# Weakly Labeled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction

**Philippe Thomas**[1*], **Tamara Bobić**[2,3*], **Martin Hofmann-Apitius**[2,3], **Ulf Leser**[1], **Roman Klinger**[2]

[1]Institute for Computer Science
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany

[2]Fraunhofer Institute for Algorithms
and Scientific Computing (SCAI)
Schloss Birlinghoven
53754 Sankt Augustin
Germany

[3]Bonn-Aachen Center for
Information Technology (B-IT)
Dahlmannstraße 2
53113 Bonn
Germany

{tbobic,klinger,hofmann-apitius}@scai.fraunhofer.de
{thomas,leser}@informatik.hu-berlin.de

## Abstract

Relation extraction is frequently and successfully addressed by machine learning methods. The downside of this approach is the need for annotated training data, typically generated in tedious manual, cost intensive work. Distantly supervised approaches make use of weakly annotated data, which can be derived automatically. Recent work in the biomedical domain has applied distant supervision for protein-protein interaction (PPI) with reasonable results, by employing the IntAct database. Training from distantly labeled corpora is more challenging than from manually curated ones, as such data is inherently noisy. With this paper, we make two corpora publicly available to the community to allow for comparison of different methods that deal with the noise in a uniform setting. The first corpus is addressing protein-protein interaction (PPI), based on named entity recognition and the use of IntAct and KUPS databases, the second is concerned with drug-drug interaction (DDI), making use of the database DrugBank. Both corpora are in addition labeled with 5 state-of-the-art classifiers trained on annotated data, to allow for development of filter methods. Furthermore, we present in short our approach and results for distant supervision on these corpora as a strong baseline for future research.

**Keywords:** Distant Supervision, Relation Extraction, Silver Standard

## 1. Introduction

Relation Extraction (RE) in the biomedical domain is a discipline that is under extensive examination in the past decade, with a goal to automatically extract interacting pairs of entities from free text. Currently, a lot of relation extraction systems rely on machine learning, namely classifying pairs of entities to be related or not (Airola et al., 2008; Miwa et al., 2009; Kim et al., 2010). Despite the fact that machine learning has been most successful in identifying relevant relations in text, a drawback is the need for manually annotated training data. Domain experts have to dedicate time and effort to this tedious and labor-intensive process.

As a consequence of the overall scarcity of annotated corpora for relation extraction in the biomedical domain, the approach of distant supervision, *e. g.* automatic labeling of a training set is emerging. Many approaches follow the distant supervision assumption (Mintz et al., 2009; Riedel et al., 2010): "If two entities participate in a relation, all sentences that mention these two entities express that relation." Obviously, this assumption does not hold in general, and therefore exceptions need to be detected.

To allow the community to compare different approaches for distant supervision, we make two corpora, one for protein-protein interaction (PPI) and one for drug-drug interaction (DDI) publicly available.[1] In addition, we present our results on this task as a strong baseline. To complete the purpose of a silver standard, annotations of well-established supervised models on this corpus are included.

### 1.1. Related Work

Distant supervision approaches have received considerable attention in the past few years. However, most of the work is focusing on domains other than biomedical texts. Mintz et al. (2009) use distant supervision to learn to extract relations that are represented in Freebase (Bollacker et al., 2008). Yao et al. (2010) use Freebase as a source of supervision, dealing with entity identification and relation extraction in a joint fashion. Riedel et al. (2010) argue that distant supervision leads to noisy training data that hurts precision and suggest a two step approach to reduce this problem. Vlachos et al. (2009) tackle the problem of biomedical event extraction. The scope of their interest is to identify different event types without using a knowledge base as a source of supervision, but explore the possibility of inferring relations from the text based on the trigger words and dependency parsing, without previously annotated data. Thomas et al. (2011b) make use of a distantly labeled corpus for protein-protein interaction extraction. Different strategies are evaluated to select informative training instances. Buyko et al. (2012) examine the usability of knowledge from a database to generate training sets that capture gene-drug, gene-disease and drug-disease relations.

The CALBC project asks for automated annotation of entity classes in a common corpus to generate a silver standard by combining different predictions (Rebholz-Schuhmann and Ş. Kafkas, 2011). The usability of automatically derived corpora has been recently demonstrated for the task of noun-phrase chunking (Kang et al., 2012). The EVEX data set is the result of applying named entity recognition, parsing and event extraction on full MEDLINE (Landeghem et al., 2011).

[*] These authors contributed equally.

