Learning to Extract Protein–Protein Interactions using Distant Supervision

Philippe Thomas [1] **Illés Solt** [1,2] **Roman Klinger** [3] **Ulf Leser** [1]

[1]Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
{thomas,leser}@informatik.hu-berlin.de

[2]Dept. of Telecommunications and Media Informatics, Budapest University of Technology, Magyar tudósok körútja 2, 1117 Budapest, Hungary
solt@tmit.bme.hu

[3]Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Fraunhofer-Gesellschaft, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
roman.klinger@scai.fraunhofer.de

Abstract

Most relation extraction methods, especially in the domain of biology, rely on machine learning methods to classify a co-occurring pair of entities in a sentence to be related or not. Such an approach requires a training corpus, which involves expert annotation and is tedious, time-consuming, and expensive.

We overcome this problem by the use of existing knowledge in structured databases to automatically generate a training corpus for protein-protein interactions. An extensive evaluation of different instance selection strategies is performed to maximize robustness on this presumably noisy resource. Successful strategies to consistently improve performance include a majority voting ensemble of classifiers trained on subsets of the training corpus and the use of knowledge bases consisting of proven non-interactions. Our best configured model built without manually annotated data shows very competitive results on several publicly available benchmark corpora.

1 Introduction

Protein function depends, to a large degree, on the functional context of its interaction partners, *e.g.* other proteins or metabolites. Accordingly, getting a better understanding of protein-protein interactions (PPIs) is vital to understand biological processes within organisms. Several databases, such as IntAct, DIP, or MINT, contain detailed information about these interactions. To populate such databases, curators extract experimentally validated PPIs from peer reviewed publications (Ceol et al., 2010). Therefore, the automated extraction of PPIs from publications for assisting database curators has attracted considerable attention (Hakenberg et al., 2008; Airola et al., 2008; Tikk et al., 2010; Bui et al., 2010).

PPI extraction is usually tackled by classifying the $\binom{n}{2}$ undirected protein mention pairs within a sentence, where n is the number of protein mentions in the sentence. Classification of such pairs is often approached by machine learning (Airola et al., 2008; Tikk et al., 2010) or pattern-based methods (Fundel et al., 2007; Hakenberg et al., 2008) both requiring manually annotated corpora, which are costly to obtain and often biased to the annotation guidelines and corpus selection criteria. To overcome this issue, recent work has concentrated on distant supervision and multiple instance learning (Bunescu and Mooney, 2007; Mintz et al., 2009). Instead of manually annotated corpora, such approaches infer training instances from non-annotated texts using knowledge bases, thus allowing to increase the training set size by a few orders of magnitude. Corpora derived by distant supervision are inherently noisy, thus benefiting from robust classification methods.

1.1 Previous work

Distant supervision for relation extraction has recently gained considerable attention. Approaches usually focus on non-biomedical relations, such as "author wrote book" (Brin, 1999) or "person born in city" (Bunescu and Mooney, 2007). This work highlighted that it is feasible to train a classifier using distant supervision, which culminated in ideas

Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing, pages 25–32,
Hissar, Bulgaria, 15 September 2011.

to learn literally thousands of classifiers from relational databases like Freebase (Mintz et al., 2009; Yao et al., 2010), Yago (Nguyen and Moschitti, 2011), or Wikipedia infoboxes (Hoffmann et al., 2010).

So far, approaches in the biomedical domain on distant supervision focused on pattern learning (Hakenberg et al., 2008; Abacha and Zweigenbaum, 2010; Thomas et al., 2011). This is surprising as statistical machine learning methods are most commonly used for relation extraction. For example, only one of the five best performing systems in the BioNLP 2011 shared task relied on patterns (Kim et al., 2011).

The approaches described by Hakenberg et al. (2008) and Thomas et al. (2011) are those most related to our work. Both approaches learn a set of initial patterns by extracting sentences from MEDLINE potentially describing protein-protein interactions. Both methods use a knowledge base (IntAct) as input and search sentences containing protein pairs known to interact according to the knowledge base. However, these approaches generate patterns only for positive training instances and ignore the information contained in the remaining presumably negative instances.

