THE EXTRACTION OF PHARMACOGENETIC AND PHARMACOGENOMIC RELATIONS – A CASE STUDY USING PHARMGKB

EKATERINA BUYKO, ELENA BEISSWANGER, and UDO HAHN

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
E-mail: {*surname.name*}*@uni-jena.de*
http://www.julielab.de

In this paper, we report on adapting the JREX relation extraction engine, originally developed for the elicitation of protein-protein interaction relations, to the domains of pharmacogenetics and pharmacogenomics. We propose an intrinsic and an extrinsic evaluation scenario which is based on knowledge contained in the PHARMGKB knowledge base. Porting JREX yields favorable results in the range of 80% F-score for *Gene-Disease*, *Gene-Drug*, and *Drug-Disease* relations.

Keywords: Information Extraction, Pharmacogenetics, Pharmacogenomics, PharmGKB

1. Introduction

While molecular biology and clinical medicine have pursued their research agenda in a fairly isolated manner for a long time in the past, this attitude has changed dramatically in the last years and new research themes have been defined which combine the efforts of both scientific camps. This change is mirrored in terms such as 'translational medicine', an area that is explicitly devoted to bringing results from genetic research labs to the bedside in a clinical environment, or 'personalized medicine' which aims at exploiting genetic data from an individual (or a genetically homogeneous cohort) for use in fine-tuned drug dosage recommendations or highly individualized drug development and risk assessments.

Unfortunately, this thematic shift has not found parallels in the field of BioNLP, up until now. Still, the overwhelming amount of work is concerned with proteins and genes (for named entity tagging) and protein-protein interactions (for relation extraction). Although this research has by no means become obsolete or irrelevant (quite the contrary is true given the current performance ceilings of state-of-the-art recognition devices), we here argue for complementing these activities by research efforts that are intended to support the collaborative work of molecular biologists and clinical physicians in terms of suitable information extraction technology. Our special targets are the domains of *pharmacogenetics* (studying the influence of single genes on drug response) and *pharmacogenomics* (studying the influence of multiple genes on drug response, typically using high-throughput techniques).

In order to contribute to this goal, we here consider possible association relations between genes, drugs, and diseases as reported in the scientific literature. Accordingly, we no longer focus on *Protein/Gene-Protein/Gene* relations but extend the scope of our investigations to cover *Gene-Drug*, *Gene-Disease*, and *Drug-Disease* relations. Such a fundamental change in the scope of entities and relations under scrutiny poses new problems. Both, for training (machine learning-based taggers) as well as testing suitable gold standards have to be provided. Since it is not an easy task to set up such metadata, we here build on previous work, the Phar-

macogenomics and Pharmacogenetics Research Network and Knowledge Base[a] (PharmGKB) (see Section 2 for a detailed description). Given this repository, we automatically generate gold standard data out of the knowledge provided by PharmGKB (see Section 4.1), and use it for two evaluation scenarios, an *intrinsic* and an *extrinsic* one (see Section 5.1). Based on experiments using our relation extraction engine JReX (see Section 4.3), we have evidence that the thematic extension of named entities as well as relations yields favorable results for the three types of relations being gleaned on, in the range of 80% F-score for *intrinsic* settings and up to 77.5% in an *extrinsic* configuration (see Section 5.2).

2. PharmGKB Database

The development of the PharmGKB repository represents a major step towards an interdisciplinary biomedical information store. PharmGKB incorporates data on genetic variations and associated phenotypic manifestations. The latter mainly concern the pharmacokinetics of therapeutic drugs (how drugs are absorbed, metabolized and excreted by an organism) and the pharmacodynamics of drugs (how drugs act in an organism). The repository also covers certain non-pharmacological aspects of phenotypes, including susceptibility to disease. Currently (as of May 2011) PharmGKB contains information about 27,000 pharmacogenes, 2,500 drugs, 3,200 diseases and 23,870 relations between them (amongst others) and it continues to grow as PharmGKB curators screen the scientific literature for new facts to be inserted. Yet, keeping up with the large amount of articles published every day is virtually impossible. Thus, the development of automatic support tools for the curation process might be a rewarding endeavor. The key task will be to develop a system that is able to recognize genes, drugs and diseases in text and to detect PharmGKB-relevant relations between them.