[1] These two corpora are publicly at: http://www.scai.fraunhofer.de/ppi-ddi-silverstandard.html.

| Corpus | Positive pairs | Negative pairs | Total |
|---|---|---|---|
| AIMed | 1000 (0.17) | 4,834 (0.82) | 5,834 |
| BioInfer | 2,534 (0.26) | 7,132 (0.73) | 9,666 |
| HPRD50 | 163 (0.38) | 270 (0.62) | 433 |
| IEPA | 335 (0.41) | 482 (0.59) | 817 |
| LLL | 164 (0.49) | 166 (0.50) | 330 |
| DDI train | 2,400 (0.10) | 21,411 (0.90) | 23,811 |
| DDI test | 755 (0.11) | 6,275 (0.89) | 7,030 |

Table 1: Basic statistics of the five PPI and two DDI corpora. Ratios are given in brackets.

### 1.2. Interaction Databases

The IntAct database (Kerrien et al., 2012) contains protein-protein interaction information. It consists of 290,947 binary interaction evidences, including 39,235 unique pairs of interacting proteins for human species.[2] KUPS (Chen et al., 2010) is a database that combines entries from three manually curated PPI databases (IntAct, MINT (Chatr-aryamontri et al., 2007) and HPRD50 (Prasad et al., 2009)) and contains 185,446 positive pairs from various model organisms, out of which 69,600 belong to human species.[3] Enriching IntAct interaction information with the KUPS database leads to 57,589 unique pairs.[4]

The database DrugBank (Knox et al., 2011) combines detailed drug data with comprehensive drug target information. It consists of 6,707 drug entries. Apart from information about its targets, for certain drugs known interactions with other drugs are given. Altogether, we obtain 11,335 unique DDI pairs.

### 1.3. Manually Curated Corpora

Pyysalo et al. (2008) made five corpora for protein-protein interaction available in the same XML-based file format. Their properties, like size and ratio of positive and negative examples, differ greatly, the latter being the main cause of performance differences when evaluating on these corpora. Moreover, annotation guidelines and contexts differ: AIMed (Bunescu et al., 2005) and HPRD50 (Fundel et al., 2007) are human-focused, LLL (Nedellec, 2005) on Bacillus subtilis, BioInfer (Pyysalo et al., 2007) contains information from various organisms, and IEPA (Ding et al., 2002) is made of sentences that describe 10 selected chemicals, majority of which are proteins, and their interactions.

Segura-Bedmar et al. (2011b) published a drug-drug interaction corpus where the drug mentions have been automatically detected with MetaMap and their pair-wise relations are manually annotated. The corpus is divided into a training and testing set, generated from web-documents describing drug effects.

An overview of the corpora is given in Table 1.

## 2. Methods

In this section, the workflow to prepare the two corpora is presented.

### 2.1. Automatically Labeling a Corpus

One of the most important source of publications in the biomedical domain is MEDLINE[5], currently containing more than 21 million citations.[6] The initial step is annotation of named entities and entity normalization against the databases mentioned in Section 1.2. – in our case performed by ProMiner (Hanisch et al., 2005), a tool proving state-of-the-art results in *e. g.* the BioCreative competition (Fluck et al., 2007). Based on the named entity recognition, only sentences containing co-occurrences of relevant entities are further processed. Based on the distant supervision assumption, each pair of entities is labeled as related if mentioned so in a structured interaction database. Following the closed world assumption, all remaining entity pairs are labeled as non-interacting. To avoid information leakage and biased classification, all documents which are contained in the test corpus are removed from the distantly labeled corpus. Each corpus is sub-sampled to a size of 200,000 entity-pairs, which is more than an order of magnitude larger than any manually annotated PPI or DDI corpus.

### 2.2. Corpus Preprocessing

Sentences are parsed using the Charniak-Lease parser (Lease and Charniak, 2005) with a self-trained re-ranking model specialized for biomedical texts (McClosky, 2010). Resulting constituent parse trees are converted into dependency graphs using the Stanford converter (Marneffe et al., 2006). We create an augmented XML following the recommendations of Airola et al. (2008). This XML encompasses tokens with respective part-of-speech tags, constituent parse tree, and dependency parse tree information. The pairs are augmented with class labels predicted from five different relation extraction methods (see Section 2.3.). For interacting pairs in the PPI corpus we provide the original source (IntAct or KUPS) along with the information if the pair is made of self-interacting proteins. For sentences of the PPI corpus we include the information if an interaction (trigger) word is present. However, in case of DDI trigger-based filtering is not applied (see Bobić et al. (2012)).

### 2.3. Pair Annotation

Labeling two large corpora with database knowledge is the main contribution of this paper. Additionally, we supplement the corpus with predictions of five state-of-the-art relation extraction approaches to provide a supplementing layer of information. (An assessment of the used methodologies for relation extraction was performed by Tikk et al. (2010).)

This includes the shallow linguistic (SL) (Giuliano et al., 2006), all-paths graph (APG) (Airola et al., 2008), subtree (ST) (Vishwanathan and Smola, 2002), subset tree SST (Collins and Duffy, 2001), and spectrum tree (SpT) (Kuboyama et al., 2007) method, which exploit different views on the data. Parameter optimization was performed as

[2] As of January 27th, 2012.