PPI extraction is one of the most extensively studied relation extraction problems in the biomedical domain and is perfectly suited for a study on distant supervision as several corpora have been published in a common format (Pyysalo et al., 2008). Pyysalo et al. showed that the corpora differ in many aspects, *e.g.* annotation guidelines, average sentence length, and most importantly in the ratio of positive to negative training instances which accounts for about 50 % of all performance differences. Related work by Airola et al. (2008) and Tikk et al. (2010) revealed that the relation extraction performance substantially decreases when the evaluation corpus has different properties than the training corpus. A basic overview of the five most commonly used benchmark PPI corpora is given in Table 1.

So far, it is unclear how distant supervision performs on the difficult tasks of PPI extraction. For example Nguyen and Moschitti (2011) achieve a F_1 of 74.3 % on 52 different Yago relations using distant supervision. On the other hand, completely supervised state-of-the-art PPI extraction using manually labeled corpora achieve F_1 ranging from 56.5 % (AIMed) to 76.8 % (LLL) depending on the complexity of the corpus (Airola et al., 2008).

Corpus	Pairs		Class ratio
	positive	negative	$\frac{\text{positive}}{\text{negative}}$
AIMed	1,000	4,834	0.21
BioInfer	2,534	7,132	0.35
HPRD50	163	270	0.60
IEPA	335	482	0.73
LLL	164	166	0.99

Table 1: Overview of the 5 corpora used for evaluation. For state-of-the-art results on these corpora, see Table 3.

The contribution of the work described herein is as follows: We present different variations of strategies to utilize distant supervision for PPI extraction in Section 2. The potential benefit for PPI extraction is evaluated. Parameters taken into account are the number of training instances as well as the ratio of positive to negative examples. Finally, we assess if an ensemble of classifiers can further improve classification performance.

2 Methods

In this section, the workflow to extract interaction pairs from the databases and to generate training instances is described. Additionally, the configuration of the classifier applied to this corpus is given followed by the outline of the experimental setting.

2.1 Generation of training data

Training instances are generated as follows. All MEDLINE abstracts published between 1985 and 2011 are split into sentences using the sentence segmentation model by Buyko et al. (2006) and scanned for gene and protein names using GNAT (Hakenberg et al., 2011). In total, we find 1,312,059 sentences with 8,324,763 protein pairs. To avoid information leakage between training and test sets, articles contained in any of the benchmark evaluation corpora have been removed. This procedure excludes 7,476 ($< 0.1\%$) protein mention pairs from the training set. Protein pairs that are contained in the PPI knowledge base IntAct[1] (Aranda et al., 2010) are labeled as positive instances. Following a *closed world assumption*, protein pairs not contained in IntAct are considered as negative instances.

[1] As of Mar 24, 2010.

It is very likely, that both negative and positive instances contain a certain amount of mislabeled examples (false positives, false negatives). Therefore, we utilize different heuristics to minimize the amount of mislabeled instances. Firstly, we generate a list of words, which are frequently employed to indicate an interaction between two proteins [2]. This list is used to filter positive and negative instances such that positive instances contain at least one interaction word (*pos-iword*) and negative contain no interaction word (*neg-iword*). Application of both filters in combination is referred to as *pos/neg-iword*. Secondly, we assume that sentences with only two proteins are more likely to describe a relationship between these two proteins than sentences which contain many protein names. This filter is called *pos-pair*. For the sake of completeness, it is tested on negative instances alone (*neg-pair*) and on positive and negative instances in combination (*pos/neg-pair*). All seven experiments are summarized in Table 2.

2.2 Classification and experimental settings

For classification, we use a support vector machine with the shallow linguistic (SL) kernel (Giuliano et al., 2006) which has been previously shown to generate state-of-the-art results for PPI extraction (Tikk et al., 2010). This method uses syntactic features, *e.g.* word, stem, part-of-speech tag and morphologic properties of the surrounding words to train a classifier, but no parse tree information.