Other than common relation extraction tasks in the biomedical domain, such as the detection of protein-protein interactions or gene regulatory relations, PharmGKB-relevant relations draw on much broader types of semantic relationships. This may be due to the fact that the focus of interest in pharmacogenetics and pharmacogenomics lies on links between genetic and phenotypic variations. Yet, evidence for these links can originate from different sources, ranging from genetic research papers to clinical studies. Another reason might be that while genes are referred to by a specific name or symbol in text, references to phenotypes are verbalized much more loosely. At the same time, phenotype variation descriptions often mention the drugs and diseases involved. Now, if *Gene-Drug* or *Gene-Disease* relations are extracted from text, this opens up a wide range of interpretations of possible association types involved, from a real physical interaction to correlation with clinical outcomes. In contrast, relations between drugs and diseases often (but not always) adhere to some kind of *treated-by* relation. Given this underspecification of PharmGKB semantic relationships, we consider PharmGKB relations as coarse association relations, or merely relations, in the rest of this paper.

The following examples illustrate the association relations we here focus on:

- ***Gene-Drug*** — "*It was found that the genotype of* ***CYP2C19*** *had a significant effect on the N-demethylation of* ***citalopram***."

[a] `http://www.pharmgkb.org/`

- ***Gene-Disease*** — "*The urocortin (UCN) gene resides at this interval, and its protein decreases appetite behavior, suggesting that* ***UCN*** *may be a candidate gene for susceptibility to* ***obesity***."
- ***Drug-Disease*** — "*Decreased expression of BRCA2 mRNA predicts favorable response to* ***docetaxel*** *in* ***breast cancer***."

The first example contains a mention of the association between the *CYP2C19* gene and the demethylation process of the drug *citalopram*. In the second example, an assumption about the impact of the *UCN* gene on *obesity* is made. The third example contains facts about the role of the gene *BRAC2* in response to the drug *docetaxel* related to *breast cancer*, thus we encounter a relationship between a drug and a disease.

PharmGKB provides a hierarchically organized category system for different levels of evidence that also classify the significance of the data. The lowest category is "Genotype" (comprising simple genetic variations), the highest "Clinical Outcome" (covering clinically relevant data). In between are the categories "Pharmacodynamics and Drug Response", "Pharmacokinetics", and "Molecular and Cellular Functional Assays". In our relation extraction approach, we do not distinguish between those categories for relations but rather focus on the extraction of generic *Gene-Drug*, *Gene-Disease* and *Drug-Disease* relations for the evaluation of the results on the available PharmGKB relationship data set.

3. Related Work

In BioNLP, research on (binary) protein-protein interactions (PPIs) was and still is predominant. This is reflected in lots of PPI-annotated corpora (e.g., LLL,[1] AIMed,[2] or BioInfer[3]), but also the large variety of methodologies to tackle this problem (pattern-based (e.g.,[4]), rule-based (e.g.,[5]), and machine learning-based approaches (e.g.,[6])). Binary PPIs constitute a fairly general abstraction from the complex interactions between genes and proteins, so requests for finer-grained representations were issued. The Genia event corpus[7] and the BioNLP 2009 Shared Task data[8] contain such detailed annotations of PPIs (amongst others). The BioNLP Shared Task was a first step towards the extraction of specific pathways with precise information about the molecular events involved. The winner system, Turku,[9] achieved with 51.95% F-score the milestone result in that competition, but was outperformed in the BioNLP 2011 Shared Task with 56% F-score.[10]

Considering relation extraction (RE) in the pharmacogenetics and pharmacogenomics domain, to the best of our knowledge, there are only few studies which deal primarily with phenotype-genotype-drug relations. This may be due to the fact that no large-scale annotated corpora are available, up until now, for phenotype-genotype-drug relationships. Thus, the developed systems mostly focused on named entity recognition (gene, drug, and disease names) and on small-scale evaluations for RE. Rindflesch *et al.* developed the Edgar system for the extraction of gene and drug names and gene-drug relations relevant for cancer.[11] Edgar exploits underspecified syntactic parse trees and applies syntactico-semantic rules for the extraction of relationships. An explicit evaluation for the Edgar system is missing though. Chang and Altman's system recognizes relations between genes and drugs in PubMed abstracts with a co-occurrence-based approach.[12] They further classify the relations into five

categories, as specified by PharmGKB, using a Maximum Entropy based machine learning approach. The relation recognition step is evaluated against a small data set of 215 gene-drug relations manually extracted from a review article, while the classification step is assessed against human-curated articles from PharmGKB. Evaluation results for all five categories range from 88% recall with 75% precision for the predictions of *Pharmacokinetics*, to 9% recall with 27% precision for the *Clinical Outcome* category. The authors concede that the selected PharmGKB data set was small, including 325 gene-drug pairs, and that the evaluation results heavily depend upon the size of training data. Chun *et al.* describe a system for disease-gene relation extraction that is based on the co-occurrence of gene and disease name mentions (found via dictionary look-up) and additional filtering of false positives with a machine learning classifier.[13] In the filtering mode, the system achieves 78.5% precision and 87.1% recall on a manually annotated corpus with 1,000 co-occurrences of gene and disease names.

