# *EXTRACTION OF DRUG-DRUG INTERACTIONS USING ALL PATHS GRAPH KERNEL.*

Shreyas Karnik[1,2], Abhinita Subhadarshini[1,2], Zhiping Wang[2], Luis M. Rocha[3,4], and Lang Li*[2,5]

[1] School of Informatics, Indiana University, Indianapolis, IN, USA, 46202
[2] Center for Computational Biology & Bioinformatics, School of Medicine, Indiana University , Indianapolis, IN, USA, 46202
[3] School of Informatics & Computing, Indiana University, Bloomington, IN, USA, 47408
[4] Instituto Gulbenkian de Ciencia, Oeiras, Portugal
[5] Department of Medical & Molecular Genetics, School of Medicine, Indiana University , Indianapolis, IN, USA, 46202
Corresponding Author: lali@iupui.edu

**Abstract.** Drug-drug interactions (DDIs) cause nearly 3% of all hospital admissions. Regulatory authorities such as the Food and Drug Administration (FDA) and the pharmaceutical companies keep a rigorous tab on the DDIs. The major source of DDI information is the biomedical literature. In this paper we present a DDI extraction approach based on all paths graph kernel [1] from the DrugDDI corpus [2]. We also evaluate the method on an in-house developed clinical *in vivo* pharmacokinetic DDI corpus. When the DDI extraction model was evaluated on the test dataset from both corpora we recorded a F-score of 0.658 on the clinical *in vivo* pharmacokinetic DDI corpus and 0.16 on the DrugDDI corpus.

## 1 Introduction

Polypharmacy has been a general clinical practice. More than 70% of old population (age >65) take more than 3 medications at the same time in US and some European countries. Since more than 80% of the drugs on the market are metabolized by the Cytochrome P450 enzyme system, and many of these drugs are inhibitors and inducers of CYP450 enzyme system, drug interactions have been extensively investigated *in vitro* and *in vivo* [3,4,5]. These DDIs in many ways affect the overall effectiveness of the drug or at some times pose a risk of serious side effects to the patients [6]. Thus, it becomes very challenging to for the successful drug development and clinical patient care. Regulatory authorities such as the Food and Drug Administration (FDA) and the pharmaceutical companies keep a rigorous tab on the DDIs. Major source of DDI information is the biomedical literature. Due to the unstructured nature of the free text in the biomedical literature it is difficult and laborious process to extract and analyze the DDIs from biomedical literature. With the exponential growth of the



Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), pages 83–88, Huelva, Spain, September 2011.

biomedical literature, there is a need for automatic information extraction (IE) systems that aim at extracting DDIs from biomedical literature. The use of IE systems to extract relationship among biological entities from biomedical literature has experienced success to a great scope [7] for example protein-protein interaction extraction. Researchers have now started to investigate DDI IE from biomedical literature. Some early attempts include retrieval of DDI relevant articles from MEDLINE [8] ;DDI extraction based on reasoning approach [9]; DDI extraction based on shallow parsing and linguistic rules[10];and DDI extraction based on shallow linguistic kernel [2].

BioCreAtIvE has established the standard of evaluation methods and datasets in the area of information extraction [7,11,12,13] which has been a asset for the community. To encourage the involvement of the community in the DDI extraction Segura-Bedmar *et al.*[2] released an annotated corpus of DDIs (DrugDDI corpus) from the biomedical literature and organized the DDIExtraction2011 challenge.

In this article, we implement the all paths graph kernel [1] to extract DDIs from the DrugDDI corpus. We also test the all paths graph kernel approach on in-house corpus that has annotations of pharmacokinetic DDIs from MEDLINE abstracts.

The paper is organized as follows, section 2.1 and 2.2 describe the datasets, section 2.3 describes the all paths graph kernel approach and section 3 describes the results.

## 2 Methodology

**DrugDDI Corpus** We used the unified format [14] of the DrugDDI corpus [2] of the DrugDDI corpus. Detailed description of the corpus can be found at DrugDDI Corpus.

