Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions

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

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

[2]Bonn-Aachen Center for Information Technology
Dahlmannstraße 2
53113 Bonn
Germany

[3]Computer Science Institut
Humboldt-Universität
Unter den Linden 6
10099 Berlin
Germany

{tbobic,klinger,hofmann-apitius}@scai.fraunhofer.de
thomas@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, like automatically annotated corpora.

Recent work in the biomedical domain has applied distant supervision for protein-protein interaction (PPI) with reasonable results making use of the IntAct database. Such data is typically noisy and heuristics to filter the data are commonly applied. We propose a constraint to increase the quality of data used for training based on the assumption that no self-interaction of real-world objects are described in sentences. In addition, we make use of the University of Kansas Proteomics Service (KUPS) database. These two steps show an increase of 7 percentage points (pp) for the PPI corpus AIMed. We demonstrate the broad applicability of our approach by using the same workflow for the analysis of drug-drug interactions, utilizing relationships available from the drug database DrugBank. We achieve 37.31 % in F_1 measure without manually annotated training data on an independent test set.

1 Introduction

Assuming co-mentioned entities to be related is an approach of extracting relations of real-world objects with limited precision. Extracting high quality interaction pairs from free text allows for building networks, *e. g.* of proteins, which need less manual curation to serve as a model for further knowledge processing steps. Nevertheless, just assuming co-occurrence to model an interaction or relation is common, as the development of interaction extraction systems can be time-consuming and complex.

Currently, a lot of relation extraction (RE) 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.

Specific biomedical domains have been explored more extensively than others, thus creating an imbalance in the number of existing corpora for a specific RE task. Protein-protein interactions (PPI) have been investigated the most, which gave rise to a number of available corpora. Pyysalo et al. (2008) standardized five PPI corpora to a unified XML format. Recently, a drug-drug-interaction (DDI) corpus is made available in the same format, originally for the DDI Extraction Workshop[1] (Segura-Bedmar et al., 2011b).

As a consequence of the overall scarcity of annotated corpora for RE in the biomedical domain, the approach of distant supervision, *e. g.* to automatically label a training set is emerging. Many approaches make use of the distant supervision assumption (Mintz et al., 2009; Riedel et al., 2010):

*These authors contributed equally.

[1]Associated with the conference of the spanish society for natural language processing (SEPLN) in 2011, http://labda.inf.uc3m.es/DDIExtraction2011/

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 35–43,
Avignon, France, April 23 - 27 2012.

> 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 which are not used for training a model. Thomas et al. (2011b) successfully used simple filtering techniques in a distantly supervised setting to extract PPI. In contrast to their work, we introduce a more generic filter to detect frequent exceptions from the distant supervision assumption and make use of more data sources, by merging the interaction information from IntAct and KUPS databases (discussed in Section 2.1). In addition, we present the first system (to our knowledge), evaluating distant supervision for drug-drug interaction with promising results.

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. Entity types are restricted to those compatible with selected relations. 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. They identify the sentences which express the known relations ("expressed-at-least-once" assumption) and thus frame the problem of distant supervision as an instance of constraint-driven semi-supervision, achieving 31 % of error reduction.

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) develop a distantly labeled corpus for protein-protein interaction extraction. Different strategies are evaluated to select valuable training instances. Competitive results are obtained, compared to purely supervised methods.

Very recent work examines the usability of knowledge from PharmGKB (Gong et al., 2008) to generate training sets that capture gene-drug, gene-disease and drug-disease relations (Buyko et al., 2012). They evaluate the RE for the three interaction classes in intrisic and extrinsic experimental settings, reaching F_1 measure of around 80 % and up to 77.5 % respectively.

