The DDI$_{NCBI}$ Corpus — Towards a Larger Resource for Drug-Drug Interactions in PubMed

Lana Yeganova[1], Sun Kim[1], Grigory Balasanov[1], Kristin Bennett[2], Haibin Liu[1], W. John Wilbur[1]
[1]National Center for Biotechnology Information, NLM, NIH, Bethesda, MD, USA
[2]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA
E-mail: {lana.yeganova, sun.kim, grigory.balasanov, haibin.liu, john.wilbur}@nih.gov, bennek@rpi.edu

Abstract

Manually annotated corpora are of great importance for the development of NLP systems, both as training and evaluation data. However, the shortage of annotated corpora frequently presents a key bottleneck in the process of developing reliable applications in the health and biomedical domain and demonstrates a need for creating larger annotated corpora. Utilizing and integrating existing corpora appears to be a vital, yet not trivial, avenue towards achieving the goal. Previous studies have revealed that drug-drug interaction (DDI) extraction methods when trained on DrugBank data do not perform well on PubMed articles. With the ultimate goal of improving the performance of our DDI extraction method on PubMed(®) articles, we construct a new gold standard corpus of drug-drug interactions in PubMed that we call the DDI$_{NCBI}$ corpus. We combine it with the existing DDIExtraction 2013 PubMed corpus and demonstrate that by merging these two corpora higher performance is achieved compared to when either source is used separately. We release the DDI$_{NCBI}$ corpus and make it publicly available for download in BioC format at: http://bioc.sourceforge.net/. In addition, we make the existing DDIExtraction 2013 corpus available in BioC format.

Keywords: Cross-corpus text mining, DDI$_{NCBI}$ corpus, Drug-drug interactions, BioC format

1. Introduction

Several studies have attempted to combine corpora on a given topic and analyse cross-corpus text mining (Pyysalo, Airola et al. 2008, Tikk, Thomas et al. 2010, Ayvaz, Horn et al. 2015). While it appears to be promising, two groups studying these issues did not show improvement in predictive performance of classifiers (Tikk, Thomas et al. 2010, Ayvaz, Horn et al. 2015).

Our interest in this study was motivated by an objective to improve the performance of the drug-drug interaction identification system (Kim, Liu et al. 2015) on PubMed abstracts. Drug-drug interactions represent a major but potentially preventable medical issue that accounts for over 30% of all adverse drug reactions. (Strandell, Bate et al. 2008, Iyer, Harpaz et al. 2014). Many DDI resources exist (Knox, Law et al. 2011, Takarabe, Shigemizu et al. 2011, Baxter and Claire L 2013), yet they cover only a fraction of knowledge available. A significant amount of up-to-date information is hidden in the text of PubMed journal articles. That is why mining PubMed data for the DDI signal is essential.

The series of DDIExtraction challenges (Segura-Bedmar, Martinez et al. 2011, Segura-Bedmar, Martinez et al. 2013) sparked community-wide competitions addressing the DDI extraction problem and provided annotated data from DrugBank and PubMed (Herrero-Zazo, Segura-Bedmar et al. 2013). While the DDIExtraction 2011 corpus was composed of texts describing DDIs from the DrugBank only (Knox, Law et al. 2011), the DDIExtraction 2013 corpus also integrated PubMed abstracts in order to deal with different type of texts and language styles. The challenges revealed that the performance of DDI detection classifiers is substantially lower for texts from PubMed than it is for DrugBank. The difference in performance could be due to different characteristics of texts (Chowdhury and Lavelli 2013, Kim, Liu et al. 2015) and the small number of training examples provided for PubMed. Indeed, the PubMed portion of the DDIExtraction 2013 dataset, which is referred to as DDI-Medline, contains 233 annotated abstracts.

In trying to address these points, we develop a new corpus for PubMed that we call the DDI$_{NCBI}$ corpus and examine whether or not the performance of the classifier can be improved by integrating the sources. We present the DDI$_{NCBI}$ corpus as a step towards a more comprehensive DDI resource for PubMed which calls for combining the existing and new resources for achieving better predictive power.