### 2.1 Clinical *in-vivo* pharmacokinetic DDI corpus

Our research group has been studying clinical DDIs reported in biomedical literature (MEDLINE abstracts) and extraction of numerical pharmacokinetic (PK) data from them[15]. During this process, we have collected MEDLINE abstracts that contain clinical PK DDIs, and further develop them into a PK DDI corpus. We decided that the ultimate goal of this task is extraction of DDIs from biomedical literature and it will be interesting to use this corpus as an additional resource. This corpus comprises of 219 MEDLINE abstracts which contains one or more of PK DDIs in same sentences. Here we call it PK-DDI corpus. Please note that a PK DDI means that one drugs exposure is changed by the co-administration of the other drug. As DrugDDI corpus focuses mainly on DDIs that change drug effects, our PK-DDI corpus is a good complementary source. In order to identify drugs in our PK-DDI corpus, we developed a dictionary based tagging approach using all the drug name entries in DrugBank [16]. The corpus was converted into the unified format as proposed in [14]. The DDI instances

were annotated based on guidelines from in-house experts. We split the corpus into training (80%) and testing (20%) fractions. This corpus will also be made public on the lines of the DrugDDI corpus. There are 825 true DDI pairs present in our corpus.

### 2.2 All paths graph kernel

We implemented the approach described by Airola *et al.* [1] for DDI extraction. This approach centers around the drugs, where a graph representation of the sentence is generated. Sentences are described as dependency graphs with interacting components (drugs). The dependency graph is composed of two unconnected sub-graphs: i) One sub-graph explores the dependency structure of the sentence; ii) the other explores the linear order of the words in the sentence. We used the Stanford parser [17]to generate the dependency graphs for both corpora. IIn the dependency graph, the shortest path between two entities was given higher weight as compared to other edges, this is because the shortest path contains important keywords which are indicative of interaction between two entities. In the linear sub-graph, all the edges have the same weight and the order in which words occur before, in the middle, or after drug mentions was considered. The all paths graph kernel algorithm [18] was subsequently implemented to compute the similarity between the graphical representations of the sentences. In particular, all paths graph kernels will be generated for tagger positive DDI sentences and negative DDI sentences. We then used Support Vector Machines (SVM) for classification. More details about the all paths graph kernel algorithm can be found in [1]. A pictorial representation of the approach is presented in figure 1.

## 3 Results

In this study we used an in-house corpus in addition to the DrugDDI corpus; both corpora contain training and testing subsets. We generated DDI extraction models based on both the training datasets individually and combined, and evaluated the performance of the DDI extraction models on the respective testing datasets.

Table 1 illustrates the summary of training and testing data in two corpuses. For the purpose of evaluation we used precision, recall and the balanced F-Score measure. We also performed 10-fold cross-validation during the training phase.

Table 2 displays the DDI extraction performance on DDI-PK corpus testing data. It suggests that using the DDI-PK corpus training data either with or without the DrugDDI corpus training data, led to the precision above 0.78 and recall above 0.64. On the other hand, if only the DrugDDI corpus was used, both precision and recall were around 0.41.

Table 3 displays the DDI extraction performance on DrugDDI corpus testing data. It suggests that all there models had similar perform in F-score, which was between 0.13 and 0.16, although using DDI PK corpus generated slightly better F-score than the other two approaches.