2 Resources

2.1 Interaction Databases

The IntAct database (Kerrien et al., 2012) contains protein-protein interaction information. It is freely available, manually curated and frequently updated. It consists of 290,947 binary interaction evidences, including 39,235 unique pairs of interacting proteins for human species.[2]

In general, PPI databases are underannotated and the overlap between them is marginal (De Las Rivas and Fontanillo, 2010). Combining several databases allows to cover a larger fraction of known interactions resulting in a more complete knowledge base. KUPS (Chen et al., 2010) is a database that combines entries from three manually curated PPI databases (IntAct, MINT (Chatraryamontri 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.

2.2 Corpora

For evaluation of protein-protein interaction, the five corpora made available by Pyysalo et al. (2008) are used. Their properties, like size and ratio of positive and negative examples, differ greatly,

[2] As of January 27th, 2012.

[3] As of August 16th, 2010.

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

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.

Methods	P	R	F_1
Thomas et al. (2011a)	60.54	**71.92**	**65.74**
Chowdhury et al. (2011)	58.59	70.46	63.98
Chowdhury and Lavelli (2011)	58.39	70.07	63.70
Björne et al. (2011)	58.04	68.87	62.99
Minard et al. (2011)	55.18	64.90	59.65
Our system (*lex*)	63.30	52.32	57.28
Our system (*lex+dep*)	**66.46**	56.69	61.19

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 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, the majority of which are proteins, and their interactions.

For the purposes of DDI extraction, the corpus published by Segura-Bedmar et al. (2011b) is used. This corpus is generated from web-documents describing drug effects. It is divided into a training and testing set. An overview of the corpora is given in Table 1.

3 Methods

In this section, the relation extraction system used for classification of interacting pairs is presented. Furthermore, the process of generating an automatically labeled corpus is explained in more detail, along with specific characteristics of the PPI and DDI task.

3.1 Interaction Classification

We formulate the task of relation extraction as feature-based classification of co-occurring entities in a sentence. Those are assigned to be either related or not, without identifying the type of relation. Our RE system is based on rich feature vectors and the linear support vector machine classifier LibLINEAR, which has shown high performance (in runtime as well as model accuracy) on large and sparse data sets (Fan et al., 2008).

The approach is based on lexical features, optionally with dependency parsing features created using the Stanford parser (Marneffe et al., 2006). Lexical features are bag-of-words (BOW) and n-grams based, with $n \in \{1, 2, 3, 4\}$. They encompass the local (window size 3) and global (window size 13) context left and right of the entity pair, along with the area between the entities (Li et al., 2010). Additionally, dictionary based domain specific trigger words are taken into account.

The respective dependency parse tree is included through following the shortest dependency path hypothesis (Bunescu and Mooney, 2005), by using the syntactical and dependency information of edges (*e*) and vertices (*v*). So-called *v-walks* and *e-walks* of length 3 are created as well as n grams along the shortest path (Miwa et al., 2010).

3.2 Automatically Labeling a Corpus in General

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 – 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 are further processed. Based on the distant supervision assumption, each pair of entities is labeled as related if mentioned so in a structured interaction databases. Note that this requires the step of entity normalization.

3.3 Filtering Noise

A sentence may contain two entities of an interacting pair (as known from a database), but does not describe their interaction. Likewise, a sentence

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

[6] As of January, 2012.

may talk about a novel interaction which has not been stored in the database. Therefore, filtering strategies need to be employed to help in deciding which pairs are annotated as being related and which not.

Thomas et al. (2011b) propose the use of trigger words, *i. e.*, an entity pair of a certain sentence is marked as *positive* (related) if the database has information about their interaction and the sentence contains at least one trigger word. Similarly, a *negative* (non-related) example is a pair of entities that does not interact according to the database and their sentence does not contain any trigger word. Pairs which do not fulfil both constraints are discarded.

Towards improvement of the heuristics for reducing noise, we introduce the constraint of "auto-interaction filtering" (AIF): If entities from an entity pair both refer to the same real-world object, the pair is labeled as not interacting. Even though self-interactions are known for proteins and drugs, such pairs can rarely be observed to describe an interaction but rather are repeated occurences or abbreviations. Moreover, the fundamental advantage of AIF is that it requires no additional manual effort.

