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

Nate Sutton, Laura Wojtulewicz, Neel Mehta, Graciela Gonzalez
Department of Biomedical Informatics
Arizona State University, Tempe, Arizona, USA
{nate.sutton, whitz, nbmehta2, graciela.gonzalez}@asu.edu

Abstract

Publications that report genotype-drug interaction findings, as well as manually curated databases such as DrugBank and PharmGKB are essential to advancing pharmacogenomics, a relatively new area merging pharmacology and genomic research. Natural language processing (NLP) methods can be very useful for automatically extracting knowledge such as gene-drug interactions, offering researchers immediate access to published findings, and allowing curators a shortcut for their work.

We present a corpus of gene-drug interactions for evaluating and training systems to extract those interactions. The corpus includes 551 sentences that have a mention of a drug and a gene from about 600 journals found to be relevant to pharmacogenomics through an analysis of gene-drug relationships in the PharmGKB knowledgebase.

We evaluated basic approaches to automatic extraction, including gene and drug co-occurrence, co-occurrence plus interaction terms, and a linguistic pattern-based method. The linguistic pattern method had the highest precision (96.61%) but lowest recall (7.30%), for an f-score of 13.57%. Basic co-occurrence yields 68.99% precision, with the addition of an interaction term precision increases slightly (69.60%), though not as much as could be expected. Co-occurrence is a reasonable baseline method, with pattern-based being a promising approach if enough patterns can be generated to address recall. The corpus is available at http://diego.asu.edu/index.php/projects

1 Introduction

Pharmacogenomics is a relatively new area of biomedical research that merges pharmacology and molecular genomics, among other disciplines, and focuses on studying the effects of genetic variability on drug toxicity and efficacy, on the discovery of novel genomic targets for drug development, and on the identification and functional characterization of polymorphisms relevant to drug action. Thus, publications that report genotype-drug findings and manually curated databases that collect such findings, like PharmGKB and DrugBank (Hewett et al., 2002; Wishart, 2006) are of paramount importance to the field. However, manual curation is expensive and time consuming, and cannot keep up with the ever increasing number of publications. Natural language processing (NLP) methods can be very useful for automatically extracting such gene-drug interactions, offering researchers immediate access to published findings, and allowing curators a shortcut for their work.

Consider for example a sentence containing an interaction NLP can help extract: "Only the epsilon4 allele of APOE was found to make a significant (P = 0.002) but small contribution to warfarin dose requirement." (PMID: 16847429). We can easily see that in the sentence, an APOE allele interacts with the drug warfarin in its dose requirement. Furthermore, at a higher level of abstraction, the sentence can help researchers infer that APOE affects the metabolic processes targeted by the drug warfarin.

NLP researchers attacking an interaction extraction project such as this one, will usually start by identifying the entities involved in the extractions and the terms that indicate such interactions. Assuming named entity recognition (NER) systems exist for the entities in question (or a dictionary is available for direct match), the main concern becomes extracting true interactions. A gold standard corpus would then need to be identified or created in order to evaluate and develop interaction extraction approaches, starting with the

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 214–222,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

simplest ones. We aim to support advancement in the area of gene-drug interaction extraction through the construction of a corpus for that task that offers advantages not available in another similar corpus. Also for that support we report on a study of the capabilities of different methods for that form of extraction.

To achieve our aim, we describe a new corpus of gene-drug interactions, and compare the performance of two basic approaches plus the re-implementation of a more advanced pattern-based approach measured against this corpus. We do not seek in this publication to advance the extraction methods themselves, but allow a side-to-side comparison of approaches on a single corpus.

The sentences in the corpus (a total of 551) were randomly selected from sentences that include both a gene and a drug mention from the abstracts published on a selection of journals that have articles relevant to pharmacogenomics. In general, annotations include interactions evident from the sentence, if any, also noting when mentioned genes or drugs are *not* involved in interactions. All sentences were annotated by the main author, with a second and third annotator verifying 26% of the corpus. The corpus is publicly available online along with other supplementary materials including the annotation guide[1].