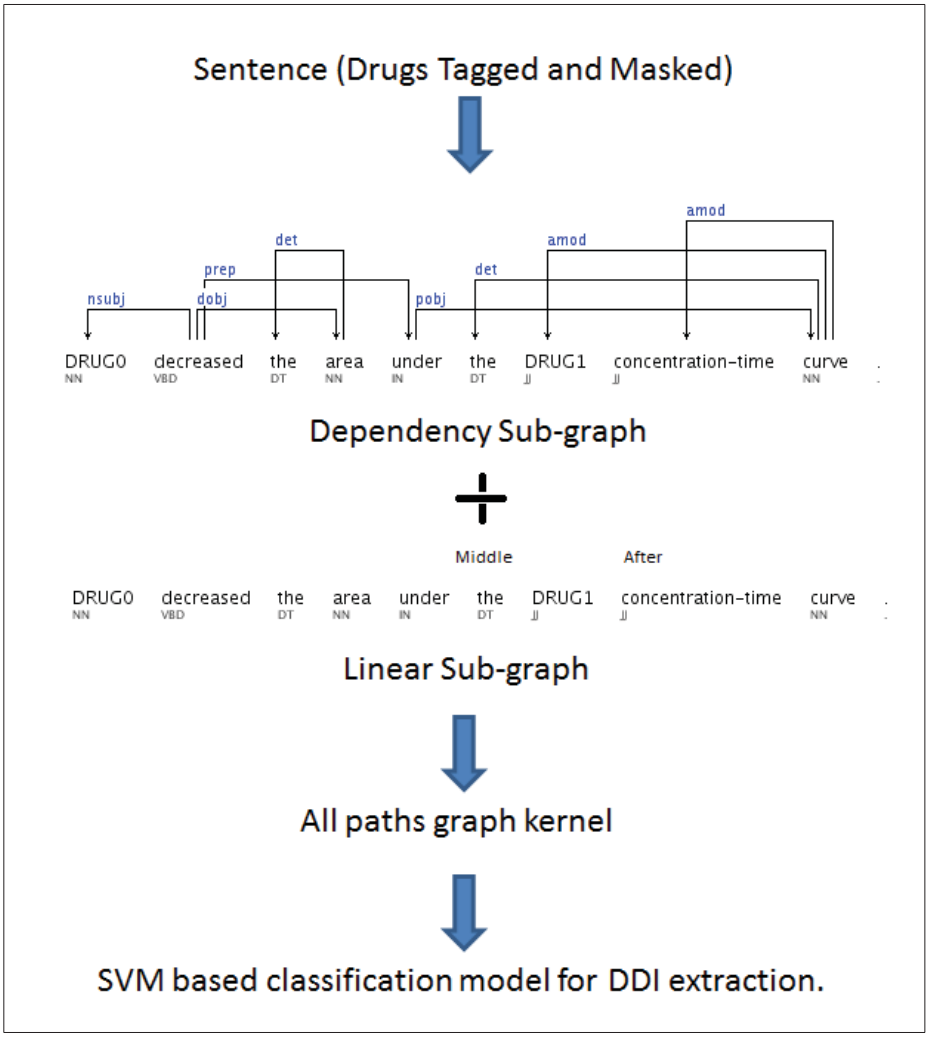


**Fig. 1.** Description of the methodology

| Dataset | Number of sentences | Number of DDI Pairs |
|---|---|---|
| PK DDI Corpus (Train) | 1939 | 2411 |
| PK DDI Corpus (Test) | 498 | 606 |
| DDI Corpus (Train) | 3627 | 20888 |
| DDI Corpus (Test) | 1539 | 7026 |

**Table 1.** Summary of the corpora used in this study

| Dataset | F-Score | Precision | Recall |
|---|---|---|---|
| PK DDI Corpus (Train) + DDI Corpus (Train) | 0.64 | 0.53 | 0.8 |
| PK DDI Corpus (Train) | 0.658 | 0.567 | 0.7857 |
| DDI Corpus (Train) | 0.415 | 0.417 | 0.414 |

**Table 2.** Performance of the different models on PK DDI Corpus (Testing dataset)

| Dataset | F-Score | Precision | Recall |
|---|---|---|---|
| PK DDI Corpus (Full) + DDI Corpus (Train) | 0.1346 | 0.1250 | 0.1457 |
| PK DDI Corpus (Full) | 0.1605 | 0.1170 | 0.2556 |
| DDI Corpus (Train) | 0.1392 | 0.1187 | 0.1682 |

**Table 3.** Performance of the different models on DrugDDI corpus test data

## 4 Discussion and Conclusion

There is large room for improvement in the DDI extraction from the biomedical literature. We also learned that the in-house DDI PK corpus and Drug DDI corpus have different DDI structures. It seems the all paths graph kernel method performed better in DDI PK corpus than the Drug DDI corpus.