[3] As of August 16th, 2010.

[4] 45,684 out of 69,600 human PPI pairs are available from the KUPS web service due to computational and storage limitations (personal communication).

[5] http://www.ncbi.nlm.nih.gov/pubmed/

[6] As of January, 2012.

described by Tikk et al. (2010). For a detailed description of the feature setting and approach, we refer to the original publications. Entities were blinded by replacing the entity name with a generic string to ensure the generality of the approach. Constituent parse trees have been reduced to the shortest-enclosed parse following the recommendations from Zhang et al. (2006). All five methods are trained on the union of all five PPI corpora and the DDI training and test set respectively. Note that the predictions coming from the five methods are biased towards these training corpora: Models trained on the resulting silver standard (excluding the database annotation) are likely to obtain a too optimistic result, even though the respective sentences from the test set are not used in the training process.

## 3. Results

In this section, we start with an overview of state-of-the-art results for fully supervised relation extraction on PPI and DDI corpora (see Table 1). Section 3.2. gives a statistical outline of the two distantly labeled corpora. Subsequently we present the results of the five relation extraction methods trained on manually annotated data and applied on the distantly labeled corpora. Finally, we present our results for models trained on distantly labeled PPI and DDI data, when evaluated on manually annotated corpora, as a strong baseline for future research.

### 3.1. Performance Overview of Supervised RE Systems

Protein-protein interactions have been extensively investigated in the past decade because of their biological significance. Machine learning approaches have shown the best performance in this domain (*e. g.* BioNLP (Cohen et al., 2011; Tsujii et al., 2011) and DDIExtraction Shared Task (Segura-Bedmar et al., 2011a)).

Our relation extraction system is based on the linear support vector machine classifier LibLINEAR (Fan et al., 2008). The approach employs lexical and dependency parsing features, as explained by Bobić et al. (2012).

Table 4 shows a comparison of state-of-the-art relation extraction systems' performances on 5 PPI corpora, determined by document level 10-fold cross-validation. In Table 2, results of the five best performing systems on the DDI test data set of the DDI extraction workshop are shown. Note that the first three systems use ensemble based methods combining the output of several different classifiers. In addition, the performance of our system, which is later used for distant supervision, is shown in both tables.

### 3.2. Distantly Labeled Corpora for DDI and PPI

The file format of the corpora is by large self explanatory and strongly follows an established file format (Airola et al., 2008; Pyysalo et al., 2008). A short excerpt of the DDI corpus is shown in the appendix. The example consists of one sentence with two annotated drugs that participate in a relation according to DrugBank.

Basic statistics of the two distantly labeled corpora are shown in Table 3. The Charniak-Lease parser does not produce results for nine sentences in the PPI corpus and 14 sentences in the DDI corpus. In general, most methods fail to predict class labels for instances contained in these sentences, leading to a reduced number of predictions per corpus. However, the effect is only marginal as <1 % of all entity pairs are affected by this problem.

| Methods | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| Thomas et al. (2011a) | 60.5 | **71.9** | **65.7** |
| Chowdhury et al. (2011) | 58.6 | 70.5 | 64.0 |
| Chowdhury and Lavelli (2011) | 58.4 | 70.1 | 63.7 |
| Björne et al. (2011) | 58.0 | 68.9 | 63.0 |
| Minard et al. (2011) | 55.2 | 64.9 | 59.6 |
| Our system (*lex*) | 62.7 | 52.1 | 56.9 |
| Our system (*lex+dep*) | **66.9** | 57.9 | 62.1 |

Table 2: Comparison of fully supervised relations extraction systems for DDI. (*lex* denotes the use of lexical features, *lex+dep* the additional use of dependency parsing-based features.) The first three systems are based on ensemble learning.

|  | PPI | DDI |
|---|---|---|
| Abstracts | 49,958 | 76,859 |
| Sentences | 51,934 | 79,701 |
| Pos. Sent. | 19,891 | 5,587 |
| Tokens | 1,608,899 | 2,520,545 |
| Entities | 150,886 | 203,315 |
| Pairs | 200,000 | 200,000 |
| Pos. Pairs | 37,600 | 8,705 |

Table 3: Statistics of the distant PPI and DDI corpora. (pos. sent. denotes the number of sentences with at least one related entity pair.)

### 3.3. Pair Annotation

As shown in Table 5, relation extraction methods tend to classify between 10.9 % and 16.8 % of all protein pairs as interacting. However, the overall ratio of positive instances across all five PPI corpora is greater, measuring up to 32.6 %. We observe similar values for the distant DDI corpus with ratios ranging from 12.7 % to 19.6 %.