The extraction methods evaluated include co-occurrence of a gene and a drug, co-occurrence of a gene and a drug plus a recognized interaction term, and one that uses specific linguistic patterns for classification based on (Coulet, Shah, Garten, Musen, & Altman, 2010). The linguistic pattern method had the highest precision (96.61%) but lowest recall (7.30%), for an f-score of 13.57%. Basic co-occurrence yields 68.99% precision, with the addition of an interaction term increasing precision slightly (69.60%), though not as much as could be expected. Analysis of our results show that performance could be immediately improved by improving the fundamental entity-recognition of drugs and genes.

2 Related Work

A good portion of the work presented here follows prior approaches to high quality protein-protein interaction (PPI) corpora development and extraction. Given that our corpus contains genes and proteins as entities, procedures used to create PPI corpora were a useful resource. A variety of annotation decisions made were informed by the work of Pyysalo et. al. on their BioInfer corpus (Pyysalo et al., 2007). A detailed annotation guide used in their work was referenced in annotation rules in this work. Other corpora, such as the ones used in Biocreative challenges, have also made valuable contributions to PPI extraction progress (Hakenberg et al., 2010; Krallinger, Leitner, Rodriguez-Penagos, & Valencia, 2008).

Unlike for PPI interaction extraction, there are very limited currently available corpora that can be used for automatic gene-drug interaction extraction system development and evaluation. One corpus that contains those interactions is a 300 sentence corpus by Ahlers et al. (Ahlers, Fiszman, Demner-Fushman, Lang, & Rindflesch, 2007). The Ahlers et. al. corpus include the biological interaction categories of inhibit, and stimulate in addition to interaction annotations for genes and drugs. Our corpus does not contain those additional categories directly, but the interaction words that are annotated in our corpus can indicate such categories as well as others. All in all, our focus was on creating a corpus that could be used for evaluation of basic as well as complex approaches, and allow machine-learning based systems to be trained on it.

Current systems for extracting gene-drug interactions are based on entity co-occurrence and some include matching of relationship terms. Those systems commonly use statistical formulas for ranking the relevance of results. Polysearch, Pharmspresso, and others are examples of such systems (Cheng et al., 2008; Garten & Altman, 2009). Some systems integrate linguistic patterns into their methods, such as those by Coulet et. al. and Tari et. al. (Luis Tari, Jörg Hakenberg, Graciela Gonzalez, & Baral, 2009). The system by Coulet et al. explores the value of dependency graph information for relationship extraction. Another result of Coulet et. al.'s work was the *Phare* ontology that includes concepts relevant to those relationships, which we utilize in this work. The value of such collections of interaction-indicating terms has been highlighted before in the biomedical relationship extraction context (Bui, Nualláin, Boucher, & Sloot, 2010; Chowdhary, Zhang, & Liu, 2009).

[1] http://diego.asu.edu/index.php/projects

3 Materials and Methods

3.1 Corpus design.

The purpose for the creation of the new corpus was to create a resource that NLP developers can use to train and test gene-drug interaction extraction systems. The corpus was based on articles from journals that are known to contain pharmacogenomic relationships. Genes and drugs were automatically tagged and then 551 sentences that contain both a gene and drug were randomly selected for annotation. The corpus and sentence selection process is described in the following subsections.

Journal Selection. A list of journals relevant to pharmacogenomics was generated by extracting the journal names from articles that have been curated in PharmGKB as containing evidence of gene-drug relationships. This list was generated from their downloadable "relationships" file, which contains the abstract IDs of articles with manually curated gene-drug relationships. 591 journal names were obtained this way. The goal of using only those journals is to make the corpus representative of typical sentences containing a gene and drug from literature known to report pharmacogenomic findings.

Sentence processing. All abstracts in PubMed from the relevant journal names were downloaded. A sentence splitter program from OpenNLP was used to extract sentences from the abstracts ("The OpenNLP Homepage," n.d.). A total of 22,601,402 sentences were processed.

Identification of entites. Previous work in pharmacogenomics relationship extraction has shown effective results by classifying relationships after identifying sentences with entities of interest through dictionary matching techniques (Garten & Altman, 2009; Rebholz-Schuhmann et al., 2007). Our work takes a similar approach, but utilizes a machine-learning based method, BANNER, for gene recognition, as it was shown to have better performance than a dictionary-based method (Leaman & Gonzalez, 2008). Drugs were recognized through the use of dictionary matching. The dictionaries used for drugs were based on drug names available at DrugBank. Exact full token matching of drug terms was used to identify them in sentences. Although incorrectly tagged (false entity) genes and drugs were corrected by annotators, they did not add entities missed by NER recognition. A second round of annotation will correct this when we shift focus to NER.