Setting	Feature:	Interaction word count		Pairs in sentence	
	Condition:	≥ 1	$= 0$	$= 1$	$= 1$
	Applied to:	positive	negative	positive	negative
baseline					
pos-iword		•			
neg-iword			•		
pos/neg-iword		•	•		
pos-pair				•	
neg-pair					•
pos/neg-pair				•	•

Table 2: Our experiment settings. Based on the number of interaction words and protein mention pairs in the containing sentence, we filter out automatically generated positive or negative example pairs not meeting the indicated heuristic condition. The dots indicate which filter is applied for which setting. For instance no filtering takes place for the baseline setting.

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

Classifiers are trained with a small subset from all 8 Million pairs, using 50,000 instances in all experiments except when stated differently. This allows us to investigate systematic differences between settings instead of generating and comparing only one prediction per setting.

Classifiers often tend to keep the same positive to negative ratio seen during the training phase. Class imbalance is therefore often acknowledged as a serious problem (Chawla et al., 2004). In our first experiments, we set the positive to negative ratio according to the overall ratio of positive to negative instances of all five corpora excluding the test corpus. This allows us to compare the results with the performance of various state-of-the-art kernel methods. As few publications provide results for the so-called cross-learning scenario, where a classifier is trained on the ensemble of four corpora and tested on the fifth corpus, we take the results from the extensive benchmark conducted by Tikk et al. (2010).

The influence of training class imbalance is evaluated separately by varying training set positive to negative ratios from 0.001 to $1,000$ using the best filtering strategy from the previous experiment.

As a sentence may describe a true protein interaction not present in the knowledge base, the closed world assumption is likely to be violated. Furthermore, not all mentions of a pair of proteins known to interact will describe an interaction. Thus both positively and negatively inferred training instances can be considered noisy. We therefore experimented with another filtering technique by using the Negatome database[3] (Smialowski et al., 2010) as an additional source to infer negative examples. Negatome contains a reference set of *non-interacting* protein pairs and is thus better suited to infer negative examples than our current method, which infers a negative example for all protein pairs not contained in the knowledge base according to the closed world assumption. However, reliable information about non-interaction is substantially more difficult to obtain and therefore the database contains far less entries than IntAct. From our 8 million protein pairs only 6,005 pairs could be labeled as negative. Additional negative training instances required for the training phase are therefore inferred using the closed world assumption.

[3] As of April 30, 2011.

Further, we evaluate how much training data is required to successfully train a classifier and if the classifier reaches a steady state after a certain number of training instances.

Finally, we evaluate whether a majority voting ensemble of 11 classifiers trained on randomly drawn training instances can further improve extraction quality. This strategy loosely follows a bagging strategy (Breiman, 1996), however, training instances are suspected to be less overlapping than using the standard bagging strategy.

2.3 Evaluation

For evaluation, we use the five benchmark PPI corpora listed in Table 1. Each training procedure, except for the ensemble experiments, is repeated 10 times randomly, thus resulting in 10 independent estimates for precision, recall, F_1, and area under the ROC curve (AUC). This allows for robust estimation of all evaluation metrics. Using single sided MannWhitney U test (Mann and Whitney, 1947) p-values for F_1 and AUC between two different models are calculated, with the null hypothesis that median of two samples is equal. Significance of Kendall correlation is determined using Best and Gipps (1974) with the null hypothesis that correlation equals zero. For all tests we assume a p-value of 0.01 to determine significance.

3 Results

Mean values for the seven different instance selection strategies (introduced in Table 2) are shown in Table 3. All strategies, except *neg-pair* filtering, lead to a higher AUC than 0.5. Thus six of seven settings perform better than randomly guessing. The advantage over random guessing is generally significant, except for three experiments in LLL. Many instance selection strategies for AIMed, BioInfer and HPRD50 outperform co-occurrence in terms of F_1. Several experiments outperform or at least perform on a par with the results from Thomas et al. (2011).

Co-occurrence outperforms significantly all seven settings for the two remaining corpora IEPA and LLL in F_1. This might have several reasons: First, these two corpora have the highest fraction of positive instances, therefore co-occurrence is a very strong baseline. Second, IEPA describes chemical relations instead of PPIs, thus our training instances might not properly reflect the syntactic property of such relations.