The contributions of this article are: 1. Introduction of the new DDI$_{NCBI}$ corpus as a resource to build and evaluate new and existing DDI recognition methods, 2. Providing evidence that leveraging labeled data by integrating multiple resources could lead towards better predictive power of classifiers, 3. Public release of the DDI$_{NCBI}$ corpus as well as conversion of both corpora, DDI$_{NCBI}$ and DDI-Medline, into BioC format.

2. The DDI$_{NCBI}$ Corpus

The DDI$_{NCBI}$ corpus consists of 535 sentences, each containing a pair of pharmacological substances, and is annotated for the presence or absence of information describing the interaction between them, resulting in 122 positive and 413 are negative sentences. In this section, we briefly describe the process followed in the annotations of drugs and their interactions in the DDI$_{NCBI}$ corpus. The DDI$_{NCBI}$ corpus is freely available for download in BioC format.

2.1. Selecting Candidate DDI Sentences

We selected a subset of 5 million PubMed abstracts covering documents dated between December 2008 and July 2014, and divided them into sentences using the MedPost part of speech tagger (Smith, Rindflesch et al. 2004). Then, a complete list of all drug names was downloaded from DrugBank (Knox, Law et al. 2011) and PubMed sentences from the 5 million that contain exactly two drug name entities where collected. DrugBank was chosen for this purpose because of its broad inclusion of drugs (Ayvaz, Horn et al. 2015), which along with pharmaceuticals includes other natural substances for instance *glycine* or *estradiol*. As such, the drug entity recognition was assumed and the annotations for drugs as found in DrugBank provided to the annotators.

Previous studies have consulted the MeSH® ontology for selecting candidate documents from PubMed for annotations. MeSH is a controlled vocabulary of terms that is used for indexing PubMed articles. A detailed explanation of MeSH can be found at http://www.nlm.nih.gov/mesh/. The candidate documents were required to have the MeSH term *"Drug Interactions"* (Herrero-Zazo, Segura-Bedmar et al. 2013) or its derivatives, such as *"Drug Hypersensitivity"*, *"Drug Antagonism"* (Duda, Aliferis et al. 2005) assigned to a document. We chose a data-driven approach and selected sentences that along with a pair of drug entities contain a trigger word or phrase typically used for describing drug interactions. The set of triggers was identified by manually examining a group of DDI sentences in PubMed and consists of 108 patterns presented in Supplement 1 (http://bioc.sourceforge.net/). This process resulted in 10,467 sentences that contained a pair of drug entities and a trigger word or phrase.

The list of sentences was further scored using the rich feature-based linear kernel approach (Kim, Liu et al. 2015) and a set of 600 sentences chosen for manual review. Positive score indicates the DDI information is present in a sentence, while negative signals the opposite. The selected sentences represent a mix between moderately scoring positive sentences (we excluded the range of high scoring positives) and high scoring negatives. The intention was to choose more challenging instances which could potentially be of more value when annotated.

2.2. Annotating Candidate DDI Sentences

The annotation work on the corpus was performed in three rounds. The first round took place in Spring of 2015, when a class of 30 students was distributed 600 sentences to annotate. Students were split into twelve groups, each consisting of two or three students, and every group was assigned to annotate 50 sentences. Students within each group were instructed to work together to come up with the answer reflecting whether or not the sentence describes the interaction between the two drugs. The students were working towards a bachelor's degree in data science.

The second round of annotations took place in Fall of 2015, when the same set of 600 sentences was annotated by a group of six scientists with backgrounds in biomedical informatics research. Each scientist annotated 100 sentences. Out of 600 sentences that have been annotated, the parties agreed on 372 sentences (with 118 judged positive and 254 judged negative for DDIs), disagreed on 145 sentences, and at least one of the sides could not make a decision on 83 sentences. For those sentences where decision has been reached by both sides, the inter-annotator agreement was 72%.

The 228 sentences that received different annotations from student groups and scientists were flagged for the third round of reviews. The third round of reviews was conducted by three scientists (among the original group of six scientists). Each one of the three reviewed sentences that were different from those offered at Round 2. At that stage a decision about the sentence has been reached. With that every sentence has been looked at by a group of students and at least one scientist.