3.4 Application on Protein-Protein Interaction and Drug-Drug Interaction

In biomedical texts there are often mentions of multiple proteins in the same sentence. However, this co-occurrence does not necessarily signal that the sentence is talking about their relation. Hence, to reduce noise, a list of trigger words specific to the problem is required. The rationale behind this filter is that the interaction between two entities is usually expressed by a specific (trigger) word. For protein-protein-interactions, we use the trigger list compiled by Thomas et al. (2011b)[7]. In addition to using IntAct alone, we introduce the use of KUPS database (as described in Section 2.2).

For drug-drug-interaction, to our knowledge, no DDI-specific trigger word list developed by domain experts is available. Therefore, filtering via such term occurrences is not applied in this case.

[7] `http://www2.informatik.hu-berlin.de/~thomas/pub/2011/iwords.txt`

4 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). Furthermore, experimental settings for distant supervision are explained. Finally, we present specific results for models trained on distantly labeled data, when evaluated on manually annotated PPI and DDI corpora.

4.1 Performance overview of supervised RE systems

Protein-protein interactions has 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) and DDIExtraction Shared Task (Segura-Bedmar et al., 2011a)). Table 3 gives a comparison of RE systems' performances on 5 PPI corpora, determined by document level 10-fold cross-validation.[8] The use of dependency parsing-based features increases the F_1 measure by almost 4 pp.

Table 2 shows results of the five best performing systems on the held out test data set of the DDI extraction workshop (Segura-Bedmar et al., 2011b). In addition, the result of our system is shown. Note that the first three systems use ensemble based methods combining the output of several different systems.

The results presented in Table 2 and 3 give a performance overview of the RE system used in distant learning strategies.

4.2 Experimental Setting

To avoid information leakage and biased classification, all documents which are contained in the test corpus are removed. For each experiment we sample random subsets to reduce processing time. This allows us to evaluate the impact of different combinations of subset size and the ratio of related and non-related (pos/neg) entity pairs, having in mind the problem of imbalanced datasets (Chawla et al., 2004). All experiments are performed five times to reduce the influence of sampling different subsets. This leads to more reliable precision, recall, and F_1 values.

[8] Separating into training and validation sets is performed on document level, not on instance (entity pair) level. The latter could lead to an unrealistiallly optimistic estimate (Van Landeghem et al., 2008)

	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 s. (*lex*)	62.3	46.3	53.1	59.1	54.3	56.6	**69.7**	69.4	69.6	67.5	73.2	70.2	66.9	84.6	74.7
Our s. (*lex+dep*)	**65.1**	48.6	55.7	64.7	57.6	61.0	69.3	69.8	69.5	67.0	72.5	69.7	71.2	86.3	78.0

Table 3: Comparison of fully supervised relations extraction systems for PPI.

Strategy	Pairs	Positive pairs	Sentences
1	3,304,033	511,665 (0.155)	842,339
2	5,560,975	1,389,036 (0.250)	1,172,920
3	2,764,626	359,437 (0.130)	780,658
4	3,454,805	650,455 (0.188)	896,344

Table 4: Statistics of the fours strategies used in distant supervision for PPI task: 1) IntAct, 2) IntAct + KUPS, 3) IntAct + AIF, 4) IntAct + KUPS + AIF. Ratios are given in brackets.

4.3 Protein-protein interaction

We explore four strategies to determine the impact of using additional database knowledge (IntAct and KUPS) and to test the utility of our novel condition (AIF).

Table 4 shows the difference in retrieved number of sentences and protein pairs, including the percentage of positive examples in the whole data set. As expected, by using more background knowledge, the number of sentences and instances retrieved from MEDLINE rises. An increase of both negative and positive pairs is observed, since a relevant sentence can have negative pairs along with the positive ones. After applying additional interaction knowledge, the fraction of positive examples (see 3rd column in Table 4) increases from 15.5 % (IntAct) to 25 % (IntAct+KUPS). However, employment of the AIF condition to both IntAct and IntAct+KUPS strategies leads to a reduction of these values (*e. g.* fraction of positive examples reduces from 15.5 % to 13 % and from 25 % to 18.8 %).