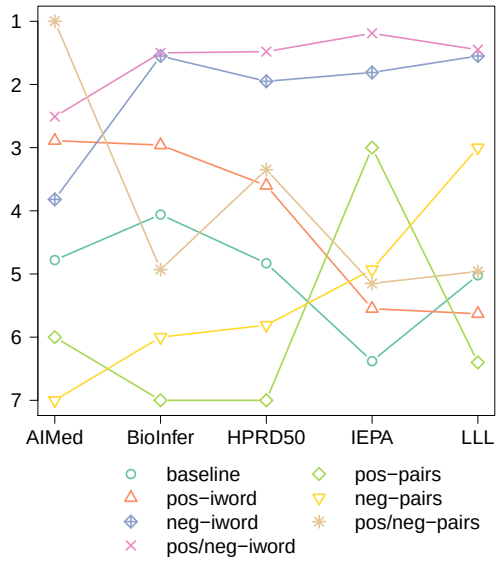


Figure 1: Average rank in F_1 for each experiment setting on the five corpora.

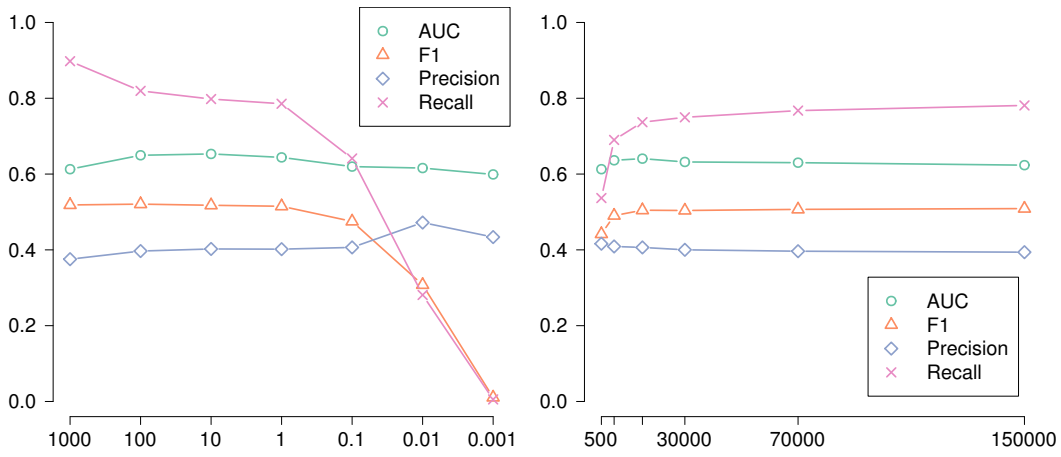
It is encouraging that on two corpora (BioInfer and HPRD50) the best setting performs about on par with the best cross-learning results from Tikk et al., which have been generated using manually annotated data and are therefore suspected to produce superior results.

For each corpus, we calculate and visualize the average rank in F_1 for the seven different strategies (see Figure 1). This figure indicates that pos/neg-iword and neg-iword filtering perform very well.

Repeating the previously described instance selection strategies (see Table 2) using Negatome to infer negative training instances lead to a small increase of 0.5 percentage points (pp) in F_1, due to an average increase of 1.1 pp in precision over all five corpora and seven settings (Results shown at bottom of Table 3). We also observe a tendency for increased AUC (0.9 pp). The largest gain in precision (3.5 pp) is observed between the two baseline results where no instance filtering is applied. Results for varied positive to negative ratios and for various amounts of training instances are also contained in the same table and visualized in Figure 2a and 2b respectively.

4 Discussion

The various settings introduced to filter out likely noisy training instances either improved precision or recall or both over the baseline using all automatically labeled instances for training (data shown in Table 3). In the following, we analyze and compare these settings.



(a) Positive to negative ratio (b) Sample size

Figure 2: Distribution of mean precision, recall, F_1, and AUC depending for the evaluation of class imbalance and sample size.