The apparent low precision and recall in the Drug DDI corpus may result from the fact that the number of real DDIs is much less than the number of false DDIs in both corpus, but a comparison with the results of other teams is forthcoming once those get released. It is also possible that the weights on the sub-graph need to be further adjusted to get a better performance. We noticed a marked performance difference between the training corpora. The sentences in the DrugDDI corpus were long and complex. On the other hand, our DDI PK corpus has a simply sentence structure, and there is an average of one to two DDI pairs per abstract. Even with the same algorithm, these major differences between two corpora resulted in different DDI extraction performances.

DrugDDI corpus focuses on DDIs that affect the clinical outcomes (i.e. pharmacodynamics DDI); while PK DDI corpus focuses on DDIs that change the drug exposure. They are complementary to each other. Therefore, our work enriches the set of resources and analysis available to this community.

## References

1. Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F., Salakoski, T.: All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics **9**(Suppl 11) (2008) S2
2. Segura-Bedmar, I., Martnez, P., de Pablo-Snchez, C.: Using a shallow linguistic kernel for drug-drug interaction extraction. Journal of Biomedical Informatics **In Press, Corrected Proof** (2011) –

3. Zhou, S., Yung Chan, S., Cher Goh, B., Chan, E., Duan, W., Huang, M., McLeod, H.L.: Mechanism-based inhibition of cytochrome P450 3A4 by therapeutic drugs. Clinical Pharmacokinetics **44**(3) (2005) 279–304
4. Cupp, M., Tracy, T.: Cytochrome P450: New nomenclature and clinical implications. American Family Physician **57**(1) (1998) 107
5. Lin, J., Lu, A.: Inhibition and induction of cytochrome P450 and the clinical implications. Clinical pharmacokinetics **35**(5) (1998) 361–390
6. Sabers, A., Gram, L.: Newer anticonvulsants: Comparative review of drug interactions and adverse effects. Drugs **60**(1) (2000) 23–33
7. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of biocreative: critical assessment of information extraction for biology. BMC Bioinformatics **6**(Suppl 1) (2005) S1
8. Duda, S., Aliferis, C., Miller, R., Statnikov, A., Johnson, K.: Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases. In: American Medical Informatics Association (AMIA) Annual Symposium proceedings, American Medical Informatics Association (2005) 216–220
9. Tari, L., Anwar, S., Liang, S., Cai, J., Baral, C.: Discovering drug-drug interactions: A text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics **26**(18) (2010) i547–i553
10. Segura-Bedmar, I., Martinez, P., de Pablo-Sanchez, C.: A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics **12**(Suppl 2) (2011) S1
11. Yeh, A., Morgan, A., Colosimo, M., Hirschman, L.: Biocreative task 1a: gene mention finding evaluation. BMC Bioinformatics **6**(Suppl 1) (2005) S2
12. Hirschman, L., Colosimo, M., Morgan, A., Yeh, A.: Overview of biocreative task 1b: normalized gene lists. BMC Bioinformatics **6**(Suppl 1) (2005) S11
13. Blaschke, C., Leon, E., Krallinger, M., Valencia, A.: Evaluation of biocreative assessment of task 2. BMC Bioinformatics **6**(Suppl 1) (2005) S16
14. Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F., Salakoski, T.: Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics **9**(Suppl 3) (2008) S6
15. Wang, Z., Kim, S., Quinney, S.K., Guo, Y., Hall, S.D., Rocha, L.M., Li, L.: Literature mining on pharmacokinetics numerical data: A feasibility study. Journal of Biomedical Informatics **42** (2009) 726–735
16. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., Wishart, D.S.: DrugBank 3.0: A comprehensive resource for 'Omics' research on drugs. Nucleic Acids Research (2010)
17. De Marneffe, M., MacCartney, B., Manning, C.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC. Volume 6., Citeseer (2006) 449–454
18. Gartner, T., Flach, P., Wrobel, S.: On graph kernels: Hardness results and efficient alternatives. In: Learning theory and Kernel machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003: proceedings, Springer Verlag (2003) 129