For simplicity reasons all runs are performed using only lexical features.

Table 5 shows the average values of distant supervision experiments carried out for the PPI task. A significant correlation between pos/neg ratio and precision/recall holds. This clearly indicates the tendency of classifiers to assign more test instances to the class more often observed during training. In accordance with their class distribution, AIMed reaches highest performance in case of lower fraction of positive instances (*i. e.* 30 % or 40 %), while for IEPA and LLL the optimal ratio is in favor of the positive class (*i. e.* 70 % or 80 %).

Comparative results of the distant learning strategies IntAct and IntAct+KUPS tested on five PPI corpora indicate that additional knowledge bases do not help per se. Supplementary employment of the KUPS database leads to a drop in performances seen in four out of five test cases (a decrease of 1.7 pp in F_1 measure is most notably observed in case of HPRD50). However, introduction of the novel filtering condition, in both strategies IntAct+AIF and IntAct+KUPS+AIF, shows a favorable effect on the precision and leads to an increase of up to 6 pp in F_1 measure, compared to IntAct and IntAct+KUPS.

Applying AIF to the baseline IntAct increases F_1 measure of AIMed and HPRD50 from 34.4 % to 37.8 % and from 56.1 % to 59.1 %, respectively. An even larger impact is observed when comparing IntAct+KUPS and IntAct+KUPS+AIF. For AIMed, HPRD50 and IEPA an increase of around 6 pp is achieved, while F_1 measure of BioInfer and LLL is improved around 3 pp. Table 5 clearly shows that IntAct+KUPS+AIF is outperforming other strategies in all five test cases by achieving F_1 measures of 39.0 % for AIMed, 52.0 % for BioInfer, 60.2 % for HPRD50, 63.4 % for IEPA and 69.3 % for LLL.

Analysis of the database (IntAct+KUPS) pairs reveals that in total there are 5,550 (around 10 %) proteins that interact with themselves, with 4,918 (89 %) originating from the KUPS database. This indicates a number of instances that represent auto-interacting proteins which contribute to increase of false positives. Such proportion where a majority of them come from KUPS explains the decrease

