructed from extending the sentence ID. The entity tag also consists of an attribute called as type which indicates the type of the drug for example whether the identified drug belongs to *group*[Term used to describe different drugs with similarities with each other.] or *brand*[manufacturer of drug] etc. This information played a vital role in acquiring SMILES files.

- **`<pair>`**Pair tag consists of all possible combinations of the drug names identified and present in the entity tag.Pair tag contains an attribute called DDI which indicates the presence of an interaction between the two drugs and if DDI is *true* then it also consists of an extra attribute called type which indicates the type of interaction(Mechanism, Int, Effect, Advice).

Now as discussed earlier we will also be incorporating some domain information and for that we decided to go forward with making use of SMILES files which is one of many kinds of files used to hold molecule's structural information. We extracted the SMILES files from drugbank dataset however, drug bank didn't have smile files for all the drugs that were present in DDI corpus hence we only used the drug pairs from training and test for which we found smile files to assess the performance. The statistics about this dataset is shown in Table 3. One of the issues with the dataset was the extremely low number of *true* instances compared to *false* instances as shown in table 2. Hence activities pertaining to balancing the dataset were performed which is discussed in next sections.

3 Preprocessing

As per the feature sets decided[which will be explained in further sections] a great deal of pre-

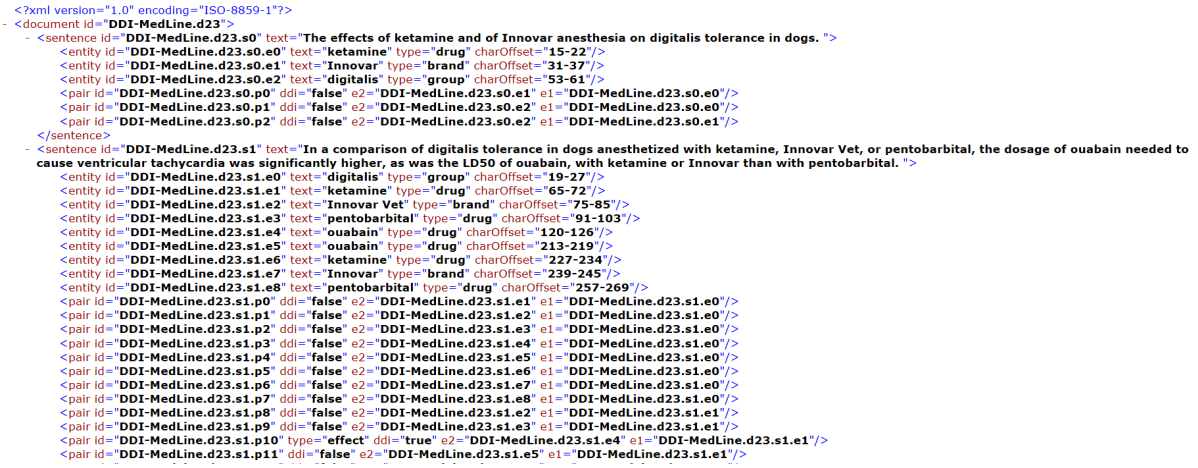

Figure 1: Small snippet of XML file from MedLine dataset.

processing activities were required to be performed. To extract the drug names from the sentences present, the sentences were tokenized which also separated any punctuation's concatenated to the names and also since we needed to replace the drug names present in the sentences we performed tokenization of drug names as well so that uniformity would be maintained. However, there were many inconsistencies pertaining to the way drug names were annotated. For example, some sentences had *alpha-blockers*, some had *alpha-blocker*, some had *alpha blockers* while some had *alpha blocker*. But, they all referred to the same drug. Hence we performed selective lemmatization where lemmatization of only the drug names was done by using a very simple method involving computing the similarity between pair of drugs.The similarity threshold was decided by trial and error technique on multiple threshold values and the value of 0.88 was used as a threshold in final version. Also replacing drug names in the sentences by a universal ID solved many of the processing issues. The pair of drugs were assigned same universal ID which were deemed as same by the above technique.

As from table 2 it's clear that number of false instances were four times more than number of true instances. Training a model on such a type of dataset gives rise to a very biased model so we decided to experiment with the unbalanced dataset and also we balanced the dataset. To balance the dataset we used two approaches. One was replicating the true records multiple times so that the number of false and true instances are almost the same and the second was removing the false records to match the number of true and false instances.The removal of false records were done randomly instead of manually selecting instances to be discarded.

Now to generate the SMILES files, the drugbank structural file available on their website was used which consisted of molecule's structural information. This information was saved into a SMILES file for teach of the drugs. These files were passed to PaDEL descriptor which generated a csv file with vectors pertaining to 1D and 2D information of drug and also another csv file with vectors pertaining to fingerprints of drugs.

