Is Drug Drug Interaction a NLP Task?

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

Abstract

Drug Drug interaction information extraction is important yet a complex task. Interaction between two administered drugs could lead to minor side effects or a major health threatening condition. Such types of adverse effects is found to be 8th leading cause of death is the US (Goldstein et al., 2004). Hence we will be exploring options to detect whether given two drugs would there be any interaction among them and also we would be exploring whether DDI is a pure NLP task as we believe detecting drug drug interactions purely depends upon the training data set available and also on ability to utilize information and techniques outside the realm of NLP. We came across articles which stated that the results can be improved by making use of some external factors not related to NLP such as molecular structure in case of DDI (Asada et al., 2018).

1 Introduction

Drug Drug interaction (DDI) task involves identification of drugs from the input sentence and predicting whether any combination of the identified drugs when taken, causes any noxious effect. Identification of drugs names from the sentence falls under Named Entity Recognition (NER) task which was already done in the data set. 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 relied on the data set given.

In 2013, SemEval had formulated a task called "Extraction of Drug Drug Interactions from BioMedical Texts" (Segura-Bedmar et al., 2013)[1] where they provided a Drug Drug Interaction corpus (Herrero-Zazo et al., 2013). It consisted of XML files of drugs from DrugBank database and Medline abstracts on DDI. Based on the training data set provided we could train a model and then using the same model we predict interactions for the test set. 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 real life implementation came across it or it can be the case that when the model was trained and developed a particular drug didn't exist but was later on bought into market. 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 DDI was purely a natural language processing (NLP) task or whether it relied heavily on getting expert knowledge from the field of medicine.

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 (GCN). The GCN has been used to predict interaction between two drugs using their molecular structures (Asada et al., 2018). 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.

2 A Glimpse into Data Set

The data set consisted of a separate XML files for each drugs. The XML files were generated based on drug drug interactions documented on the DrugBank database (Wishart et al., 2006) and

[1] https://www.cs.york.ac.uk/semeval-2013/task9

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

on Medline abstracts which were based on the drug drug interactions. Below table provides a glimpse into the statistics about the data set.

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 a sentence tag which has an attribute id used to uniquely identify the sentence in the corpus and contains text which has the actual sentence itself. The sentence tag contains entity tag and pair tag (Figure 2).

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

`<pair>` tag consists of all possible combination 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.

3 DDI Extraction

We implemented a system to only identify whether two drugs interact when administered in a particular time frame and to explore on whether DDI is purely a NLP task we implemented multiple models and checked the results. Following were the models we implemented.

- Using only the Drug-Drug Pair and their interactions from data set for training the model with the help of Word2Vec.
- Using only the Drug-Drug Pair and their interactions from data set for training the model with the help of one-hot encoding technique.
- Using Drug-Drug pair along with their neighbouring words from the sentences for training the model and and using one-hot representation technique.
- Using Co-sine Similarity between the two drug pairs.
- Using entire sentences by generating Word2Vec representation.

We now look at these methodologies one by one and discuss the results achieved at the end.

3.1 Drug-drug pair with Word2Vec representation

We extracted the drug pairs provided in the XML files form the DDI corpus along with their interactions represented by attribute. We generated Word2Vec representation of drug names and then trained SVM on this data. We did a 80/20 split on the training data set.

3.2 Drug-drug pair with one-hot encoding representation

For this task we extracted all the drug drug pairs from the XML files present in the DDI corpus and generated one-hot encoding. Once we obtained all these drug drug pairs and their labels we generated one-hot encoding for them and using them trained a SVM model. We did a 80/20 split on the training data set.

3.3 Using neighbouring words of the drugs

For this task apart from extracting the drug drug names we decided to make use of the surrounding words present in the sentence where the drug occurred. Once we acquired all the info from the XML files we generated one-hot encoding for them and trained a SVM model. We did a 80/20 split on the training data set here as well.

3.4 Using cosine similarity

For this task we reused the data set which we generated in method in §3.1 and just generated the cosine similarity between the drugs. Then, based on the mean value of all the cosine similarities we decided on a threshold value which could be used to classify other drugs based on their cosine similarities.

3.5 Results

Using Word2Vec to convert drug pairs in training files to feature vectors so that could be fed to SVM for training got 99 percent accuracy on test as most

```
<document id="DrugDDI.d327" origId="Acetohydroxamic Acid">
<sentence id="DrugDDI.d327.s0" origId="s0" text="Concomitant use with iron supplements may result in the
reduced absorption of iron.">
<entity id="DrugDDI.d327.s0.e0" origId="s0.p1" charOffset="21-36" type="drug" text="iron supplement"/>
<entity id="DrugDDI.d327.s0.e1" origId="s0.p5" charOffset="78-82" type="drug" text="iron"/>
<pair id="DrugDDI.d327.s0.p0" e1="DrugDDI.d327.s0.e0" e2="DrugDDI.d327.s0.e1" interaction="false"/>
</sentence>
</document>
```