During manual annotation we found that some chemicals downloaded from DrugBank are not drugs or substances that could be used as drugs. We dropped the sentences which contained such chemicals from consideration. Our final analysis resulted in a set of 535 sentences of which 122 are annotated positive and 413 negative.

2.3. The DDI$_{NCBI}$ Corpus in BioC format

When choosing to use more than one corpus, the text miners frequently need to deal with more than one format for the text documents and annotations and write specific parsers for each of them. This has been a problem that the BioC initiative (Comeau, Islamaj Dogan et al. 2013) aimed to solve with the recent introduction of the BioC XML format. The BioC project attempts to address the interoperability among existing natural language processing tools by providing a unified BioC XML format. The newly annotated DDI$_{NCBI}$ corpus is distributed in BioC format with the goal to promote high corpus usage. This shared format follows the standoff annotation principle in which the original sentence text is preserved and all entities are stored as offsets, an example is presented in Figure 1. We also make the DDI-Medline corpus available for download in BioC format from http://bioc.sourceforge.net/.

```
<document>
<id>22900583</id>
<passage>
<infon key="DDI">Yes</infon>
<offset>0</offset>
<text>
These data demonstrate that ritonavir is able
to block prasugrel CYP3A4 bioactivation.
</text>
<annotation id="0">
<infon key="type">DrugName</infon>
<location offset="28" length="9"/>
<text>ritonavir</text>
</annotation>
<annotation id="1">
<infon key="type">DrugName</infon>
<location offset="55" length="9"/>
<text>prasugrel</text>
</annotation></passage></document>
```

Figure 1. A fragment from the annotated DDI$_{NCBI}$ corpus in the BioC format.

3. Merging the Corpora – Experiments and Results

We perform experiments to test if merging DDI-Medline & DDI_{NCBI} datasets improves the performance of the existing state-of-the-art linear SVM classifier developed in our earlier work (Kim, Liu et al. 2015). As described in the paper, we first apply the standard tokenization step, and to ensure generalization of the features, drug mentions are anonymized with "DRUG" for drug entities, numbers are replaced by a generic tag "NUM", and other tokens normalized into their corresponding lemmas by the BioLemmatizer (Liu, Christiansen et al. 2012).

In that study we outlined five types of features (words with relative positions, pairs of non-adjacent words, dependency relations, syntactic structures and noun phrase-constrained coordination tags) and demonstrated that the words with relative positions and pairs of non-adjacent words provide the greatest contribution to the performance of the classifier. When using only these two types of features on DDI-Medline set the classifier has achieved an F1 score of 0.738 as compared to the best F1 score of 0.752 when all five types of features were used. Taking into consideration that there is only 1.4% decrement in performance using a much simpler representation, we proceed by constructing only these two types of features to test the performance of the classifier on the new dataset.

Two experiments are conducted to examine the contribution of the DDI_{NCBI} dataset. In the first experiment, we compared the 10-fold cross validation on the DDI-Medline dataset with exactly the same 10-fold cross-validation on the DDI-Medline dataset with each training fold augmented with the DDI_{NCBI} dataset. In the second experiment, we compared the 10-fold cross validation on the DDI_{NCBI} dataset with exactly the same 10-fold cross-validation on the DDI_{NCBI} dataset with each training fold augmented with the DDI-Medline dataset. Table 1 presents the basic statistics of the corpora. Tables 2 and 3 demonstrate the results of these tests and report the Average Precision, Precision, Recall and F-1 scores.

Sent per Corpus	DDI-Medline	DDI_{NCBI}
# of Positive Sent	338	122
# of Neg Sent	1,688	413
Total Sentences	2,026	515

Table 1: Basic Statistics of the DDI_{NCBI} Corpus and DDI-Medline corpora in terms of number of sentences included.

10-fold CV	Avg Prec	Prec	Recall	F1
DDI-Medline	0.7473	0.7445	0.6308	0.6829
DDI-Medline + DDI_{NCBI}	0.7495	0.7610	0.6385	0.6943

Table 2: Performance comparison between DDI-Medline and the augmented DDI-Medline+ DDI_{NCBI} corpus. Results are based on 10-fold cross validation and evaluate Precision, Recall and F1 on DDI-Medline when additional DDI_{NCBI} corpus is made available during training.