Terms indicative of an interaction for adding to basic co-occurrence relationship extraction were extracted from the *Phare* ontology. The terms acquired were from rdfs labeled text in the "object properties" in the ontology. Object properties are elements of the ontology that describe relationships between classes such as gene and drugs, yielding 168 unique terms after stemming.

Sentence selection. The initial annotation effort that is the focus of this paper was aimed at completing around 500 sentences as a proof of concept, with a total of 1,500 to be completed in the second phase of this project. Random selection of sentences that include a gene and a drug, in contrast to balanced positive and negative selection, was used to make the corpus reflect typical sentences potentially containing an interaction that can be easily extracted from the source articles after simple (drug and gene) concept tagging, which is the most basic approach to interaction extraction. The randomized ratio of positive and negative interactions in the corpus is useful for training classification systems that operate on similarly pre-processed sentences to account for that naturally occurring ratio.

3.2 Annotation.

An annotation tool named STAV was used to create annotations ("stav," n.d.). Customization of the tool was performed to match the types of annotations needed for the corpus. The identified entities were formatted for use with the tool. Annotations created with the tool were stored in the BioNLP shared task file format. That format is compatible with a variety of existing systems for relationship extraction.

Annotation guidelines. Based on a review of literature on related annotation guidelines for relationships such as PPIs, an initial annotation guideline was created based on a small sample of sentences. The guide was iteratively refined through annotation of additional sentences, until considered sufficiently stable for release to additional annotators.

The guideline was refined to achieve a balance of complexity and clarity to assist annotators.

Only a few (5-10) example sentences per annotator have been discussed in person. The explicit written instructions in the guide were relied on more than in-person example sentence discussions to train annotators to handle the complicated content of the corpus and avoid over-influencing the annotators, as noted that is possible with the overuse of those examples (Hovy & Lavid, 2008).

The first annotator, a student with a Bachelor of Science (BS) in Biology, was the main annotator and author of the guidelines. The second and third annotators are PhD students in Biomedical Informatics, the second with a BS in Biology and 10 years nursing experience, and the other with a Bachelor of Technology in Bioinformatics. Weekly annotation meetings were done on individual bases. A short checklist of things to look for in annotations was distributed in addition to the guidelines.

Annotations. The following describes major annotation categories and subcategories in the corpus:

- **Interaction** Genes and drugs are annotated simply as "having an interaction" broadly understood as having an "action, effect, or influence" on each other. All gene-drug interactions annotated must have at least one interaction term that helps explain the interaction. Additional properties that were annotated and a brief explanation of their purpose include:
    - **Direct/Indirect:** Describes the complexity in the interaction statements. An "indirect" interaction is one where the presence of an intermediary entity is needed for semantic understanding of the interaction.
    - **Explicit/Inferred:** Records if an inference had to be made on whether the interaction was present because an interaction was not explicitly stated.
- **Non-interaction**
    - **Shared Entity:** An entity connected to both a gene and a drug that don't interact with each other. In contrast to an intermediary entity.
- **Interaction Term** Terms that are descriptive of the interaction (as defined earlier). These terms are helpful for capturing more specifically the type of interaction present.
- **Intermediary Entity** These are non-gene, non-drug entities that are closely connected to the interaction. They are entities that are needed for understanding of the full semantic meaning of gene-drug interactions. These entities are not annotated themselves but they are used to determine the indirectness property.

Examples of these categories can be seen in the sentence: "Using standard steady-state kinetic analysis, it was demonstrated that *paclitaxel* was a possible uncompetitive inhibitor to NAT activity in cytosols based on the decrease in apparent values of K(m) and V(max)." (PMID: 11955677). This sentence includes an interaction between the drug *paclitaxel* and gene *NAT*. An interaction term that helps establish that the interaction is present is "inhibitor". "Cytosols" is where the NAT inhibition activity can occur and represents an intermediary entity that is needed in the semantic meaning of the interaction.