Strategy	pos/neg	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
IntAct	30-70	22.3	75.8	**34.4**	41.7	54.1	46.9	42.6	73.8	53.9	44.6	70.3	54.5	58.9	63.5	61.0
	40-60	21.5	83.5	34.2	40.0	61.9	48.5	42.0	81.7	55.5	44.4	78.0	56.6	55.7	73.3	63.2
	50-50	20.8	87.0	33.5	38.7	67.1	49.0	41.4	86.9	**56.1**	43.7	82.2	57.1	54.6	80.7	65.1
	60-40	20.0	90.8	32.8	37.3	72.6	**49.2**	40.5	91.2	**56.1**	43.2	85.6	57.4	52.4	86.7	65.3
	70-30	19.0	94.5	32.1	35.4	79.5	48.9	39.6	93.4	55.6	42.6	89.3	57.7	50.7	92.1	**65.4**
	80-20	18.6	96.8	31.2	33.5	86.5	48.3	38.6	96.2	55.1	42.1	93.3	**58.1**	49.4	96.7	65.0
IntAct + KUPS	30-70	20.6	48.9	29.0	37.5	30.0	33.3	38.6	45.8	41.8	33.1	25.3	28.6	55.3	25.4	34.6
	40-60	21.6	70.3	33.0	39.3	47.4	42.9	40.7	70.2	51.5	41.0	49.6	44.9	58.6	49.3	53.2
	50-50	20.8	81.6	**33.2**	38.2	59.4	46.5	39.6	80.4	53.0	42.9	65.3	51.8	58.5	61.1	59.5
	60-40	20.0	89.0	32.7	37.0	68.8	48.2	38.9	87.4	53.8	43.4	76.8	55.4	55.2	74.4	63.2
	70-30	19.2	94.3	31.9	35.2	79.1	**48.7**	38.6	92.3	**54.4**	42.9	86.2	57.2	52.8	88.5	66.1
	80-20	18.3	97.5	30.9	32.2	88.6	47.3	37.8	96.1	54.2	41.9	92.7	**57.8**	50.8	97.0	**66.6**
IntAct + AIF	30-70	25.1	76.7	**37.8**	42.8	54.1	47.7	45.7	75.7	57.0	49.9	77.2	60.6	58.4	69.5	63.4
	40-60	24.5	78.9	37.4	42.3	56.5	48.3	46.1	79.2	58.3	49.2	79.0	60.7	58.2	72.8	64.6
	50-50	23.9	81.1	36.9	42.3	59.2	49.2	45.9	83.1	**59.1**	49	81.6	61.2	57.8	75.5	65.3
	60-40	23.1	83.8	36.1	41.8	63.3	50.3	44.9	85.3	58.8	48.4	84.7	61.6	56.8	79.2	66.1
	70-30	22.1	85.8	35.2	40.8	66.4	**50.5**	43.9	86.5	58.2	47.6	87.9	**61.8**	56.3	82.1	**66.7**
	80-20	21.3	88.3	34.3	39.6	69.9	**50.5**	42.9	89.8	58.1	46.0	91.6	61.3	54.0	84.9	66.0
IntAct + KUPS + AIF	30-70	26.6	72.1	38.8	43.8	50.8	47.0	48.1	78.6	59.7	51.1	75.3	60.9	60.2	63.7	61.8
	40-60	26.0	77.8	**39.0**	43.2	55.4	48.5	47.6	82.5	60.4	50.7	80.6	62.2	58.8	68.7	63.3
	50-50	25.5	81.6	38.8	44.8	56.2	49.8	46.0	83.9	59.4	51.4	78.7	62.2	60.3	72.2	65.6
	60-40	24.6	84.1	38.0	44.5	60.0	51.1	45.6	88.6	**60.2**	50.6	83.8	63.1	59.4	77.8	67.3
	70-30	23.6	86.7	37.1	43.3	64.4	51.8	44.3	90.5	59.5	49.3	88.8	**63.4**	59.4	83.3	**69.3**
	80-20	22.1	90.4	35.5	41.0	71.3	**52.0**	42.5	93.4	58.4	46.8	91.8	62.0	56.2	88.2	68.6
Thomas et al. (2011b)		22.3	81.3	35.0	38.7	76.0	51.2	45.6	92.9	61.2	42.6	88.3	57.3	53.7	93.3	68.1
Tikk et al. (2010)		28.3	86.6	42.6	62.8	36.5	46.2	56.9	68.7	62.2	71.0	52.5	60.4	79.0	57.3	66.4
Our system		34.3	74.0	46.9	70.8	22.5	34.2	63.3	61.3	62.3	70.0	46.0	55.5	82.4	45.7	58.8
Co-occurrence		17.1	100	29.3	26.2	100	41.5	37.6	100	54.7	41.0	100	58.2	49.7	100	66.4

Table 5: Results achieved with lexical features, trained on 10,000 distantly labeled instances and tested on 5 PPI corpora.