10-fold CV	Avg Prec	Prec	Recall	F1
DDI_{NCBI}	0.5541	0.6744	0.2769	0.3922
DDI_{NCBI} + DDI-Medline	0.6335	0.7043	0.4240	0.5291

Table 3: Performance comparison between DDI_{NCBI} and the augmented DDI_{NCBI}+DDI-Medline corpus. Results are based on 10-fold cross validation and evaluate Precision, Recall and F1 on the DDI_{NCBI} corpus when additional DDI-Medline corpus is made available during training.

These experiments demonstrate that adding more training data improves the performance in the last row of both tables. As seen in Table 2, we observe an increase in F1 score from 0.6829 to 0.6943 when tested on the DDI-Medline set, and an improvement of F1 score from 0.3922 to 0.5291 when tested on the DDI_{NCBI} set. Interestingly, the last row of Table 3 involves slightly more training data than the last row of Table 2, but shows significantly lower performance. This could mean different characteristics of DDIs covered in the two corpora, or more difficult cases in the DDI_{NCBI} corpus. We believe the overall quality of DDI_{NCBI} is good because DDI_{NCBI} leads to improvement when added as training to the DDI-Medline, especially in precision. We also hypothesize that the characteristics of the sentences describing the DDIs are somewhat different and by combining the sets we get an enriched corpus.

4. Conclusion

Inherent complexity of natural language and convoluted style of scientific writing make the DDI extraction problem from PubMed a challenge. With the goal to improve the performance of a drug-drug interaction identification system (Kim, Liu et al. 2015) on PubMed abstracts, we create and release DDI_{NCBI}, a corpus of 535 sentences manually annotated for drug-drug interaction information. We further combine our corpus with the DDI-Medline corpus and demonstrate that adding more training improves the performance of the classifier.

In the future, we intend to extend our study on facilitating cross-corpus text mining by leveraging additional resources, such as the corpus of pharmacokinetic interactions (Kolchinsky, Lourenco et al. 2015) in PubMed.

5. Acknowledgements

The authors thank Isabel Segura-Bedmar for her facilitation in releasing the DDI-Medline dataset in the BioC format. Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

6. Bibliographical References

Ayvaz, S., et al. (2015). Toward a complete dataset of drug–drug interaction information from publicly available sources. Journal of Biomedical Informatics 55: 206-217.

Baxter, K. and P. Claire L (2013). Stockley's Drug Interactions, 10th edition. London, Pharmaceutical Press.

Chowdhury, M. F. M. and A. Lavelli (2013). FBK-irst : A

Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, Georgia.

Comeau, D., et al. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. Database 2013.

Duda, S., et al. (2005). Extracting Drug-Drug Interaction Articles from MEDLINE to Improve the Content of Drug Databases AMIA Annu Symp Proc: 216–220.

Herrero-Zazo, M., et al. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of Biomedical Informatics 46: 914-920.

Iyer, S. V., et al. (2014). Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association 21(2): 353–362.

Kim, S., et al. (2015). Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. Journal of Biomedical Informatics 55: 23-30.

Knox, C., et al. (2011). DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Research (Database issue) 39(Suppl 1): D1035–D1041.

Kolchinsky, A., et al. (2015). Extraction of Pharmacokinetic Evidence of Drug–Drug Interactions from the Literature. PLOS one 10(5).

Liu, H., et al. (2012). BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. Journal of Biomedical Semantics 3(3).

Pyysalo, S., et al. (2008). Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics 9(Suppl 3): S6.

Segura-Bedmar, I., et al. (2013). SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). Second Joint Conference on Lexical and Computational Semantics (*SEM), Atlanta, GA.

Segura-Bedmar, I., et al. (2011). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011): 1-9.

Smith, L., et al. (2004). MedPost: A part of speech tagger for biomedical text. Bioinformatics 20(14): 2320-2321.

Strandell, J., et al. (2008). Drug-drug interactions—a preventable patient safety issue? Br J Clin Pharmacol 65(1): 144-146.

Takarabe, M., et al. (2011). Network-based analysis and characterization of adverse drug-drug interactions. Journal of Chemical Information and Modeling 51(11).

Tikk, D., et al. (2010). A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. Plos Computational Biology 6(7).