Method		AIMed				BioInfer				HPRD50				IEPA				LLL			
		AUC	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1
co-occurrence			17.8	(100)	30.1		26.6	(100)	41.7		38.9	(100)	55.4		40.8	(100)	57.6		55.9	(100)	70.3
supervised (Tikk et al.)		77.5	28.3	86.6	42.6	74.9	62.8	36.5	46.2	78.0	56.9	68.7	62.2	75.6	71.0	52.5	60.4	79.5	79.0	57.3	66.4
semi-supervised (Thomas et al.)			25.8	62.9	36.6		43.4	50.3	46.6		48.3	51.5	49.9		67.5	58.2	62.5		70.3	70.7	70.5
Setting	baseline	65.1	21.0	82.8	33.5	63.2	33.3	64.2	43.8	64.4	42.8	75.4	54.6	52.2	40.9	11.6	18.0	51.8	51.3	39.2	44.4
	pos-iword	66.6	21.8	82.6	34.5	67.5	38.4	60.8	47.1	67.5	45.5	76.5	57.1	53.8	48.6	12.3	19.6	51.6	50.0	37.0	42.2
	neg-iword	65.3	21.1	**91.1**	34.2	68.1	37.3	70.9	48.9	**73.4**	43.9	**93.6**	59.8	54.7	43.9	49.9	46.7	53.9	49.9	**77.4**	60.7
	pos/neg-iword	65.1	21.4	89.8	34.6	68.6	38.6	67.0	**49.0**	73.3	44.8	93.2	**60.5**	54.6	43.8	**53.2**	**48.0**	53.5	50.7	75.8	**60.8**
	pos-pairs	64.2	**29.3**	33.4	31.2	**69.8**	**57.8**	18.0	27.5	62.7	**47.9**	35.6	40.8	**66.6**	**54.9**	26.3	35.5	**63.2**	**68.2**	27.8	39.5
	neg-pairs	46.9	17.2	85.5	28.6	37.3	24.4	**85.6**	37.9	50.8	39.0	80.9	52.6	36.5	22.4	18.6	20.3	38.2	44.7	66.2	53.3
	pos/neg-pairs	**69.7**	23.6	82.3	**36.6**	62.0	32.8	60.6	42.5	69.2	46.5	75.2	57.5	56.0	43.4	13.3	20.3	54.3	54.5	37.9	44.6
Train pos/neg ratio	1,000	60.6	19.0	89.8	31.3	64.2	31.3	**84.6**	45.7	62.5	41.1	92.9	57.0	57.9	42.6	**88.3**	**57.3**	61.2	53.7	**93.3**	**68.1**
	100	63.9	20.0	88.7	32.7	69.0	35.5	77.8	48.7	71.5	44.2	91.9	59.6	**58.9**	45.6	65.6	53.7	**61.5**	53.1	85.8	65.6
	10	65.5	20.9	91.0	33.9	**71.2**	38.7	76.0	**51.2**	74.1	44.2	**95.8**	60.5	57.9	45.7	55.5	50.1	57.9	51.8	80.7	63.1
	1	65.6	21.4	**91.1**	34.7	70.0	38.6	71.3	50.1	**74.5**	44.3	95.5	**60.6**	56.1	45.0	55.5	49.7	55.7	51.6	79.3	62.5
	0.1	65.4	22.3	81.3	**35.0**	67.9	40.9	57.9	48.0	72.1	46.9	84.7	60.4	53.5	43.1	37.9	40.3	51.0	50.0	58.7	53.9
	0.01	**66.0**	26.9	46.7	34.1	66.5	46.9	24.7	32.4	70.4	59.7	48.5	53.4	52.8	**48.2**	8.3	14.2	52.2	**54.3**	12.3	19.7
	0.001	61.5	**41.4**	0.9	1.8	63.2	**63.0**	0.3	0.6	67.8	**72.5**	1.3	2.6	53.0	30.0	0.1	0.2	54.1	10.0	0.1	0.1
Train set size	500	63.4	**21.8**	71.5	33.4	65.9	39.8	44.6	41.9	67.6	**48.4**	67.4	56.2	55.5	45.4	31.5	36.7	54.0	**52.6**	53.4	52.7
	5,000	65.3	21.4	84.3	34.2	69.0	**39.9**	63.5	48.9	72.6	45.7	89.0	60.4	**56.8**	**46.1**	**41.9**	43.8	54.5	51.3	66.3	57.8
	15,000	**65.5**	**21.6**	87.9	**34.6**	**69.1**	39.7	65.1	**49.3**	**74.2**	45.6	92.9	**61.2**	55.8	44.5	47.4	45.9	**55.7**	51.9	75.1	61.3
	30,000	65.3	21.5	89.4	**34.6**	68.8	39.2	66.5	**49.3**	73.0	44.6	**93.1**	60.3	55.0	44.0	50.7	47.1	53.8	50.9	75.2	60.7
	70,000	65.1	21.3	90.7	**34.6**	68.6	38.1	67.4	48.7	73.2	44.2	92.1	59.8	54.2	43.7	55.0	48.7	53.9	50.9	78.6	61.8
	150,000	64.7	21.3	**91.3**	34.5	68.2	37.5	**68.1**	48.4	73.1	44.1	92.8	59.8	53.0	43.0	57.1	**49.1**	52.7	51.1	**81.3**	**62.7**
Setting (+Negatome)	baseline	65.9	22.2	79.6	34.7	65.7	36.8	58.6	45.2	67.6	46.7	74.0	57.3	54.9	47.5	12.7	20.0	54.8	53.6	36.3	43.2
	pos-iword	67.4	22.9	81.4	35.8	69.1	41.1	56.3	47.5	69.2	**47.9**	75.4	58.5	57.4	52.6	12.9	20.6	52.3	51.2	37.5	43.1
	neg-iword	65.3	21.1	**90.7**	34.3	68.8	38.1	69.6	**49.2**	**73.6**	44.6	92.1	60.1	55.6	44.4	51.7	47.8	55.2	51.3	**78.9**	62.2
	pos/neg-iword	65.1	21.4	89.4	34.6	68.8	38.8	66.9	49.1	73.2	44.8	**92.2**	**60.3**	55.3	44.2	**53.8**	**48.5**	54.9	52.2	77.9	**62.5**
	pos-pairs	64.6	**29.6**	33.7	31.5	**69.7**	**58.2**	18.3	27.8	62.2	48.5	35.5	41.0	**66.9**	**56.6**	30.7	39.7	**63.4**	**68.8**	28.1	39.9
	neg-pairs	47.0	17.2	84.9	28.6	37.0	24.3	**85.0**	37.8	50.9	38.4	79.8	51.9	36.0	22.4	18.5	20.3	38.5	45.1	66.0	53.5
	pos/neg-pairs	**69.8**	23.8	81.1	**36.8**	63.9	34.6	58.6	43.5	69.5	47.5	74.2	57.9	57.0	44.3	13.9	21.1	54.7	53.2	34.5	41.7