The most recent systems for the extraction of gene-drug relationships are Pharmspresso[14] and GenDrux.[15] Pharmspresso builds on the Textpresso tool,[16] a full-text search engine for biological entities and facts such as PPIs. Pharmspresso has been extensively evaluated concerning the detection of gene and drug names, but with respect to relationships it yields only 50% recall on gene-drug *'association'* relations on 45 full-text articles which contain 178 gene name mentions and 142 drug names mentions. GenDrux is a web-based tool developed for the analysis of documents in the breast cancer domain. Its document collection consists of 4,000 PubMed abstracts collected using gene and drug name filters. GenDrux's focus is on the retrieval of documents with gene and drug names related to breast cancer, while RE is based only on the co-occurrence of relevant terms in the titles of documents; an evaluation, however, is lacking. A first large-scale evaluation study was carried out by Coulet *et al.* who extracted PharmGKB-relevant relationships using a lexicon of key phramocogenomic entities and syntactic parses of 17 million Medline abstracts.[17] The extracted relationships are reported to have a precision up to 87.7%. Still this work does not evaluate the recall of the presented relationship extraction.

With the exception of Coulet *et al.*, we here provide the first large-scale evaluation study of phenotype-genotype-drug relationship extraction for *Gene-Drug*, *Gene-Disease* and *Drug-Disease* relations, using a high-performance relation extractor. We automatically gather a large-scale gold standard based on human-curated texts from the PharmGKB database. Furthermore, we provide two evaluation scenarios, an *intrinsic* one based on corpus cross-validation and an *extrinsic* one based on PharmGKB relationship data.

4. Methods

4.1. PharmGKB *as Gold Standard*

Usually, manually annotated gold standard corpora are used to evaluate RE systems. Manual annotation, however, is a time-consuming and costly process. In contrast, we here capitalize on previous curation efforts and derive large-scale gold data for genes, drugs, diseases, and

binary relationships among them automatically from the PharmGKB knowledge base.[b] For all relationships, PharmGKB specifies the two participants' names and IDs and zero to many references to PubMed abstracts that provide textual evidence for each relation. In total, references to 5,241 different PubMed abstracts are given. We extracted all available abstracts from the Medline Baseline Repository 2011 to form our initial corpus.[c] Next we limited our focus to those PharmGKB relations that hold between entities of different semantic types (*Gene-Drug*, *Gene-Disease*, *Drug-Disease*). For each entity involved in a relation we compiled a list of its names and all alternative names, plus for genes its symbol and all alternative symbols, as specified in PharmGKB. Finally, for each relation, we matched the names of its participants (case-insensitively) against the abstracts from the referenced PubMed abstracts. Matches covering partial tokens were skipped in this process to avoid false positives. If in a sentence at least one name of the first participant and at least one name of the second participant matched, we marked these sentences as containing a gold PharmGKB relation, and incorporated the abstract as a gold corpus document. The numbers of the unique relations we detected in this way are specified in Table 1.

Table 1. First line: The number of distinct binary *Gene-Drug*, *Gene-Disease*, *Drug-Disease* relations as specified in PharmGKB. Second line: The subset of relations for which at least one PubMed reference is given in PharmGKB. Third line: The fraction of relations with PubMed reference that could be retrieved by our extraction machinery from at least one of the specified PubMed abstracts.

# PharmGKB Relations	*Gene-Drug*	*Gene-Disease*	*Drug-Disease*	Total
all	11,476	8,028	2,639	22,143
with PubMed reference	6,628	7,163	2,634	16,425
retrieved from PubMed abstract(s)	1,686	1,711	673	4,070

The gold corpus we collected using PharmGKB references contains 1,980 PubMed abstracts, where 522 abstracts incorporate *Drug-Disease* relations, 1,414 abstracts hold *Gene-Disease* relations, and 1,262 abstracts contain *Gene-Drug* (see Table 2). In a later step, the collected corpus of 1,980 abstracts, called here gold PharmGKB corpus, was used for the evaluation of our relation extraction approach (see Section 5).