The distribution of confidence scores (distance to the hyperplane) for all methods on both corpora is shown in Figure 1. Instances with a negative sign are classified as non-interacting and instances with a positive sign are classified as interacting. The linear association between different methods is assessed using Pearson correlation for all instances contained in the distantly supervised corpus. We observe correlation coefficients ranging from 0.29 (APG versus SpT) to 0.59 (APG versus SL) for PPI and between 0.34 (APG vs ST) to 0.71 (ST vs SST) for DDI. Significance of all pairwise correlations is assessed using a t-test and is in all cases highly significant (p-value < 0.01). Correlation is exemplarily depicted as scatterplot for SL and APG on PPI in Figure 2. Both methods agree on the predicted class label on instances contained in the first and third quadrant, whereas the two methods have conflicting results for instances in the second and fourth quadrant. The figure indicates that some instances can be confidently classified by one method

|  | AIMed |  |  | BioInfer |  |  | HPRD50 |  |  | IEPA |  |  | LLL |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| Airola et al. (2008) | 52.9 | 61.8 | 56.4 | 56.7 | 67.2 | 61.3 | 64.3 | 65.8 | 63.4 | 69.6 | **82.7** | **75.1** | 72.5 | 87.2 | 76.8 |
| Kim et al. (2010) | 61.4 | 53.2 | 56.6 | 61.8 | 54.2 | 57.6 | 66.7 | 69.2 | 67.8 | **73.7** | 71.8 | 72.9 | 76.9 | 91.1 | **82.4** |
| Fayruzov et al. (2009) |  |  | 39.0 |  |  | 34.0 |  |  | 56.0 |  |  | 72.0 |  |  | 76.0 |
| Liu et al. (2010) |  |  | 54.7 |  |  | 59.8 |  |  | 64.9 |  |  | 62.1 |  |  | 78.1 |
| Miwa et al. (2009) | 55.0 | **68.8** | **60.8** | 65.7 | **71.1** | **68.1** | 68.5 | **76.1** | 70.9 | 67.5 | 78.6 | 71.7 | **77.6** | 86.0 | 80.1 |
| Tikk et al. (2010) | 47.5 | 65.5 | 54.5 | 55.1 | 66.5 | 60.0 | 64.4 | 67 | 64.2 | 71.2 | 69.3 | 69.3 | 74.5 | 85.3 | 74.5 |
| Our system (*lex*) | 62.9 | 50.0 | 55.7 | 59.3 | 55.1 | 57.1 | **72.4** | 75.6 | **73.9** | 67.7 | 73.3 | 70.4 | 66.6 | 88.6 | 76.1 |
| Our system (*lex+dep*) | **63.6** | 52.0 | 57.2 | **65.8** | 62.9 | 64.3 | 70.8 | 74.0 | 72.4 | 70.4 | 76.1 | 73.2 | 70.4 | **91.6** | 79.6 |

Table 4: Comparison of fully supervised relation extraction systems for PPI.

|  | PPI |  | DDI |  |
|---|---|---|---|---|
| Method | positive | negative | positive | negative |
| SL | 33,677 (16.8) | 166,219 | 25,344 (12.7) | 174,539 |
| SpT | 21,971 (10.9) | 177,921 | 29,324 (14.6) | 170,558 |
| ST | 28,885 (14.4) | 171,112 | 39,286 (19.6) | 160,597 |
| SST | 24,840 (12.4) | 175,157 | 25,841 (12.9) | 174,039 |
| APG | 26,313 (13.1) | 173,686 | 25,357 (12.7) | 174,643 |

Table 5: Distribution of positive and negative instances for the different methods on both distantly labeled corpora. The ratio of positive examples is given in brackets.

(high distance to the hyperplane), but the other method is comparably inconfident. This suggests a great variability between the methods.

Even though the correlation between the methods is lower than expected, the inter-classification agreement (accuracy) is comparably high and ranges between 80.7 % to 86.4 % and 78.2 % to 84.6 % for all PPI and DDI instances respectively. We observe a large agreement between the distantly labeled corpus and the classification methods with approximately 76 % overall agreement for PPI and 80 % for DDI. The association between distantly labeled corpora and all classification methods is significant according to a fisher test (p-value $< 0.01$), except for SpT where we observe a p-value of 0.04. However, the large overall agreement is due to the high number of negative instances in the distant corpora and predicted by the different methods. For positive PPI instances alone we observe an agreement of approximately 27 % between instances labeled as interacting by our knowledge base and the classification methods. Similar effects can be observed for the DDI corpus. We assessed the overall agreement between methods and the two distantly labeled corpora using Cohens $\kappa$. For PPI we observe values ranging between 0.07 to 0.19 and for DDI we observe $\kappa$ values of 0.03. The low $\kappa$ values show a comparably small agreement between classification methods and distantly labeled corpora and more sophisticated filtering techniques might be required to make optimal use of the corpus. Results in terms of precision, recall and $F_1$ can be seen in Table 6.