Table 3: Results of different instance selection strategies, different positive to negative ratios in the training set, sample size and employing Negatome as negative knowledge base.

Method	AIMed				BioInfer				HPRD50				IEPA				LLL			
	AUC	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1	AUC	P	R	F_1
co-occurrence		17.8	(100)	30.1		26.6	(100)	41.7		38.9	(100)	55.4		40.8	(100)	57.6		55.9	(100)	70.3
supervised (Tikk et. al)	77.5	28.3	86.6	42.6	74.9	62.8	36.5	46.2	78.0	56.9	68.7	62.2	75.6	71.0	52.5	60.4	79.5	79.0	57.3	66.4
semi-supervised (Thomas et. al)		25.8	62.9	36.6		43.4	50.3	46.6		48.3	51.5	49.9		67.5	58.2	62.5		70.3	70.7	70.5
mean of 11 runs	65.5	21.4	90.9	34.6	69.9	70.7	38.9	50.2	74.0	44.4	94.7	60.4	55.5	44.7	54.6	49.1	55.2	50.6	78.0	61.4
bagging over 11 runs		21.4	91.3	34.7		70.9	39.3	50.6		44.3	95.1	60.4		44.4	53.1	48.3		49.8	77.4	60.6

Table 4: Result of bagging over 11 classifier trained on different subsets. For comparison we show the average results for these 11 runs.