Table 2. Overvuew of gold relations in the gold PhramGKB corpus.

Relation	Abstracts With Gold Relation Annotations	Gold Relation Annotations
Gene-Drug	1,262	9,914
Gene-Disease	1,414	6,626
Drug-Disease	522	3,439
Total	1,980	19,979

[b]The data files genes.zip, drugs.zip and relationships.zip were downloaded on May 17 and diseases.zip on May 18, 2011, from `http://www.pharmgkb.org/resources/downloads_and_web_services.jsp`.
[c]All but six of the referenced PubMed abstracts could be retrieved from the Baseline Repository 2011.

4.2. *Recognition of Relevant Named Entities (Relation Participants)*

For each entity type (gene, drug, disease) we compiled dictionaries from PharmGKB exploiting preferred names and alternate names of entities, as well as preferred symbols and alternate symbols for genes. The drug dictionary was further extended by terms taken from the *Orange Book Dictionary*[d] extended with MeSH[e] term variants, while the disease dictionary was also further extended by headings plus alternate entry terms from the MeSH disease branch (starting with the top node "Diseases [C]"). For gene mention recognition, we used GeNo[18] and for the remaining NER tasks we applied the Lingpipe Chunker.[f]

4.3. *Extraction of PharmGKB-relevant Relations — JReX*

The relation extraction experiments were run with the relation and event extraction system JReX (Jena Relation eXtractor). Generally speaking, the JReX system classifies pairs of genes in sentences as *interaction pairs* using various forms of syntactic and semantic evidence (see Buyko *et al.*[19] for a deeper account). JReX (under the name of JulieLab) scored on 2nd rank among 24 competing teams in the *BioNLP 2009 Shared Task on Event Extraction*, with 45.8% precision, 47.5% recall and 46.7% F-score. After the competition, this system was further streamlined and now peaks at 57.6% precision, 45.7% recall and 51.0% F-score (,[2021]), and thus considerably narrowed the gap to the winner of the BioNLP'09 Shared Task who scored at 51.95% F-score.[g]

As far as pre-processing is concerned, JReX uses JCore tools[23] such as JulieLab's sentence splitter and tokenizer. For shallow syntactic analysis it applies the OpenNLP POS Tagger and Chunker, both re-trained on the Genia corpus. For dependency parsing, the MST parser[24] was retrained on the Genia Treebank and the parses subsequently converted to the CoNLL'07 representation.[h]

The JReX relation extractor accounts for two major subtasks – first, the structural trimming of dependency graphs, and, second, the identification and ordering of arguments for the relation under scrutiny. Trimming dependency graphs amounts to eliminating informationally irrelevant and to enriching informationally relevant lexical nodes by concept overlays. For example, JReX prunes auxiliary and modal verbs which govern the main verb in syntactic structures such as passives, past or future tense. Accordingly, (see Figure 1), the verb "*activate*" is promoted to the Root in the dependency graph and governs all nodes that were originally governed by the modal "*may*". An example of semantic enrichment is also given in Figure 1, where the lexical item *"TNF-alpha"* is turned into the conceptual label *Gene*. This abstraction avoids over-fitting of dependency structures for the machine learning mechanisms on which JReX is based. For a more detailed explanation and evaluation of this approach to syntactic simplification and semantic decoration, see Buyko *et al.*[19] and.[21]

---

[d] `http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm`

[e] `http://www.nlm.nih.gov/mesh/`

[f] `http://alias-i.com/lingpipe/`

[g] The Turku system was also improved after the competition and now performs at 52.9 F-score.[22]

[h] We used the Genia Treebank version 1.0, available from `http://www-tsujii.is.s.u-tokyo.ac.jp`. The conversion script is accessible via `http://nlp.cs.lth.se/pennconverter/`.

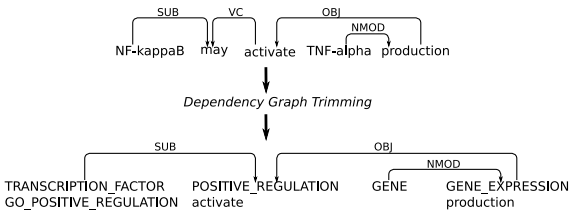

Fig. 1. Trimming of dependency graphs.