### 3.4. Baselines for Distantly Supervised Models

For each experiment we sample random subsets of 10,000 entity pairs from the proposed corpora. All experiments are performed five times to reduce the influence of sampling different subsets. We apply the method proposed by Bobić et al. (2012), with dependency parsing based features and filtering auto-interacting entities. For PPI, trigger-based filtering is applied (compare to Section 2.2.). Table 7 shows the average performance trained on the distantly labeled PPI and DDI corpora.

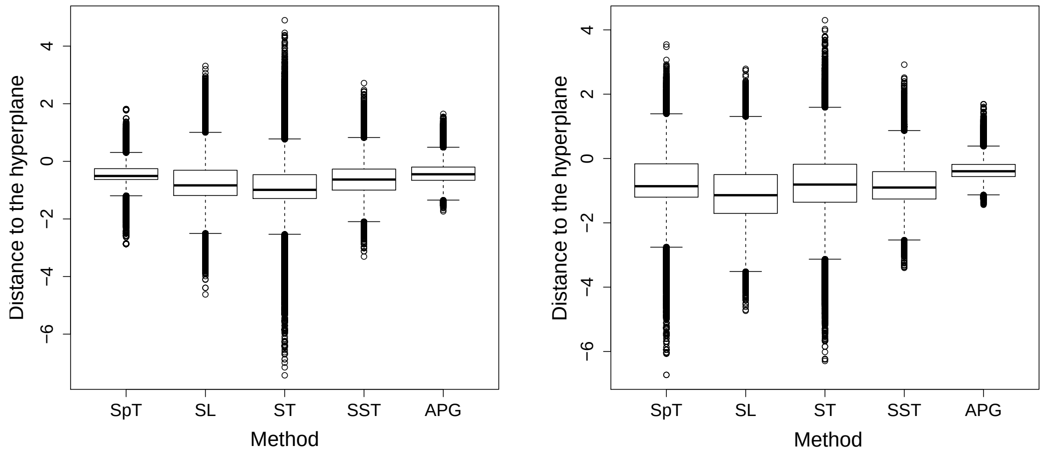
|  | PPI |  |  | DDI |  |  |
|---|---|---|---|---|---|---|
| Method | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| SL | 35.1 | 31.4 | 33.2 | 6.4 | 18.7 | 9.5 |
| SpT | 27.4 | 16.0 | 20.2 | 4.5 | 15.3 | 7.0 |
| ST | 35.2 | 27.1 | 30.6 | 5.5 | 25.1 | 9.1 |
| SST | 32.3 | 21.4 | 25.7 | 6.2 | 18.6 | 9.3 |
| APG | 36.0 | 25.1 | 29.6 | 5.8 | 16.7 | 8.6 |

Table 6: Comparison of all methods on both distantly labeled corpora. ($P$ denotes precision, $R$ recall and $F_1$ the harmonic mean of $P$ and $R$ )

Note that the instance labels used for training the model are based solely on database knowledge. The information provided by five supervised methods (addressed in Section 2.3.) are not taken into account for generating baseline results, although they are available to be used in future work.

Our system outperforms co-occurrence results for all five PPI corpora, as shown in Table 7. $F_1$ measure of AIMed and BioInfer, for which we assume to have the most realistic pos/neg ratio, outperforms the baseline by around 9 percentage points (pp). HPRD50, IEPA and LLL have an improvement of 4.7 pp, 5.3 pp and 0.8 pp respectively, due to high fractions of positive instances (leading to a strong co-occurrence baseline).

Evaluation on corpora that have different properties than the training set leads to decreased performance (Airola et al., 2008; Tikk et al., 2010). Often, the properties of a test corpus (like MEDLINE) are not known for real world



(a) PPI

(b) DDI

Figure 1: Boxplot on distance to the hyperplane of all used methods for both corpora.

| | Our system (*lex*) | | | Our system (*lex*+*dep*) | | | Co-occ. | | | Tikk et al. (2010) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Corpus | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| AIMed | 25.6 | 78.4 | 38.6 | 25.0 | 81.9 | 38.4 | 17.1 | 100 | 29.3 | 28.3 | 86.6 | **42.6** |
| BioInfer | 40.4 | 66.7 | **50.3** | 40.3 | 66.9 | **50.3** | 26.2 | 100 | 41.5 | 62.8 | 36.5 | 46.2 |
| HPRD50 | 45.7 | 85.1 | 59.4 | 44.9 | 86.3 | 59.0 | 37.6 | 100 | 54.7 | 56.9 | 68.7 | **62.2** |
| IEPA | 50.0 | 87.2 | **63.5** | 49.9 | 85.8 | 63.1 | 41.0 | 100 | 58.2 | 71.0 | 52.5 | 60.4 |
| LLL | 56.4 | 83.1 | **67.2** | 56.3 | 83.2 | **67.2** | 49.7 | 100 | 66.4 | 79.0 | 57.3 | 66.4 |
| DDI | 33.2 | 39.2 | 36.0 | 33.0 | 44.1 | **37.7** | 10.7 | 100 | 19.4 | — | — | — |

