Automatic Approaches for Gene-Drug Interaction Extraction from Biomedical Text: Corpus and Comparative Evaluation

Nate Sutton[1*], Laura Wojtulewicz[1], Neel Mehta[1], Graciela Gonzalez[1]
[1]Department of Biomedical Informatics, Arizona State University, AZ
*Corresponding author: nate.sutton@asu.edu

1. Introduction

Drug-gene interactions (GDI) are a part of the field of pharmacogenomics, a discipline that indentifies ways that genes influence drugs and diseases. In interaction extraction NLP research, a variety of work has been done on protein-protein interaction extractions but less on GDI extractions. Very limited corpora available for developing GDI extraction systems has hindered development. To assist developers to advance work on systems for extracting GDIs, a corpus has been developed and an evaluation has been performed on different approaches for creating those extractions. The corpus is available at http://diego.asu.edu/index.php/projects

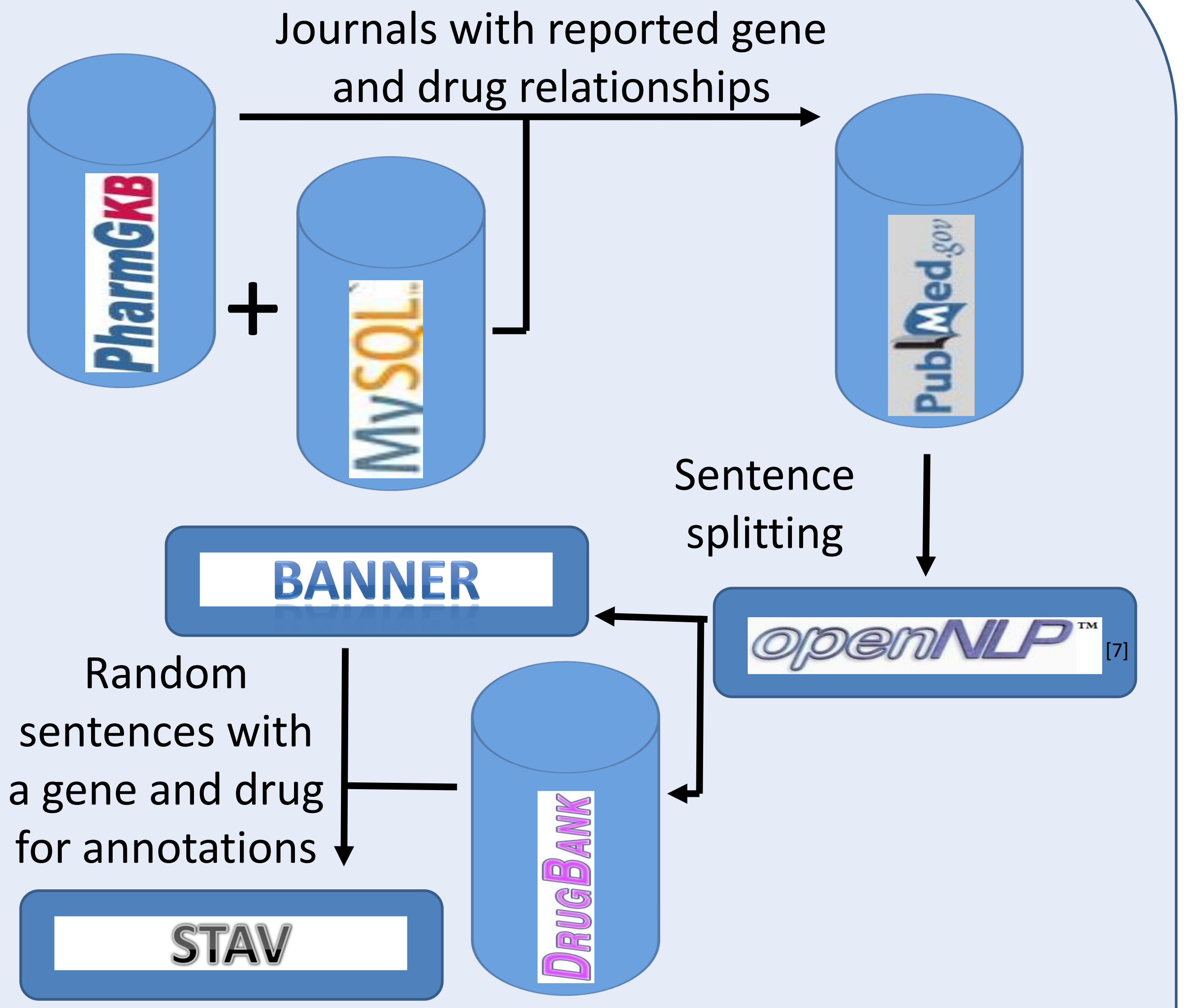
2. Corpus development methods

- 551 sentence corpus from Pubmed abstracts in 591 journals found relevant to GDIs [1].