The principal relation extraction step is argument detection and ordering. For argument extraction, JReX builds sentence-wise pairs of putative arguments (named entities) and applies two machine learning approaches, one feature-based and the other one kernel-based. In the feature-based classifier, three groups of features are distinguished: (1) *lexical* features (covering lexical items before, after and between the mentions of relation arguments; (2) *chunking* features (concerned with head words of the phrases between two mentions; (3) *dependency parse* features (considering both the selected dependency levels of the arguments, parents and least common subsumer, as well as a shortest dependency path structure between the arguments for *walk* features). For this feature-based approach, the Maximum Entropy (MaxEnt) classifier from Mallet is used.[i]

The graph kernel classifier uses a converted form of dependency graphs in which each dependency node is represented by a set of labels associated with that node. The dependency edges are also represented as nodes in the new graph such that they are connected to the nodes adjacent in the dependency graph. Subgraphs which represent, e.g., the linear order of the words in the sentence can be added, if required. The entire graph is represented in terms of an adjacency matrix which is further processed to contain the summed weights of paths connecting two nodes of the graph (see Airola *et al.*[6] for details). For the graph kernel approach, the LibSVM Support Vector Machine is used as classifier.[j]

5. Experiments and Results

5.1. *Experimental Settings*

We established two evaluation settings, an *intrinsic* and an *extrinsic* one. In the intrinsic, corpus-based scenario we used the gold PharmGKB corpus introduced in Section 4.1 for a cross-validation of JReX. In the extrinsic scenario we used the PharmGKB relationship data for evaluation of our relation extraction pipeline.

In the corpus-based evaluation scenario we performed a 10-fold cross-validation using JReX. In order to analyze the corpus settings, we provide here further information on the number of positive and negative instances (see Table 3).[k] The ratio of negative to positive instances is different for genotype/phenotype-drug and for the genotype-phenotype relations.

[i] `http://mallet.cs.umass.edu/index.php/Main_Page`

[j] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[k] Instances were generated sentence-wise. Positive instances contain pairs of entity mentions (in a sentence), which were annotated to stand in a relation in PharmGKB, negative instances contain pairs of entity mentions (in a sentence), which were not annotated to stand in a relation in PharmGKB.

Table 3. Positive and negative training instances in the gold PHRAMGKB corpus.

Relation	Positive Instances	Negative Instances	Total Instances
Gene-Drug	9,914	7,060	16,974
Gene-Disease	6,626	6,664	13,290
Drug-Disease	3,439	2,468	5,907

For example the ratio for *Gene-Drug* and *Drug-Disease* relations is 1.4, while for *Gene-Disease* relations it is about 1.0. These are fairly low numbers in comparison with the ratio of negative to positive instances in the known relationally annotated corpora which provide protein-protein or gene regulatory relations (the ratio here ranges between 5 and 10). This can be explained by the fact that when entities such as drugs, diseases and genes are mentioned in the same sentence they are likely to form a relation as well. Furthermore, our corpus, although annotated with rich dictionaries and GENO tools, contains a high number of gold PharmGKB entities and thus PharmGKB relations. The distribution of positive and negative instances of a corpus clearly influences the overall evaluation figures in terms of F-score. This is the reason why we decided on an additional, extrinsic evaluation of JREX.

In the extrinsic approach, we extracted from the PHARMGKB relationship subset data as presented in Table 1 (4,070 unique relations) about one third of unique relations for each relation type. For this set of relations, we collected the referenced abstracts from our gold PHARMGKB corpus. Thus, for every relation type we generated two splits of the PHARMGKB corpus, the training and the test part as depicted in Table 4. In contrast to the intrinsic evaluation, we focused here on the extraction of unique relations from the test corpus. Furthermore, the training corpus was not allowed to contain the PHARMGKB relations mentioned in the test corpus. The training corpus subset was used for training JREX on the corresponding relation type. The re-trained version of JReX was then used to predict relations on the test part. We generated a list of detected unique relations and compared it against the test relation subset. The figures for the test and training parts in terms of abstracts and unique relations are presented in Table 4.

Table 4. Test and train parts of the gold PHARMGKB corpus (unique relations only).