The broad definition of interaction was used to make progress toward annotations including, and in turn being representative of, the most general form of gene-drug interaction that is described in the source abstracts. We chose to first concentrate on getting good inter-annotator agreement using the general definition before considering additionally annotating specific biological interaction types. Annotated interactions are required to have at least one annotated interaction term (although terms do not have to be from the predefined list) to ensure that specific and identifiable language is present that justifies the annotation.

The subcategories included were added to record the linguistic complexity in which the interactions and non-interactions are described. Recording that complexity can help system developers handle its presence when trying to automatically recognize interaction statements. Additionally, the annotation properties of speculation, negation, and nesting were allowed but not separately annotated in interaction annotations.

Each annotator reported annotation time estimates. Total time spent on annotations including meetings but not other work (e.g. guideline development) was approximately 80 hours for the primary annotator and 20 hours combined for other annotators. Hard sentences to annotate required research into source articles and entities described.

Evaluation of the Corpus. Around 26% of the corpus was annotated by a second and third annotator. A program was created for IAA scoring, accounting for nested entities and equivalent entities including abbreviations. Manual review was used to verify the program's scores. Example sentences from the corpus discussed with annotators were not used for IAA scoring.

3.3 Relationship Extraction methods.

Three basic methods for extracting interactions were implemented for evaluation. The basic method, co-occurrence, is inherent to the corpus as all sentences are selected based on both entities being present in them. Thus, in co-occurrence, any mention of a gene and a drug together in a sentence represents an interaction between those entities.

Co-occurrence plus interaction terms, the second method tried, identifies that interactions are present only when sentences contain an interaction word from a predefined list. The list of interaction terms obtained from the *Phare* ontology was filtered by removing common stop words. Also, a filter was applied to only use terms greater than two letters in size. Those filters were used to avoid unneeded matches from common words.

The linguistic pattern based extraction method developed for this evaluation was based on the work by Coulet et. al. Specific linguistic patterns described in that work were used to classify the presence of interactions between genes and drugs. A program named Graph Spider was used to match the specified patterns within sentences (Shepherd & Clegg, 2008). The Stanford Parser was used to generate dependency graphs for use with the pattern recognition in Graph Spider.

The dependency rules designed by Coulet. et. al. were entered into Graph Spider using the metapattern language (MPL) designed by the Graph Spider authors. MPL is a pattern formalism that can be used to match dependency subgraph patterns in dependency parsed text. After dependency graphs were generated for processing in Graph Spider, text representing genes and drugs in the graphs were converted to general tags for those entity types. Those conversions were made to allow the patterns in MPL to be generalizable.

Java programs were created to reformat and score the subgraph pattern match results made by Graph Spider. Scoring used text character positions (spans) of entities included in annotations. True positives were recorded when pairs of entity spans in Graph Spider subgraph results matched annotated pairs of entity spans labeled as having interactions. False positives and false negatives were similarly assessed using entity spans. A manual evaluation of pattern matched output compared to annotations was performed to ensure accuracy.

A condition applied in the pattern based system was that the patterns can match up to four modifier words for each individual gene and drug in interaction pattern matches. Those words are additional words that modify the meaning of the gene or drug in the interaction. The limit was included for practical reasons, as hand coding of patterns in MPL is complex. The rules described by Coulet et. al. did not specify any limit on modifier words but the difference in results by including a realistic limit is predicted to be negligible.

4 Results

A total of 551 sentences are annotated, with 781 interactions present in them. There are 351 instances of non-interactive entities in the same set. The average length of sentences is 28.1 words. Table 1 describes further properties of the corpus.

Annotation Analysis. The inter-annotator agreement scores are reported as accuracy and Cohen's kappa. Kappa was chosen due to its widespread use and therefore comparability with other work in corpus creation. Accuracy is found by the number of instances agreed on divided by the total instances annotated. A total of 144 sentences were used for the scoring. Annotators 1 and 2, 1 and 3, and 2 and 3 were compared using 92, 52, and 61 sentences respectively. IAA results with the main categories of interaction vs. non-interaction are shown in Table 2.

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

Table 2. Inter-annotator agreement results.