Table 7: Results (in %) achieved when training on 10,000 distantly labeled instances and testing on 5 PPI corpora and the DDI test corpus, respectively.

applications. Thus cross-learning[7] is considered to provide a more realistic scenario to compare the performance of distantly supervised systems to fully supervised systems. Our approach outperforms the state-of-the-art cross-learning results from Tikk et al. (2010) in three out of five corpora, most notably in case of BioInfer where an increase of more than 4 pp in $F_1$ measure is observable.

In the case of drug-drug interaction, it is noteworthy that the manually annotated corpora are generated from web documents discussing drug effects which are not necessarily contained in MEDLINE. Hence, this evaluation corpus can be considered as out-domain and provides additional insights on the robustness of distant-supervision. Table 7 shows that compared to co-occurrence, a gain of more than 18 pp is achieved when training on a distantly labeled DDI corpus. Taking into account the high class imbalance of the DDI test set (see Table 1), which is most similar to the AIMed corpus, a $F_1$ measure of 37.7 % is encouraging.

Application of distant supervision to five substantially different PPI corpora and further utilization of the same workflow to DDI confirms its robustness and usability.

## 4. Discussion

This paper introduces two distantly labeled corpora created for the purpose of protein-protein and drug-drug interaction extraction. Corpus generation and the process of automatic pair labeling using database information are presented, together with strong baseline results for distantly supervised relation extraction.

In addition to entity-pair annotation based on a knowledge base, we add predictions from five relation extraction systems, trained on manually annotated corpora. These annotations can be exploited to develop better instance filtering techniques. Several assessments demonstrated the superiority of ensemble methods, hence it might be beneficial to combine classifier predictions for the sake of higher method robustness.

Our distant supervision baseline achieves competitive results and outperforms co-occurrence in all test cases. Comparison to fully supervised cross-learning results for PPI argues for the opportunities of using automatically annotated data.

This paper presents the potential of distant learning to allow a fully automated relation extraction process. The PPI and DDI corpora are made freely available to the community such that novel strategies of efficient employment of database knowledge can be compared.

[7] For five PPI corpora: train on four, test on the remaining.

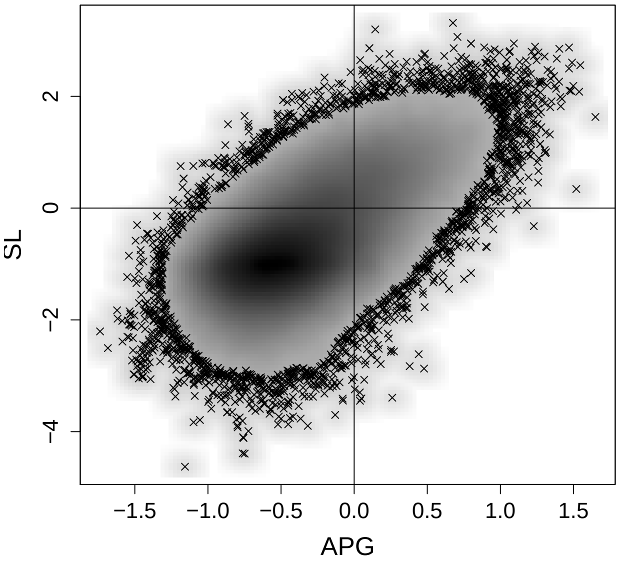


Figure 2: Scatterplot for distance to the hyperplane between APG and SL on the distantly labeled PPI corpus. Warm regions (dark) indicate an accumulation of instances whereas light regions contain no instances. The 2,000 points in areas with lowest regional density are plotted separately.

## 5. Acknowledgements

Thanks to our colleagues Juliane Fluck and Theo Mevissen for fruitful discussions and technical help. T. Bobić was partially funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School. P. Thomas was funded by the German Federal Ministry of Education and Research (grant No 0315417B). R. Klinger was partially funded by the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement no. 248726.

## 6. References

A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. All-paths Graph Kernel for Protein-protein Interaction Extraction with Evaluation of Cross-corpus Learning. *BMC Bioinformatics*, 9(Suppl 11):S2.