Relation	Test		Train		Overall	
	Abstracts	Relations	Abstracts	Relations	Abstracts	Relations
Gene-Drug	368	621	1612	1,068	1,980	1,686
Gene-Disease	513	673	1467	1038	1,980	1,711
Drug-Disease	169	198	1811	475	1,980	673

5.2. *Evaluation on the complete gold* PharmGKB *corpus*

The gold PHARMGKB corpus of 1,980 abstracts was used for a 10-fold cross-validation of our relation extraction machinery. As a lower bound to estimate performance, we selected the common co-occurrence approach, i.e., if two entities of interest co-occur in a sentence, they are marked to stand in a relation. In our case, it was particularly interesting to see the baseline results as the ratio between negative and positive instances is lower as, for example, for protein-protein interaction relations and, hence, has a potential to boost co-occurrence figures.

The results are presented in Table 5. The co-occurrence approach allows to extract *Gene-Drug* and *Drug-Disease* relations with about 73.0% F-score, while *Gene-Disease* relations perform at a lower 65.9% F-score.

Table 5. 10-fold cross-validation on PharmGKB corpus

Relation	Co-occurrence			JReX (FB)			JReX (GK)		
	R	P	F	R	P	F	R	P	F
Gene-Drug	100	57.5	73.0	85.1	79.7	**82.3**	77.6	67.4	72.1
Gene-Disease	100	49.2	65.9	76.9	75.2	**76.0**	71.0	65.7	68.2
Drug-Disease	100	58.2	73.5	81.2	77.0	**79.0**	75.3	67.2	71.0
All relations	100	54.5	70.6	82.3	78.0	**80.1**	71.0	65.7	68.2

In the next step, we evaluated JReX in its feature-based (JReX FB) and its graph kernel-based classifier variant (JReX GK). The results are also contained in Table 5. In the feature-based approach, we achieve for *Gene-Drug* relation detection a performance of 82.3% F-score, with 85.1% recall and 79.7% precision. For the *Drug-Disease* relations 79.0% F-score with 81.2% recall and 77.0% precision were determined. The lowest results are achieved for the *Gene-Disease* relation with 76.0% F-score. When JReX learns all relations in one step and performs 3-types classification at once, the overall performance for all three relations settles at 80.1% F-score. This means that learning all three relations in one step is advantageous for achieving a good overall F-score result. The performance of the graph-kernel based classifier (JReX GK) is lower than for the feature-based classifier with up to 10 percentage points difference in F-score results. The feature-based classifier is shown to perform better for all phenotype/genotype-drug relation types than the JReX GK classifier. While the main source of information for the JReX GK classifier are dependency parse trees, the JReX FB classifier considers a range of lexical, morpho-syntactic and syntactic information. It seems that the extraction of phenotype/genotype-drug association relations cannot be captured by considering dependency parse trees only and profits from richer sources of evidence.

Our intrinsic evaluation showed that the JReX based extraction of phenotype/genotype-drug relations outperform significantly the co-occurrence based approach by up to 9.3 percentage points in terms of F-score (see *Gene-Drug* relation evaluation in Table 5). This indicates that this type of relations can effectively be learned by the JReX tool.

5.3. *Evaluation against the* PharmGKB *relation subset*

In Table 4 we presented the figures of test and training parts for each relation type of interest. As the feature-based classifier was shown in the intrinsic evaluation to outperform the graph kernel classifier-based one, for the extrinsic evaluation we used only the feature-based JReX variant. For each relation type, we retrained JReX FB on the training part and performed relation extraction on the test part. The extracted list of unique entities was then compared with the gold PharmGKB relation list as described previously. In this evaluation, we chose the co-occurrence approach as the baseline, once again. The results are presented in Table 6.

The co-occurrence approach performs with a very low precision of 39.4% for the *Gene-*

Table 6. Evaluation against PharmGKB relation test sets

Relation	Test Size (Unique)	Co-occurrence			JReX		
		R	P	F	R	P	F
Gene-Drug	621	100	39.4	56.5	90.9	61.9	**73.6**
Gene-Disease	673	100	35.0	52.1	83.3	58.6	**68.8**
Drug-Disease	198	100	45.3	62.3	85.3	71.0	**77.5**
Total	1,492	100	39.9	59.9	86.5	63.8	**73.3**