Based on table 3 it can be inferred that there is an imbalance in the number of instances for each types in the dataset that consisted of PaDEL descriptor (Yap, 2011). For example, number of instances for *Int* is extremely low as compared to other 3 types associated with true interaction. Hence to balance the dataset the instances were replicated a number of times. However, for *Model 1* balancing was not required since the number of *True* and *False* instances were balanced which can be inferred from Table 3.

4 System Design

The system developed is shown in Figure 2. We first developed various feature sets[will be discussed in next section] and created a SVM *Model 1* that only predicted whether a interaction between a pair of drugs existed or not just to evaluate which of the features set were efficacious. Based on the predictions of *Model 1* all the drug pairs which had *True* predictions were passed to *Model 2* which performed multi class classifica-

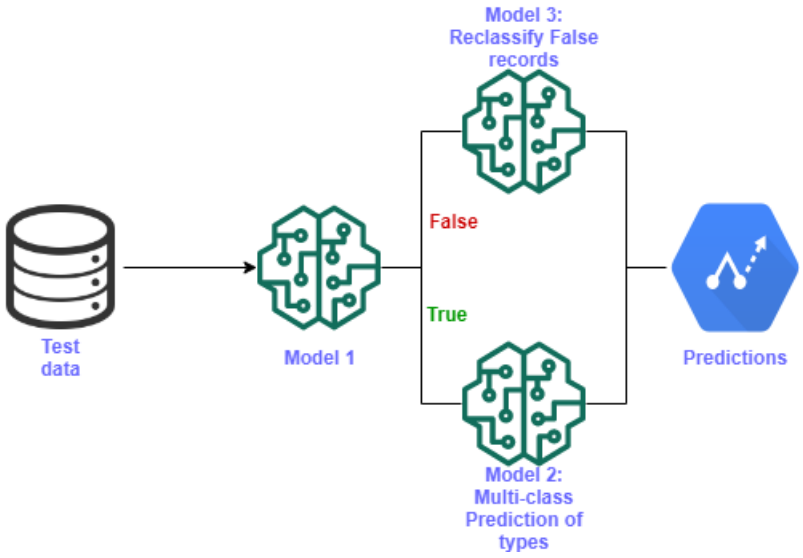

Figure 2: Prediction process using cascaded SVM models

tion. Also we created another model *Model 3* which was capable of multi-class predictions including *False* class apart from the 4 types for true cases was passed drug pairs that were categorized as *False* to re-categorize in-case they were miss-categorized by *Model 1*.The predictions from *Model 2* and *Model 3* were combined to achieve final predictions. The features that performed well with *Model 1* were only used for further classification using *Model 2* and *Model 3*. This system consists of 3 models:

Model 1 This model was trained only on the DDI attribute present in the `<pair>` tag. Hence, the model just lets us know whether an interaction between a pair of drugs exists or not[True or False].

Model 2 This model was trained on the Type attribute present in the `<pair>` tag. Hence, the model was trained only on instances that had DDI as *True* and predicted whether the type of interaction was Int, Mechanism,Advice or Effect.

Model 3 When creating the training set care was taken that whenever the DDI was *False* the Type of that pair was set to *False*. Hence when we trained for Model 3 we used the entire training set of DDI corpus and used Type attribute to get the labels. This model predicted the 4 types as well as was capable of identifying *False* instances. In other words its predictions consisted of any one of the 5 labels.

5 Methodologies

Following were the feature sets used.

- Using only the drug-drug pair and representing them in the form of Word2vectors.
- Using 1-Hot encoding of Drug-Drug pair along with 1 hot encoding of the neighbouring words of the drugs from the sentence.
- Using 1 hot encoding of the neighbouring words of the drugs from the sentence[Drug Pair encoding was not used].
- Using only the drug-drug pair and representing them in the form of 1 hot encoding.
- Using only the drug-drug pair and representing them in the form of 1 hot encoding and chemical descriptors generated by PaDEL (Yap, 2011)

We will now look at these methodologies one by one and discuss the results achieved at the end.[90/10 split was used to get the training and development set]

5.1 Drug-drug pair with word2vec representation

Word2vec representations were generated using fastText (Bojanowski et al., 2016) and these vector representations were used to train the Model. This model was trained on unbalanced data only.

5.2 Using neighbouring words of the drugs along with drug pair encoding

In this method surrounding 3 words except for stop words for each drug were extracted from the sentences and encoded using normalized 1 hot encoding technique. These encoding were appended by normalized 1 hot encoding of drug pairs. There were 3 variations of this, model 1 for unbalanced, 1 for each of the 2 balancing techniques described in section 3.

5.3 Using neighbouring words of the drugs

In this method only the surrounding 3 words from the sentences except for stop words for each drug were extracted and encoded using normalized 1 hot encoding technique.There were 3 variations of this as described above.