Figure 1: DDI interaction between two drugs *iron supplement* and *iron*

Methodology	Accuracy
drug pair using Word2Vec	99%
drug pair using one-hot encoding	90.4%
neighbouring words	77.5%
cosine Similarity	90.85%

Table 2: Experiment results

of the 24k drug pairs in training had ‘false’ interactions. Using one-hot encoding to generate feature vectors from drug pairs.We got 90.4% accuracy on test data. Using count vectorizer on not only just the drug pairs but each drugs five nearest neighbours we got 77.5% accuracy on test set.We even passed the training set on the trained model again and observed an accuracy of 92.80%. Used cosine similarity between the word vectors of drug pairs as a feature to enhance the predictions. Used a threshold of 0.0022 on the cosine measure calculated to classify as ‘true/false’. The threshold value was created by initially calculating similarity between each pair and mean of the similarity value was calculated. Once the mean was calculated we started with the mean as threshold and then through trial and error got to 0.0022. An accuracy of 90.85% was observed as most of the drug pairs had interaction as false. Table 2 summarizes these results.

4 Why We Think DDI is Not a Purely NLP Task

While development of the system we came across the notion of whether DDI is purely a NLP task? 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 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 machines however we weren’t able to get a model with better results. We even tried to make use of the surrounding words in the sentence where the drug name was encountered to see if the surrounding words play any role in identification of drug drug interaction but we couldn’t get any noteworthy improvements in results. Also we can see that in most of the methods we got accuracy above 90% because there were extremely large samples of drug pair which had false interactions. Due to this the predictions were skewed towards false classification. 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 a very restricted medley.

Due to these reasons relying completely on the annotated data set would not facilitate improvement in system performance to a greater extent. Hence leveraging domain specific knowledge such as molecular structure used in one the papers we reviewed 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 Asada et al. (2018).

We believe that apart from utilizing information encoded in the molecular structure utilizing the chemical descriptors generated by any of the multiple software’s available which would aid in augmenting the predicting capabilities of the system. We believe this would work because not only we would be making use of NLP techniques but at the same time we will not be relying solely on the linguistic aspects of the data set. Instead, we will make use of the chemical descriptor which would

```
<SentenceText>Do not administer Dobutamine Hydrochloride in 5 % Dextrose Injection simultaneously with
solutions containing sodium bicarbonate or strong alkaline solutions .</SentenceText>
<Mention id="M7" str="Do not administer" span="0 17" type="Trigger"/>
<Mention id="M6" str="sodium bicarbonate" span="112 18" type="Precipitant" code="N0000005741"/>
<Mention id="M8" str="strong alkaline solutions" span="134 25" type="Precipitant" code="NO MAP"/>
<Interaction id="I3" type="Unspecified interaction" precipitant="M6" trigger="M7"/>
<Interaction id="I4" type="Unspecified interaction" precipitant="M8" trigger="M7"/>
```

Figure 2: Example snippet of drug Dobutamine Hydrochloride in Dextrose from training set of TAC 2018

offer us countless number of features that fall under domain specific information.

5 Related Work

A more recent task that featured in TAC 2018 conference on DDI was called "Drug-Drug Interaction Extraction from Drug Labels"[2] which involved multiple tasks. Task 1 involved identification of not only the name of drugs but also the interaction triggers.Triggers are basically words that indicate a type or a kind of an interaction event or an action. For example in Figure 2 we can see that *Do not administer* is a trigger indicating that we shouldn't administer *Dobutamine Hydrochloride in Dextrose* simultaneously with *sodium bicarbonate*. Triggers could also indicate an interaction event that would occur when 2 drugs are administered together for example *lowering of blood pressure* or increase the potential of some side effect.This task involves performing NER. Task 2 involved not only identification of interacting drugs but also the type of interaction for example whether the interaction is pharmacokinetic or an unspecified one. It also involved predicting the outcome of the interaction which could be pharmacokinetic or pharmacodynamic. This task can be solved by using relationship identification evaluation. Task 3 involved normalization of terms as per the details specified on the task page. This type of task is almost *entirely* a NLP task as solutions to these are inclined towards harnessing the techniques of text processing to identify specific kind of words and normalization and also identification of relationship between the words.

6 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, 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 would not facilitate towards improvement beyond particular point. It requires employing techniques which are concocted by using not just NLP but also from realm pertaining empirical findings to which the problem statement belongs. Also a 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, Melbourne, Australia. Association for Computational Linguistics.

Joshua N Goldstein, Issa E Jaradeh, Payal Jhawar, and Thomas O Stair. 2004. ED Drug-Drug Interactions: Frequency & Type, Potential & Actual, Triage & Discharge. *The Internet Journal of Emergency and Intensive Care Medicine*, 8(2).

Mara Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drugdrug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Isabel Segura-Bedmar, Paloma Martínez, and Mara Herrero Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(suppl_1):D668–D672.

[2] https://bionlp.nlm.nih.gov/tac2018druginteractions/