- Relevant journals were identified as journals with at least one article containing a gene-drug relationship reported in a pharmacogenomics relationships collection from PharmGKB [2].
- The corpus includes only sentences with a drug and gene found with NER using Banner and a DrugBank dictionary, respectively [3,4].
- An annotation guide and customizations of the Stav annotation tool were made to enable interaction annotation [5].
- Inter-annotator scores were found from amongst 3 annotators.

3. Interaction extraction approach analyses

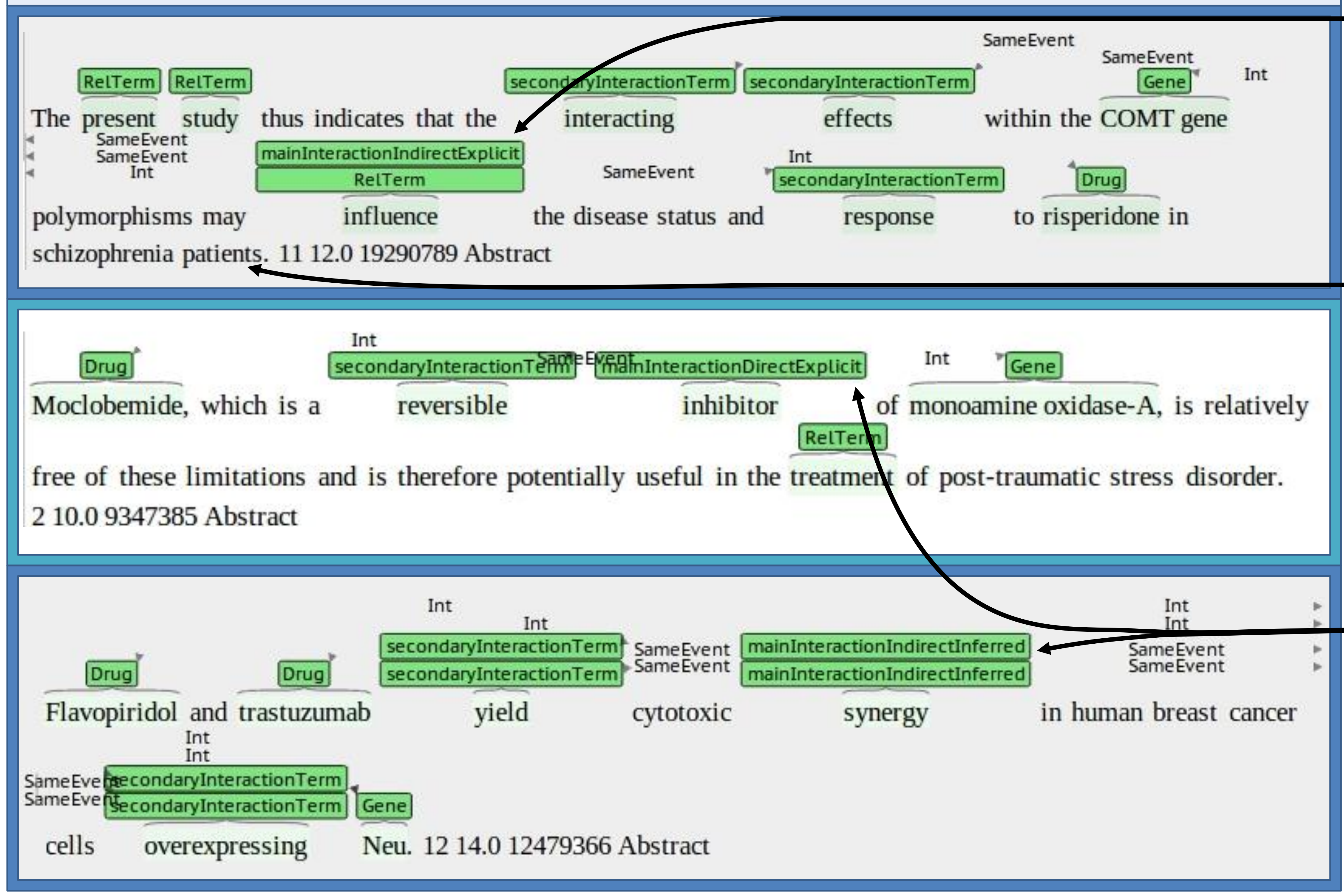
- Extraction approaches used are basic co-occurrence, co-occurrence plus interaction terms, and a re-implementation of a more advanced pattern-based approach.
- Interaction terms were extracted from a Phare pharmacogenomics ontology [6].
- Patterns for extraction were based on patterns in the work of Coulet et. al. and used linguistic structures known as dependency graphs.
- The pattern based approach showed value in high precision extractions but had low recall. Co-occurrence had the highest f-score and using interaction terms also did a reasonable job.