Drug, 35.0% for *Gene-Disease* and 45.3% for *Drug-Disease* relations. The overall F-score results range between 52.1% for the *Gene-Disease* relation and 62.3% for the *Drug-Disease* relation. The co-occurrence approach is outperformed by JReX in the extrinsic evaluation as in the intrinsic one. JReX peaks for the *Drug-Disease* relation at 77.5% F-score followed by 73.6% F-score for the *Gene-Drug* relation and 68.8% F-score for the *Gene-Disease* relation. In all three extrinsic evaluation results for JReX we see that recall is much higher than precision. This can be explained by the distribution of the positive and negative instances in the training data. The lower the ratio between the positive and the negative mentions is, the higher is the tendency of the classifier to classify an instance as a positive. Still, the precision and thus the overall F-score results show that phenotype/genotype-drug relations can successfully be extracted with JReX, with performance peaking up to 77.5% F-score in a real life extrinsic evaluation on the PharmGKB relationship data.

5.4. *Caveats – Putting Things in Perspective*

The main concern with the approach above is clearly centred around the reliability and representativeness of the semi-automatically generated gold standard. The collection of PharmGKB referenced abstracts and the automatic mapping of relationship data in the text may admittedly hide risks of running into bad gold standard data, which is not approved by human annotators. Still, the PharmGKB references to the abstract texts are a human-curated highly reliable data source for relations that hold between entities of interest.

A manual analysis of selected texts revealed that the annotations seem to reflect the knowledge represented by the PharmGKB. One student of biology analyzed 200 sentences randomly extracted from the PharmGKB gold corpus, which should contain gold PharmGKB relations (100 sentences for *Drug-Disease*, 50 sentences for *Gene-Drug* and 50 sentences for *Gene-Disease* relations). The analysis revealed that 80% of the sentences for *Drug-Disease* relations, and 90% of the sentences for the *Gene-Disease* as well as *Gene-Drug* relations, in effect, contain descriptions of these relations. The figures of the distribution of positive and negative training instances are also an indicator for the reasonable quality of the semi-automatically generated gold standard. Furthermore, numerous previous studies on learning relations from automatically generated corpora in a *distant supervision* mode[1] reinforce the outcomes of our approach (see Mintz *et al.*[25]).

[1]A distant supervision approach produces a large automatically annotated corpus using relation mentions from available databases, it considers all sentences containing those entities to stand in a relation in order to train a relation classifier.[25]

To avoid the intricacies of an intrinsic corpus-based evaluation, we exposed our approach to an extrinsic evaluation similar in spirit to the one we had already carried out on the RegulonDB database.[26] The results, up to 77.5% F-score, are promising and reflect the fact that the PharmGKB gold standard may indeed be representative for general genotype-phenotype-drug relationships. But as the PharmGKB curation process seems to involve some automatic preprocessing of the text data for named entity recognition and co-occurrence analysis, the database may lack articles that will not pass the pre-selection step. Furthermore, the PharmGKB database does not contain references to curated articles which do not contain PharmGKB-relevant relations (those are dropped after the curation process). These restrictions may explain the low ratio of negative to positive instances in the generated corpus.

6. Conclusions

While the BioNLP community's focus is still almost exclusively gene/protein-centered, both in terms of named entity and relation extraction (PPIs, in particular), this only partially matches the most recent needs at the intersection of molecular biology and clinical/pharmacological research. Our work reported in this paper aims at mitigating that mismatch.

Accordingly, we ported JULIE Lab's high-performance JReX relation extractor from the protein/gene interaction domain proper to the domains of pharmacogenetics and pharmacogenomics. Here, three novel types of relations, namely *Gene-Drug*, *Gene-Disease*, and *Drug-Disease* relations had to be targeted. For these relations we achieved over-all F-scores on the order of 80% in an intrinsic and 73% in an extrinsic evaluation. In both cases co-occurrence-based baselines were clearly outperformed. Our approach crucially relies upon training and test data that we automatically compiled from PharmGKB.

Despite these encouraging results, our experimental design needs further refinement. First, our triple relation repertoire cannot be matched straightforwardly with the PharmGKB-specific relation hierarchy (ranging from 'Genotype' to 'Clinical Outcome'). However, such a mapping is needed for directly supporting the curators of PharmGKB. Second, our internal evaluation is not based on a real gold standard but on an automatically generated substitute only. This requirement opens up the box of Pandora in that not only rarely dealt with entities (diseases, drugs, etc.) but even worse rather underspecified, if not fuzzy relations (such as 'Clinical Outcome') have to be served by massive annotation efforts.