J. Björne, A. Airola, T. Pahikkala, and T. Salakoski. 2011. Drug-drug interaction extraction with RLS and SVM classiffers. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 35–42.

T. Bobić, R. Klinger, P. Thomas, and M. Hofmann-Apitius. 2012. Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactionsp. In O. Abend, C. Biemann, A. Korhonen, A. Rappoport, R. Reichart, and A. Sgaard, editors, *ROBUS-UNSUP*.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250.

R. C. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 33(2):139–155, Feb.

E. Buyko, E. Beisswanger, and U. Hahn. 2012. The extraction of pharmacogenetic and pharmacogenomic relations–a case study using pharmgkb. *PSB*, pages 376–387.

A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M.V. Schneider, L. Castagnoli, and G. Cesareni. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–D574.

X. Chen, J. C. Jeong, and P. Dermyer. 2010. KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. *Nucleic Acids Res*, 39(Database issue):D750–D754.

F. M. Chowdhury and A. Lavelli. 2011. Drug-drug interaction extraction using composite kernels. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 27–33.

F. M. Chowdhury, A. B. Abacha, A. Lavelli, and P. Zweigenbaum. 2011. Two different machine learning techniques for drug-drug interaction extraction. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 19–26.

K. B. Cohen, D. Demner-Fushman, S. Ananiadou, J. Pestian, J. Tsujii, and B. Webber, editors. 2011. *Proceedings of the BioNLP*.

M. Collins and N. Duffy. 2001. Convolution Kernels for Natural Language. In *Proc. of Neural Information Processing Systems (NIPS'01)*, pages 625–632, Vancouver, BC, Canada.

J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. 2002. Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, pages 326–337.

E. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Machine Learning Research*, 9:1871–1874.

T. Fayruzov, M. De Cock, C. Cornelis, and V. Hoste. 2009. Linguistic feature analysis for protein interaction extraction. *BMC Bioinformatics*, 10(1):374.

J. Fluck, H. T. Mevissen, H. Dach, M. Oster, and M. Hofmann-Apitius. 2007. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In *BioCreative 2*, pages 149–151.

K. Fundel, R. Kuffner, and R. Zimmer. 2007. RelEx-Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 401–408, Trento, Italy.

D. Hanisch, K. Fundel, H. T. Mevissen, R. Zimmer, and J. Fluck. 2005. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S14.

N. Kang, E. M. van Mulligen, and J. A. Kors. 2012. Training text chunkers on a silver standard corpus: can silver replace gold? *BMC Bioinformatics*, 13(1):17, Jan.

S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R.C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger,

P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40:D841–D846.

S. Kim, J. Yoon, J. Yang, and S. Park. 2010. Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11:107.

C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Chi Guo, and D.S Wishart. 2011. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*, 39(Database issue):D1035–D1041.

T. Kuboyama, K. Hirata, H. Kashima, K. F. Aoki-Kinoshita, and H. Yasuda. 2007. A Spectrum Tree Kernel. *Information and Media Technologies*, 2(1):292–299.

S. Van Landeghem, F. Ginter, Y. Van de Peer, and T. Salakoski. 2011. EVEX: A PubMed-Scale Resource for Homology-Based Generalization of Text Mining Predictions. In *Proc. of BioNLP*, pages 28–37.

M. Lease and E. Charniak. 2005. Parsing biomedical literature. In *Proc. of IJCNLP'05*, pages 58–69.

B. Liu, L. Qian, H. Wang, and G. Zhou. 2010. Dependency-driven feature-based learning for extracting protein-protein interactions from biomedical text. In *COLING*, pages 757–765.

M. C. De Marneffe, B. Maccartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*, pages 449–454.

D. McClosky. 2010. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Brown University.

A. L. Minard, L. Makour, A. L. Ligozat, and B. Grau. 2011. Feature Selection for Drug-Drug Interaction Detection Using Machine-Learning Based Approaches. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 43–50.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP*, pages 1003–1011.

M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii. 2009. A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. In *Proc. of EMNLP*, pages 121–130.

C. Nedellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proc. of the ICML05 workshop: Learning Language in Logic (LLL'05)*, volume 18, pages 97–99.

T. S. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A.Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. 2009. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–D772.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. BioInfer: A Corpus for Information Extraction in the Biomedical Domain. *BMC Bioinformatics*, 8(50).

S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein–protein interaction corpora. *BMC Bioinformatics*, 9 Suppl 3:S6.

D. Rebholz-Schuhmann and Ş. Kafkas, editors. 2011. *Proceedings of the Second CALBC Workshop*.

S. Riedel, L. Yao, and A. McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *ECML PKDD*.

I. Segura-Bedmar, P. Martínez, and D. Sanchez-Cisneros, editors. 2011a. *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*.