4.1 Pair count based settings

From our analysis it becomes apparent that no correlation between AUC and F_1 exists (Kendall's tau = 0.23, p-value = 0.55). For example pos-pair filtering significantly outperforms on three corpora all remaining six settings in terms of AUC, but the same setting supersedes almost no other setting in terms of F_1. A closer look reveals that on all five corpora the highest average precision can be achieved with this setting, at the cost of a decrease in recall. The pos-pair selection strategy results in fairly good training instances, but the decision hyperplane is not appropriately set.

The opposing filtering strategy (neg-pair) outperforms no other method in terms of AUC with an average score often below or at least close to a random classifier. However, this is expected, as the classifier tends to assign negative class labels to all sentences with exactly two protein mentions. This filter is in direct conflict to the original motivation and demonstrates that filtering must be performed carefully.

Even though positive and negative training instance filtering alone lead to almost no increase in F_1, the filtering of both negative and positive pairs leads to an overall improvement of 1.44 pp.

4.2 Interaction word based settings

All different combinations of instance filtering using a list of interaction words lead to an overall increase in F_1 and AUC. Filtering of positive and negative instances (pos/neg-iword) leads to the highest increase in AUC and with 11.8 pp in F_1, followed with 11.3 pp by exclusively filtering negative instances (neg-iword). Finally we observe only a marginal improvement of 1.3 pp when filtering positive instances (pos-iword).

4.3 Experiments with Negatome

A clear drawback of Negatome is the comparable small sample size of protein pairs. The number of confidently negative training instances could be increased by generalizing proteins across species using, for instance, Homologene. On our data set we could infer approximately 4,200 additional training instances. However, it is unclear if these derived instances are of the same quality than the Negatome data set. Another possibility is the usage of additional text repositories.

4.4 Effect of the pos/neg ratio

Table 3 clearly indicates that positive to negative ratio on training data affects performance of a classifier. Precision and recall strongly correlate with the pos/neg ratio seen in the training set. The observed correlation between recall and pos/neg ratio (Kendall's tau ranging from 0.524 to 1 for all five corpora) is expected, as the classifier tends to assign more test instances to the majority (positive) class. This procedure works best for corpora with many positive examples. A strong correlation (Kendall's tau ranging from -0.9 to -1.0) between precision and class ratio can be observed for AIMed, BioInfer, and HPRD50. Correlation for IEPA is close to zero and for LLL the correlation is even positive but not significant (p-value of 0.13). Overall, the observed influence is less pronounced than expected. For instance F_1 remains comparably robust with an average standard deviation of 2.6 pp for ratios between 0.1 and 10 . With more pronounced differences in the training ratio, a strong impact on F_1 can be observed.

In contrast to previous work on distant supervision, more noise on positive and negative instances is expected as database knowledge is suspected to be less complete and besides incompleteness knowledge evolves faster than for example for "president of country" relations. Other ap-

proaches often deal only with a strong noise on positive data, but little noise on negative instances. To avoid the double sided noise, we experimented with one class variations of SVM (Schölkopf et al., 2001) exploring the identical feature space. In one class classification only instances for the target set are available and the classifier searches a separating boundary between instances and yet unseen outliers. It has been previously demonstrated that one class classifiers are less sensitive to highly imbalanced data (Raskutti and Kowalczyk, 2004; Dreiseitl et al., 2010). However, in our experiments one class classifiers constantly achieved results close to random classification regardless of whether we used solely positive or negative instances for training.

4.5 Effect of training set size

For all corpora except for HPRD50 a monotonic increase in recall (Kendall's Tau of 1; p-value < 0.01) can be observed while increasing the training set. The negative correlation between precision and sample size is less pronounced but still observable for all Corpora (Kendall's Tau ranges between -0.552 and -1). Subsequently F_1 increases for corpora with many positive instances. Presumably, the problem of class imbalance gets more pronounced with additional instances.