5.4 Drug-drug pair with one hot encoding representation

In this method normalized 1 hot encoding technique was used to encode drug pairs and trained the model on it. There were 3 variations of this as described above.

5.5 Augmenting neighbouring words feature set with chemical descriptors

This is the method where we incorporated the domain information in order to make the model better at prediction. We used a software called PaDEL Descriptor (Yap, 2011) which is used to calculate molecular descriptors and fingerprints. It provided us with 1D and 2D , fingerprints information encoded in vectors. Files or directory containing files that contain the molecule's structural information is passed to PaDEL. Some of the most common files are MDL mol, SMILES[We used SMILES in our experiment].We used these vectors in following ways:

5.5.1 Only PaDEL generated vectors

In this method we only used the 1D and 2D vectors or fingerprints as features and trained the model on it. Since the dataset consisted of only those instances for which we found smile files in the drugbank dataset, the instances were unbalanced [as shown by table 3] and hence balanced the dataset again by using replication.

5.5.2 Pairing PaDEL generated vectors along with textual features

In this method we only used the 1D and 2D vectors or fingerprints along with the textual feature 5.3 to train and test the model. Feature set 5.3 was decided to be used only after we compared the results obtained from feature sets 5.1, 5.2, 5.3 and 5.4. The dataset was balanced the same way as done before in 5.5.1. So we had 1 feature set that had 1D and 2D descriptors and another one which had fingerprint vectors. Each was used to train a separate model.

5.6 Results

`conlleval.pl` was used to calculate the F1 Measure. Firstly all the methods performance were evaluated based on the performance on development set. Based on it, it was found that method 2 and method 3 trained on balanced dataset by reducing the number of false instances performed better. These methods were used for further classification using *Model 3* as shown in Table 5. Also to explore whether incorporation of domain knowledge increases the efficacy, feature set 5.5 was used to train model and the prediction results are discussed in Table 4.

From table 4 we can see that there were some methods that performed better than the ones we went ahead with for *Model 3* However, they performed very bad for *True* predictions and extremely well for *False* instances.

Now after observing table 4 we can see that the methods 5.5.1 and 5.5.2 performed better.However, its not the case. They performed exceptionally well in prediction of *False* cases but were extremely poor with further categorization into 4 types. This was worst than methods that relied on just textual methods. However from Table 3 its clear that we had very few instances that could be used to train the model. For other methods based on removed false balancing techniques we had 8040 instances to train the model however for method 5.5.1 and 5.5.2 we only had 1801 instanced, which was significantly low. But from table 3 we can infer that there were 1463 *False* instances and 1100 *True* instances which was quite balanced and upon looking at table 6 which gives us the results for *Model 1*, which just classifies whether Drug-Drug Interaction is present or not, the performance was similar to model which was trained by using method 5.2 even though it was

Method	Feature set	Training	Dev	Test
Method 5.1	unBalanced	85.52%	85.6%	-
Method 5.2	Reduced False	74.12%	77.38%	71.8%
Method 5.2	Replicate True	81.84%	81.78%	-
Method 5.2	unBalanced	85.53%	85.6%	-
Method 5.3	Reduced False	49.41%	85.5%	71.48%
Method 5.3	Replicate True	52.24%	81.13%	-
Method 5.3	unBalanced	79.84%	90.89%	-
Method 5.4	Reduced False	57.15%	57.88%	-
Method 5.4	Replicate True	79.84%	70.99%	-
Method 5.4	unBalanced	85.53%	85.6%	-
Method 5.5.1	Replicate True [1D and 2D]	-	-	75.46%
Method 5.5.1	Unbalanced [1D and 2D]	-	-	88.78%
Method 5.5.2	Replicate True [1D and 2D]	-	-	77.01%
Method 5.5.2	Unbalanced [1D and 2D]	-	-	68.74%
Method 5.5.1	Replicate True [Fingerprint]	-	-	66.69%
Method 5.5.1	Unbalanced [Fingerprint]	-	-	64.13%
Method 5.5.2	Replicate True [Fingerprint]	-	-	60.19%
Method 5.5.2	Unbalanced [Fingerprint]	-	-	77.01%

Table 4: Results achieved after *Model 2* cascading

Method	Test
Method 2	NA
Method 3	70.77

Table 5: Results achieved after *Model 3* cascading

Method	Balanced
Method 2	74.18
Method 5.1	76.62
Method 5.2	77.57

Table 6: Results predicted by *Model 1*

trained on very less instances. With *Model 3* model 3 no improvement was observed for models based on method 5.5.1 and 5.5.2 which was expected behaviour since the number of *True* instances were very less to begin with.