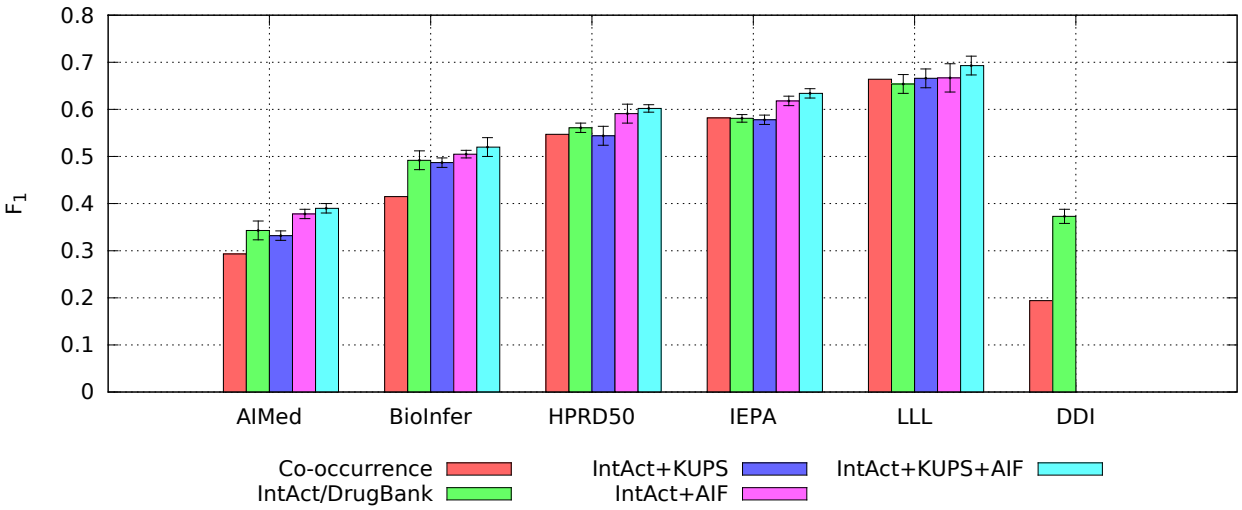

Figure 1: Comparison of four distant learning strategies with co-occurrence baseline. "IntAct/DrugBank" denotes the database used as source of supervision for PPI corpora and DDI corpus, respectively.

of performance in strategy IntAct+KUPS and the recovery after applying the AIF condition.

The strategy IntAct+KUPS+AIF results in a higher quality of data used for training and achieves the best performance in all five test cases thus proving the effectiveness of the novel condition. More knowledge is beneficial, but only when appropriate filtering of the data is applied.

Distantly supervised systems outperform co-occurrence results for all five PPI corpora. Considering the best performing strategy (IntAct+KUPS+AIF), F_1 measure of AIMed and BioInfer, for which we assume to have the most realistic pos/neg ratio, increased around 10 pp. HPRD50, IEPA and LLL have an improvement of 5.5 pp, 5.2 pp and 2.9 pp respectively, due to high fractions of positive instances (leading to a strong co-occurrence baseline).

Cross-learning[9] evaluation may be more realistic to be compared to distant-learning than cross validation (Airola et al., 2008). For AIMed and HPRD50 our approach performs on a par with Tikk et al. (2010) or better (up to 6 pp for BioInfer).

4.4 Drug-drug interaction

The problem of drug-drug interactions has not been previously explored in terms of distant supervision. It is noteworthy that DDI corpora are generated from web documents discussing drug effects which are in general not contained in MEDLINE. Hence, this evaluation corpus can be considered as out-domain and provides additional insights on the robustness of distant-supervision. The AIF setting is not evaluated for the DDI task, because only 1 of all 11,335 unique pairs describes a self interaction. In MEDLINE, only 7 sentences with multiple mentions of this drug (Sulfathiazole, DrugBank identifier DB06147) are found.

Table 6 gives an overview of the results for distant supervision on DDI, with the parameter of size of the training corpus and the pos/neg ratio. A slight increase in F_1 measure can be observed with additional training instances, both in case of using just lexical features and when dependency based features are additionally utilized (*e. g.* (*lex+dep*) from 36.2 % (5k) to 37.3 % (25k) in F_1 measure).

Accounting for dependency parsing features leads to an increase of 0.5 pp in F_1 measure, *i. e.* from 36.5 % to 37.0 % (10k) and 36.7% to 37.3 % (25k)), the latter being our best result obtained for weakly supervised DDI.