Acknowledgments. This work is funded by a grant from the German Ministry of Education and Research (BMBF) for the *Jena Centre of Systems Biology of Ageing* (JenAge) (grant no. 0315581D).

References

1. C. Nédellec, Learning Language in Logic: Genic interaction extraction challenge, in *Proceedings of the 4th Learning Language in Logic Workshop*, 2005.
2. R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani and Y. W. Wong, *Artificial Intelligence in Medicine* **33**, 139 (2005).
3. S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Jarvinen and T. Salakoski, *BMC Bioinformatics* **8** (2007).
4. C. Blaschke, M. A. Andrade, C. A. Ouzounis and A. Valencia, Automatic extraction of biological information from scientific text: Protein-protein interactions, in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, 1999.

5. K. Fundel, R. Küffner and R. Zimmer, *Bioinformatics* **23**, 365 (2007).
6. A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter and T. Salakoski, A graph kernel for protein-protein interaction extraction, in *Proceedings of the BioNLP 2008 Workshop on Current Trends in Biomedical Natural Language Processing*, 2008.
7. J.-D. Kim, T. Ohta and J. Tsujii, *BMC Bioinformatics* **9** (2008).
8. J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, Overview of BioNLP'09 Shared Task on Event Extraction, in *Proceedings BioNLP 2009. Companion Volume: Shared Task on Event Extraction*, 2009.
9. J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala and T. Salakoski, Extracting complex biological events with rich graph-based feature sets, in *Proceedings BioNLP 2009. Companion Volume: Shared Task on Event Extraction*, 2009.
10. S. Riedel, D. McClosky, M. Surdeanu, A. McCallum and C. D. Manning, Model combination for event extraction in BioNLP 2011, in *Proceedings BioNLP 2011 Shared Task Workshop*, 2011.
11. T. C. Rindflesch, L. Tanabe, J. N. Weinstein and L. Hunter, EDGAR: Extraction of drugs, genes and relations from the biomedical literature, in *Proceedings of the 2000 Pacific Symposium on Biocomputing*, 2000.
12. J. T. Chang and R. B. Altman, *Pharmacogenetics* **14**, 577 (September 2004).
13. H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki and J. ichi Tsujii, Extraction of gene-disease relations from medline using domain dictionaries and machine learning, in *Proceedings of the 2006 Pacific Symposium on Biocomputing*, 2006.
14. Y. Garten and R. B. Altman, *BMC Bioinformatics* **10** (2009).
15. C. Crasto, D. Luo, F. Yu and D. Chen, *BMC Medical Informatics and Decision Making* **11** (2011).
16. H.-M. Mueller, E. E. Kenny and P. W. Sternberg, *PLoS Biology* **2**, 1984 (2004).
17. A. Coulet, N. H. Shah, Y. Garten, M. A. Musen and R. B. Altman, *Journal of Biomedical Informatics* **43**, 1009 (2010).
18. J. Wermter, K. Tomanek and U. Hahn, *Bioinformatics* **25**, 815 (2009).
19. E. Buyko, E. Faessler, J. Wermter and U. Hahn, Event extraction from trimmed dependency graphs, in *BioNLP 2009. Companion Volume: Shared Task on Event Extraction*, 2009.
20. E. Buyko and U. Hahn, Evaluating the impact of alternative dependency graph encodings on solving event extraction tasks, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.
21. E. Buyko, E. Faessler, J. Wermter and U. Hahn, *Computational Intelligence* **27** (2011).
22. J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala and S. Tapio, *Computational Intelligence* **27** (2011).
23. U. Hahn, E. Buyko, R. Landefeld, M. Mühlhausen, M. Poprat, K. Tomanek and J. Wermter, An overview of JCoRe, the JULIE lab UIMA component repository, in *Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, 2008.
24. R. T. McDonald, F. Pereira, K. Ribarov and J. Hajic, Non-projective dependency parsing using spanning tree algorithms, in *Proceedings of the 2005 Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, 2005.
25. M. Mintz, S. Bills, R. Snow and D. Jurafsky, Distant supervision for relation extraction without labeled data, in *Proceedings of the 2009 Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009.
26. U. Hahn, K. Tomanek, E. Buyko, J. J. Kim and D. Rebholz-Schuhmann, How feasible and robust is the automatic extraction of gene regulation events? A cross-method evaluation under lab and real-life conditions, in *Proceedings of the BioNLP 2009 Workshop*, 2009.