6 Why We Think DDI is not a Pure NLP Task

While developing the system we came across the notion of whether DDI prediction could be done purely by relying on NLP techniques or not? Based on the data set we had, which consisted of sentences where the drug names were encountered, paired and indicated whether any interaction would happen in case they were consumed together, we identified cases wherein the system was bound to fail unless it made use of features which were completely outside the realm of Computational Linguistics. One such example of the case was encountering a drug or pair of drug which the system had never encountered or presence of a human error during development of the data set which can have a good probability of occurrence. We made use of various techniques for representing data and using these representations as discussed in above sections we trained a model using support vector machine algorithm. However, we weren't able to get a model with better results.Hence relying only on natural language processing would act as an impediment since you will be restricted to extracting features from the annotated data set which would cause a very restricted medley. Also we performed experiments with feature sets that didn't utilize any of the linguistics aspects of the corpus[5.5] and an unexpected outcome we observed was that relying only on the features pertaining to molecular structures didn't provide us with good results as well.However, here we would like to argue that this happened due to the limitation of DDI corpus which we used since SMILES for many of the drugs were not available. The reason for this are as follows:

- many of the drugs in pair tags were brand names[*survanta*] or group [for example *flu-*

oroquinolones] instead of actual drug names.

- There were many instances wherein the drug names consisted of a drug that belonged to a particular group as well as instance where the group name itself was one of the drugs in the pair.For example, there were instances where the drug name consisted of *Metamfetamine* which belongs to group *Amphetamine*.

- And interesting fact *Amphetamine*, apart from being a name for group of drugs similar to each other is in itself a drug.

- Hence distinguishing between them is challenging. Also many of the drugs in the DDI corpus were a kind of protein which do not have any molecular structure.

Hence, this indicated that DDI is not just NLP but also once cannot reply completely on just the domain features as contextual information helps in identification and augmentation of correct features.

Due to these reasons, relying completely on the annotated data set or just domain related features would not facilitate improvement in system performance to a greater extent. Hence leveraging domain specific knowledge such as molecular structure along with textual features used in one the papers plays a pivotal role not only in improvement of systems predicting capability but also its ability to handle cases it has never encountered before.Using molecular structures of the drugs catalyzed an improvement in the system's prediction capabilities as discussed in the paper published (Asada et al., 2018).

We believe that apart from utilizing information encoded in the molecular structure by using the chemical descriptors generated by any of the multiple software's available would aid in augmenting the predicting capabilities of the system. However, these features on its own wouldn't suffice as shown in Table 4 method 5.5.1. We believe that combining the descriptor features along with textual features would work because not only we would be making use of NLP techniques to leverage contextual information that would help in categorizing the type of interaction but at the same time we wont be relying solely on the linguistic aspects of the data set instead we will make use of the chemical descriptor which would offer us countless number of features that fall under domain specific information which would further enhance the discriminative capabilities of model. From table 6 we can see that using these features the performance of *Model 1* was almost similar to that of *Model 1* which used feature described in method 5.3. It should be noted that instances available for training *Model 1* belonging to method 5.5.1 were half that of method 5.3. Hence we believe that if we can increase the number of instances for method 5.5.1 and 5.5.2 we can certainly get better predictive performance than just relying on textual knowledge.

7 Conclusion

Based on the observations we had during the development of the system we believe that although there are NLP techniques such as lemmatization of drug names,augmenting the true instances in the training datasets with the help of other external data sources, smarter ways of pre-processing the biomedical texts, improving accuracy of named entity recognition, through which we could improve the performance of the system, relying solely on the language processing aspect of data present in training set wouldn't facilitate towards improvement beyond that particular point.It requires employing techniques which are concocted by using not just NLP but also from the realm pertaining to empirical findings to which the problem statement belongs. However, we cannot ignore the fact that ignoring textual features completely and only utilizing the domain features like chemical descriptors is also not a efficacious methodology as textual features do augment the performance as seen in our experiments. Hence finding the right combination of features which is a mix of both NLP and domain knowledge is imperative for developing a efficient predicting model.

Also experimenting with advanced machine learning algorithms like CNN's may help in increasing the predictive capabilities of the model. However, having the right set of features in right quantity is pivotal to the success of a particular model. And system can be developed which would be capable of inferring chemical properties of a drug by tokenizing the chemical structure and applying techniques similar to what NLP does to the sentences.

References

Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2018. Enhancing drug-drug interaction extraction from texts by molecular structure information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 680–685. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

J N Goldstein, I E Jaradeh, P Jhawar, and T O Stair. 2005. ED drug-drug interactions: frequency & type, potential & actual, triage & discharge. *Internet Journal of Emergency & Intensive Care Medicine*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*.

Maria Herrero Zazo Isabel Segura Bedmar, Paloma Martinez. 2013. Ddi 2013 corpus information.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2014. Lessons learnt from the DDIExtraction-2013 Shared Task. *Journal of Biomedical Informatics*.

D. S. Wishart. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*.

Chun Wei Yap. 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*.