Sentences	Tokens (with punctuation)	Words (tokens with no punctuation)
551	18,585	15,464

Table 1. Statistics describing corpus properties.

IAA scores were found for all annotated subcategories. Those subcategories are DirectExplicit, IndirectExplicit, IndirectInferred for interactions and SharedEntity for non-interactions. Their ranges of scores with all annotator pair groups using accuracy scores are 72-79%, 40-69%, 62-82%, 50-60% and kappa scores are 31-58%, 1-27%, -4-31%, 0-4% respectively. Those scores are created by selecting main category inter-annotator matches (e.g. interaction) and calculating the IAA between the annotated subcategories.

In some sentences, annotators missed doing annotations for gene-drug instances that the other annotator added. IAA scores did not include annotations made by only one annotator. Confirmation with annotators was made that annotations not made were not intended to represent non-interactions. The percentage of missed inter-annotator instances was approximately 20%. Future work will be to improve the inter-annotator annotation process so that those instances are not missed for IAA scoring. While some annotations were missed in IAA scoring, annotations by the primary annotator that are included in the corpus contain all instances (none missed) from the source text to our knowledge.

ID	Contents	Agreement	Sentence text
A	One direct explicit interaction	Y	This suggests that galantamine (GAL), a cholinesterase inhibitor, could be effective when seeking to prolong abstinence in recently detoxified alcoholics. (PMID: 16328375)
B	One indirect explicit and four shared entity non-interactions	Y	They are widely distributed and mediate all of the known biologic effects of angiotensin II (AngII) through a variety of signal transduction systems, including activation of phospholipases C and A2, inhibition of adenylate cyclase, opening of calcium channels, and activation of tyrosine kinases. (PMID: 9892138)
C	One indirect explicit interaction	N	The results of studies of perfused rat hearts with completely inhibited creatine kinase show significantly decreased work capacity and respectively, energy fluxes, in these hearts in spite of significant activation of adenylate kinase system (Dzeja et al. this volume). (PMID: 9746326)

Table 3. Example sentences from the corpus.

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 int. terms	69.60% (664/954)	85.02% (664/781)	76.54%
Pattern-based	96.61% (57/59)	7.30% (57/781)	13.57%

Table 4. 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.

The scoring methods used were instance level scoring instead of sentence level scoring. In the instance level scoring each gene-drug instance counted in performance scores.

A caveat about the pattern-based system scoring should be noted. That caveat was that the Graph Spider software used was unable to process approximately 10% (around 50) of the sentences in the corpus due to errors. The pattern-based system is likely to have scored slightly higher if it could have processed those sentences.

5 Discussion

5.1 Analyses of interaction extraction methods performance.

The f-score of co-occurrence with and without interaction terms showed better performance than the pattern-based interaction extractions, which was expected. Pattern based methods, particularly those where the patterns were manually created, are typically very high in precision and very low in recall, as they are highly dependant on the specific patterns included for recognition. Although recall was low, users who want very high confidence interaction predictions or interactions of a very specific type can benefit from the pattern-based system's demonstrated high precision. Co-occurrence can suit users who want to focus on recall.

Coulet et al. reported their system scored a precision of 70% for exact match and 87.7% for exact or incomplete match but true classification. Our results are similar to their 87.7% results in both percentage and scoring method. The method that allows incompleteness accepts matches that accurately identify core pharmacogenomic relationships but don't need to correctly match modifier words. Our scoring is similar in not needing to match modifier words. The similarity in results indicates that we correctly implemented the system that Coulet et al. designed. That indication does have the limitation that the 10% of sentences unable to be processed may have affected the results.

An example of a more complex interaction that was matched by co-occurrence with an interaction term but not the pattern-based method was "Moreover, S-nitrosylation of thioredoxin was also significantly augmented after atorvastatin treatment." (PMID: 15289372). In that sentence, an interaction occurred where thioredoxin's (gene) S-nitrosylation was augmented by atorvastatin (drug). Analysis of the dependency graphs used by the pattern-based system revealed some reasons why it was unable to identify the interaction.