4. Corpus construction


Journals with reported gene and drug relationships
PharmGKB
+
MySQL
PubMed.gov
Sentence splitting
openNLP [7]
BANNER
Random sentences with a gene and drug for annotations
DrugBank
STAV

- Filtering of articles with gene and drug relationships from the PharmGKB collection was done with MySQL [8].
- Once random sentences were collected, they and their NER identified entities were converted to the BioNLP shared task file format for use with Stav.

5. Example sentences



The present study thus indicates that the interacting effects within the COMT gene polymorphisms may influence the disease status and response to risperidone in schizophrenia patients. 11 12.0 19290789 Abstract

Moclobemide, which is a reversible inhibitor of monoamine oxidase-A, is relatively free of these limitations and is therefore potentially useful in the treatment of post-traumatic stress disorder. 2 10.0 9347385 Abstract

Flavopiridol and trastuzumab yield cytotoxic synergy in human breast cancer cells overexpressing Neu. 12 14.0 12479366 Abstract

6. Annotation types

Interaction A gene and drug broadly categorized as having an "action, effect, or influence" on another is in an interaction.

Entities involved in interactions:

Interaction Term Terms that are descriptive of the interaction (as defined earlier).

Intermediary Entity Non-gene, non-drug entities that are needed for understanding the full semantic meaning of interactions (only annotated through the indirectness property).

Properties of interactions are:

- **Direct/Indirect:** Interactions that are indirect due to intermediate entities being needed to understand them.
- **Explicit/Inferred:** The necessity of an inference due to an explicit interaction statement not being present.

Non-interaction

- **Shared Entity:** An entity connected to both a gene and a drug that don't interact with each other.

7. Inter-annotator agreement results

	Annotator 1 & 2	Annotator 1 & 3	Annotator 2 & 3
Accuracy	81.1%	74.2%	73.0%
Kappa	45.7%	30.5%	11.4%

Inter-annotator agreement percentage scores. Intentionally using random sentences resulted in having an imbalance between positive and negative instances and therefore lowered kappa scores by increasing kappa's correction of chance agreement.

8. Interaction extraction results

Interaction Extractor Type	Precision (TP/TP+FP)	Recall (TP/TP+FN)	F1-Score (2*((P*R)/(P+R)))
Co-occurrence	68.99% (781/1132)	100.00% (781/781)	81.65%
Co-occurrence plus interaction terms	69.60% (664/954)	85.02% (664/781)	76.54%
Pattern-based	96.61% (57/59)	7.30% (57/781)	13.57%

Extraction system performances. Note that sentences were selected based on co-occurrence of a gene and a drug, thus recall is 100% for that method, as it essentially defines the corpus.

9. References

[1] "Home - PubMed - NCBI." [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/. [Accessed: 29-May-2012].
[2] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein, "PharmGKB: The Pharmacogenetics Knowledge Base," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 163–165, Jan. 2002.
[3] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," *Pac Symp Biocomput*, pp. 652–663, 2008.
[4] D. S. Wishart, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. 90001, pp. D668–D672, Jan. 2006.
[5]"TsujiiLaboratory/stav - GitHub." [Online]. Available: https://github.com/TsujiiLaboratory/stav. [Accessed: 06-Dec-2011].
[6] A. Coulet, N. H. Shah, Y. Garten, M. Musen, and R. B. Altman, "Using text to build semantic networks for pharmacogenomics," *J Biomed Inform*, vol. 43, no. 6, pp. 1009–1019, Dec. 2010.
[7] "The OpenNLP Homepage." [Online]. Available: http://opennlp.sourceforge.net/projects.html. [Accessed: 26-Mar-2012].
[8] "MySQL :: The world's most popular open source database." [Online]. Available: http://www.mysql.com/. [Accessed: 29-May-2012].