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

size	pos/neg	P	R	F_1
5k	30-70	**35.4**	32.4	33.7
	40-60	33.3	37.0	34.9
	50-50	31.9	41.7	36.0
	50-50 (lex+dep)	32.7	40.7	**36.2**
	60-40	30.1	46.6	**36.5**
	70-30	27.4	**51.8**	35.7
10k	30-70	**36.0**	34.4	34.9
	40-60	34.2	38.9	36.3
	50-50	32.9	41.0	**36.5**
	50-50 (lex+dep)	33.8	41.1	**37.0**
	60-40	30.8	44.8	36.4
	70-30	28.2	**48.7**	35.6
25k	30-70	**35.8**	35.0	35.3
	40-60	34.3	38.6	36.2
	50-50	33.2	41.1	**36.7**
	50-50 (lex+dep)	32.5	43.7	**37.3**
	60-40	31.7	42.6	36.3
	70-30	28.9	**47.2**	35.7
Co-occurrence		10.7	100	19.4

Table 6: Results for distant supervision with only lexical features on the DDI test corpus.

Compared to co-occurence, a gain of around 18 pp is achieved. Taking into account the high class imbalance of the DDI test set (see Table 1), which is most similar to AIMed corpus, the F_1 measure of 37.3 % is encouraging.

Figure 1 shows the results of PPI and DDI experiments in addition. The error bars denote the standard deviation over 5 differently sampled training corpora.

5 Discussion

This paper presents the application of distant supervision on the task to find protein-protein interactions and drug-drug interactions. The first is addressed using the databases IntAct and KUPS, the second using DrugBank.

More database knowledge does not necessarily have a positive impact on a trained model, appropriate instance selection methods need to be applied. This is demonstrated with the KUPS database and the automatic curation via auto-interaction filtering leading to state-of-the-art results for weakly supervised protein-protein interaction detection.

We present the first results of applying the distant supervision paradigm to drug-drug-interaction.

The results may seem comparatively limited in comparison to protein-protein interaction, but are encouraging when taking into account the imbalance of the test corpus and its differing source domain.

Future development of noise reduction approaches is important to make use of the full potential of available database knowledge. The results shown are encouraging that manual annotation of corpora can be avoided in other application areas as well. Another future direction is the investigation of specifically difficult structures, *e. g.* listings and enumerations of entities in a sentence.

Acknowledgments

We would like to thank the reviewers for their valuable feedback. Thanks to Sumit Madan and Theo Mevissen for fruitful discussions. 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. We acknowledge financial support provided by the IMI-JU, grant agreement no. 115191 (Open PHACTS).

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.

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

R. C. Bunescu and R. J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *HLT and EMNLP*.

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.

N. V. Chawla, N Japkowicz, and A. Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6:1–6.

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*.

J. De Las Rivas and C. Fontanillo. 2010. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6:e1000807+.

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.

L. Gong, R. P. Owen, W. Gor, R. B. Altman, and T. E. Klein. 2008. PharmGKB: an integrated resource of pharmacogenomic data and knowledge. *Curr Protoc Bioinformatics*, Chapter 14:Unit14.7.

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.

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.

Y. Li, X. Hu, H. Lin, and Z. Yang. 2010. Learning an enriched representation from unlabeled data for protein-protein interaction extraction. *BMC Bioinformatics*, 11(Suppl 2):S7.

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*.

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. Saetre, Y. Miyao, and J. Tsujii. 2009. A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. *EMNLP*, 1(1):121–130.

M. Miwa, R. Saetre, J. D. Kim, and J. Tsujii. 2010. Event extraction with complex event classification using rich features. *J Bioinform Comput Biol*, 8(1):131–146.

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.

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.

S. Van Landeghem, Y. Saeys, B. De Baets, and Y. Van de Peer. 2008. Extracting protein-protein interactions from text using rich feature vectors and feature selection. *SMBM*, pages 77–84.

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*.