The pattern-based system uses a rule that applies to that sentence: a potential pattern sequence match can be "interrupted" by a dependency that does not fit accepted patterns. In the non-classified sentence, the entities "was" and "augmented" were terms that caused the pattern matching to be interrupted. Both "was" and "augmented" are not nouns or prepositions. They both also are needed in the dependency subgraph that connects the gene and drug together. Those parts of speech are not allowed to be chained together in the pattern-based system's patterns. That deviation from the allowed patterns caused the system to miss that interaction.

Adding patterns with more diversity in allowed parts of speech in series of interaction terms that connect genes and drugs in interactions can improve recall performance. A review of parts of speech (POS) in missed matches showed that some misses were due to no verb POS tags being present in interaction descriptions. That can occur when verbs are in their nominalized form or other situations. Mining the corpus for both part of speech and dependency graph patterns can identify patterns that are able to correct those misses. Also, the POS tagger included with the parser mistagged a variety of words. Using a higher performance tagger or one trained on biomedical text may help with pattern matches.

Ahlers et. al. also reported relationship extraction performance from a new system with their gene-drug corpus. That system achieved a precision of 73% and recall of 50% extracting an annotation category including gene-drug relationships. The system is built upon an earlier system and an important part of its capabilities comes from specialized linguistic rules it uses. The corpus included in this work can be useful for further development of systems that integrate such rules with other methods to improve extraction performances.

Some characteristics were notable about the results of the methods using co-occurrence with and without interaction terms. The performances found of those methods may be specific to an increased amount of gene-drug interactions found in the journals used compared to other journals. Also, the use of interaction terms from the Phare ontology was expected to increase precision because they were found from predicted pharmacogenomic relationships. The co-occurrence with interaction terms method resulted in only approximately equaling the precision of basic co-occurrence. One possible reason for that is the terms were originally found partly with disease relationships. They therefore can be less relevant to gene-drug interactions.

5.2 Analyses of annotations

Table 2 includes that the general interaction annotations had the kappa values 46%, 30%, 11% which are considered only moderate to low scores by common rating methods. Some IAA scores, such as kappa, include a correction for chance agreement probability. An intentional design choice was made in the corpus to allow an unbalanced but natural ratio of interactions to non-interactions. That imbalance increased kappa's correction. Although our reasonably high IAA scores with accuracy helped increase the kappa score, they were not enough to offset the correction and bring kappa above the moderate score.

An article by Strijbos et. al. states that kappa can have a strict chance agreement correction in the case of few categories (Strijbos, Martens, Prins, & Jochems, 2006). Given that general interaction scores were only based on the categories of present or absent, kappa may have been overly strict with the correction. If that correction in our data is not strict, but justified, than that indicates how further improving our annotation process can be valuable. Further investigation will go into understanding what statistics may be useful for scoring given the corpus properties. Exploration will also continue with talking to annotator s about what may be causing disagreement. That exploration will help reveal ways to improve IAA.

Subcategories showed mixed results in their IAA performances. The subcategories with the highest IAA scores may indicate that those subcategories are more clearly defined than others in the annotation guide.

Reviewing some annotated sentences can help clarify how the IAA results occurred. All annotators agreed the drug galantamine has a direct explicit interaction with cholinesterase in sentence A in Table 3. Such an interaction description is simply described and an annotator has reported that type of interaction being the easiest to identify.

Agreement was found with all annotators for annotations in sentence B in Table 3. It was readily understandable to annotators that calcium and other signal transduction systems do not have an interaction simply for all being a part of those types of systems.

An example of a sentence with annotator disagreement was sentence C in table 3. Although endogenously produced in this case, the nested entity creatine was considered a drug due to being relevant to creatine in its exogenous drug form.

The occurrence of multiple properties, such as inhibition and effects on hearts can make it difficult to follow the logic of the interaction between creatine and adenylate kinase (enzyme). The interaction annotation can be hard for annotators to find due to that complexity and the subtleness of the "in spite of" phrase describing the negated effect between the drug and gene. The interaction is negated but that still is considered an interaction by the annotation rules used.

5.3 Future Work

As mentioned before, the corpus will grow from around 500 sentences that it has right now to around 1,500. The larger the corpus expands to be, the more representative it will become of gene-drug interactions. Other future work includes work with more advanced interaction extraction systems.