I. Segura-Bedmar, P. Martínez, and D. Sanchez-Cisneros. 2011b. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In *Challenge Task on Drug-Drug Interaction Extraction 2011*, pages 1–9.

P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser. 2011a. Relation Extraction for Drug-Drug Interactions using Ensemble Learning. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 11–18.

P. Thomas, I. Solt, R. Klinger, and U. Leser. 2011b. Learning Protein Protein Interaction Extraction using Distant Supervision. In *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pages 34–41.

D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6:e1000837.

J. Tsujii, J.-D. Kim, and S. Pyysalo, editors. 2011. *Proceedings of the BioNLP Shared Task*.

S. V. N. Vishwanathan and A. J. Smola. 2002. Fast Kernels for String and Tree Matching. In *Proc. of Neural Information Processing Systems (NIPS'02)*, pages 569–576, Vancouver, BC, Canada.

A. Vlachos, P. Buttery, D. Ó Séaghdha, and T. Briscoe. 2009. Biomedical Event Extraction without Training Data. In *BioNLP*, pages 37–40.

L. Yao, S. Riedel, and A. McCallum. 2010. Collective Cross-Document Relation Extraction Without Labeled Data. In *EMNLP*, pages 1013–1023.

M. Zhang, J. Zhang, J. Su, and G. Zhou. 2006. A Composite Kernel to Extract Relations between Entities with Both Flat and Structured Features. In *Proc. of the 21st International Conference on Computational Linguistics*, pages 825–832, Sydney, Australia, July.

## 7. Appendix

An excerpt of the corpus in XML format:

```
<corpus source=”SilverDDICorpus”>
  <document id=”d3” origId=”10796253”>
    <sentence id=”d3.s0” origId=”10796253.s14”
        text=”In the subset with initial BUN/creatinine ratio > 20 mg/mg, 2 of 18 patients receiving furosemide
         could not complete a 3−dose course of indomethacin because of toxicity.”>
        <entity charOffset=”87−96” id=”d3.s0.e0” origId=”10796253.s14.e0” text=”furosemide”
        type=”drug”/>
        <entity charOffset=”136−147” id=”d3.s0.e1” origId=”10796253.s14.e1” text=”indomethacin”
        type=”drug”/>
        <pair e1=”d3.s0.e0” e2=”d3.s0.e1” id=”d3.s0.p0” interaction=”True” source=”DrugBank”
        APG=”0.32” SL=”0.60” ST=”−1.08” SST=”0.12” SpT=”0.34”/>
        <sentenceanalyses>
          <tokenizations>
            <tokenization tokenizer=”Charniak−Lease”>
            <token POS=”IN” charOffset=”0−1” id=”t_1” text=”In”/>
            <token POS=”DT” charOffset=”3−5” id=”t_2” text=”the”/>
            <token POS=”NN” charOffset=”7−12” id=”t_3” text=”subset”/>
            ...
            </tokenization>
          </tokenizations>
        <bracketings>
          <bracketing tokenizer=”Charniak−Lease” parser=”Charniak−Lease” bracketing=”(S1 (S (S (PP
          (IN In) (NP (NP (DT the) (NN subset)) (PP (IN with) (NP (NP (JJ initial) (NN BUN/creatinine) (NN ratio) (NN &gt;))
          (NP (CD 20) (NN mg/mg)))))) (, ,) (NP (NP (CD 2)) (PP (IN of) (NP (NP (CD 18) (NNS patients)) (VP (VBG receiving)
          (NP (NN furosemide)))))) (VP (MD could) (RB not) (VP (VB complete) (NP (NP (DT a) (JJ 3−dose) (NN course))
          (PP (IN of) (NP (NN indomethacin)))) (PP (IN because) (IN of) (NP (NN toxicity)))))) (. .)))”>
           <charOffsetMapEntry sentenceTextCharOffset=”0−1” bracketingCharOffset=”18−19”/>
           <charOffsetMapEntry sentenceTextCharOffset=”3−5” bracketingCharOffset=”34−36”/>
           <charOffsetMapEntry sentenceTextCharOffset=”7−12” bracketingCharOffset=”43−48”/>
           ...
          </bracketing>
        </bracketings>
        <parses>
          <parse tokenizer=”Charniak−Lease” parser=”Charniak−Lease”>
            <dependency id=”d_1” t1=”t_3” t2=”t_2” type=”det” origId=”det(subset−3, the−2)”/>
            <dependency id=”d_2” t1=”t_20” t2=”t_3” type=”prep_in” origId=”prep_in(complete−20, subset−3)”/>
            <dependency id=”d_3” t1=”t_8” t2=”t_5” type=”amod” origId=”amod(&gt;−8, initial−5)”/>
            ...
          </parse>
        </parses>
       </sentenceanalyses>
    </sentence>
        ...
  </document>
  ...
</corpus>
```