Along with this publication, a version of the corpus with high confidence in annotations will be released. Given that this is an initial work, a relatively modest amount of annotation revisions may occur with a few periodic later version releases of the corpus to improve its quality.

Unfortunately no tagger is perfect so as annotations proceed, drugs or genes that were missed by the tagger can be investigated to further understand why that occurred. An example of a commonly missed drug was acetylcholine. Acetylcholine was picked up as a drug if it was spelled out, but not if it was abbreviated as ACh and it is commonly abbreviated.

6 Conclusion

The extraction results indicated that the systems tested can be utilized and built upon according to user preferences in precision, recall, or specific interaction terms. The corpus presented here offers valuable utility to system developers working toward achieving favorable balances of precision and recall in gene-drug interaction extractions. The growth of that corpus will also increasingly benefit the developers working on those extractions. That type of extraction is important to advancing work in pharmacogenomics by retrieving knowledge for individuals working in the field.

Acknowledgements

The authors wish to thank Ehsan Emadzadeh for his help with the annotation tool and Robert Leaman for his help with annotation methods.

References

Ahlers, C., Fiszman, M., Demner-Fushman, D., Lang, F.-M., & Rindflesch, T. (2007). Extracting semantic predications from Medline citations for pharmacogenomics. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 209–220.

Bui, Q.-C., Nualláin, B. O., Boucher, C. A., & Sloot, P. M. A. (2010). Extracting causal relations on HIV drug resistance from literature. *BMC Bioinformatics*, *11*, 101. doi:10.1186/1471-2105-11-101

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., & Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, *36*(Web Server issue), W399–405. doi:10.1093/nar/gkn296

Chowdhary, R., Zhang, J., & Liu, J. S. (2009). Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics (Oxford, England)*, *25*(12), 1536–1542. doi:10.1093/bioinformatics/btp245

Coulet, A., Shah, N. H., Garten, Y., Musen, M., & Altman, R. B. (2010). Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, *43*(6), 1009–1019. doi:10.1016/j.jbi.2010.08.005

Garten, Y., & Altman, R. B. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*, *10 Suppl 2*, S6. doi:10.1186/1471-2105-10-S2-S6

Hakenberg, J., Leaman, R., Vo, N. H., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., et al. (2010). Efficient extraction of protein-protein interactions from full-text articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *7*(3), 481–494. doi:10.1109/TCBB.2010.51

Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., & Klein, T. E. (2002). PharmGKB: The Pharmacogenetics Knowledge Base. *Nucleic Acids Research*, *30*(1), 163–165. doi:10.1093/nar/30.1.163

Krallinger, M., Leitner, F., Rodriguez-Penagos, C., & Valencia, A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, *9 Suppl 2*, S4. doi:10.1186/gb-2008-9-s2-s4

Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 652–663.

Luis Tari, Jörg Hakenberg, Graciela Gonzalez, & Baral, C. (2009). Querying parse tree database of medline text to synthesize user-specific biomolecular networks. CiteSeerX. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.8574

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., & Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, *8*, 50. doi:10.1186/1471-2105-8-50

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., & Stoehr, P. (2007). EBIMed—text Crunching to Gather Facts for Proteins from Medline. *Bioinformatics*, *23*(2), e237–e244. doi:10.1093/bioinformatics/btl302

Sconce, E. A., Daly, A. K., Khan, T. I., Wynne, H. A., & Kamali, F. (2006). APOE genotype makes a small contribution to warfarin dose requirements. *Pharmacogenetics and Genomics*, *16*(8), 609–611. doi:10.1097/01.fpc.0000220567.98089.b5

Shepherd, A. J., & Clegg, A. B. (2008). Syntactic pattern matching with GraphSpider and MPL. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine SMBM 2008 Turku Finland*, 129–132.

stav. (n.d.).*GitHub*. Retrieved March 26, 2012, from https://github.com/TsujiiLaboratory/stav

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education*, *46*(1), 29–48. doi:10.1016/j.compedu.2005.04.002

T1.pdf. (n.d.). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/workshops/T1.pdf

The OpenNLP Homepage. (n.d.). Retrieved March 26, 2012, from http://opennlp.sourceforge.net/projects.html

Wishart, D. S. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(90001), D668–D672. doi:10.1093/nar/gkj067