4.6 Bagging

On the settings previously identified of being superior, we trained 11 classifiers using randomly sampled training sets. That is, a filtering of positive and negative instances for interaction words, a positive to negative ratio of 1, and a training size of 15,000 instances. The average results of the trained classifiers and the result of majority voting are given in Table 4. The ensemble classifier performs about on par with the mean of the individual classifiers and we observe no significant difference between the two approaches. However, a single classifier sometimes performs better or worse than the ensemble, whereas bagging always performs close to the mean result. Thus, bagging can be successfully applied for improving robustness of a classifier. Note that in our setting, all votes are of equal importance, thus neglecting the fact that some classifier perform generally better than others.

5 Conclusion

We investigated the use of distant supervision and demonstrated that it can be successfully adopted for domains where named entity recognition and normalization is still an unsolved issue and the closed world assumption might be an unsupported stretch. This is important, as named entity recognition and normalization is a key requirement for distant supervision. Distant supervision is therefore an extremely valuable method and allows training classifiers for virtually all kinds of relationships for which a database exists. We have proven here that results obtained without a manually annotated corpus are competitive with purely supervised methods, thus the tedious task of annotating a training corpus can be avoided.

Using five benchmark evaluation corpora – having diverse properties, annotated by different researchers adhering to differing annotation guidelines – offers a perfect opportunity to evaluate the robustness and usability of distant supervision. Our analysis reveals that background knowledge such as interaction words or "negative" knowledge bases such as Negatome consistently improves results across all five corpora. Also bagging had a positive impact on classifier robustness.

Surprisingly, class imbalance seams to be a less pronounced problem in distant supervision as often observed for supervised settings. One possible explanation might be that due to the noisy data, a classifier is less prone to over-fitting. So far, our experiments with one-class classification algorithms trained on positive or negative examples solely lead to disappointing results with AUC scores close to that of a random classifier. In future work, we plan to investigate if other one-class algorithms can be successfully adapted for relation extraction in a distant supervised setting.

Instance selection seems to have the largest impact for this approach. Instead of simple heuristics, we plan to investigate the usability of syntactic patterns to further discriminate positive and negative instances (Bui et al., 2010).

References

A.B. Abacha and P. Zweigenbaum. 2010. Automatic Extraction of semantic relations between medical entities: Application to the treatment relation. In *Proc. of SMBM 2010*.

A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Gin-

ter, 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.

B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. 2010. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 38:525–531, Jan.

D. J. Best and P. G. Gipps. 1974. Algorithm AS 71: The Upper Tail Probabilities of Kendall's Tau. *Journal of the Royal Statistical Society.*, 23(1):pp. 98–100.

L. Breiman. 1996. Bagging Predictors. *Machine Learning*, 24(2):123–140.

S. Brin. 1999. Extracting Patterns and Relations from the World Wide Web. Technical Report 1999-65, Stanford InfoLab, November.

Q. Bui, S. Katrenko, and P. M. A. Sloot. 2010. A hybrid approach to extract protein-protein interactions. *Bioinformatics*, Nov.

R. C. Bunescu and R. J. Mooney. 2007. Learning to Extract Relations from the Web using Minimal Supervision. In *Proc. of ACL'07*.

E. Buyko, J. Wermter, M. Poprat, and U. Hahn. 2006. Automatically Adapting an NLP Core Engine to the Biology Domain. In *Proc. of ISMB'2006*.

A. Ceol, A. Chatr-aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. 2010. MINT, the molecular interaction database: 2009 update. *Nucl. Acids Res.*, 38(suppl1):D532–539.

N.V. Chawla, N. Japkowicz, and A. Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6.

S. Dreiseitl, M. Osl, C. Scheibböck, and M. Binder. 2010. Outlier Detection with One-Class SVMs: An Application to Melanoma Prognosis. *AMIA Annu Symp Proc*, 2010:172–176.

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

C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *Proc. of EACL'06*.

J. Hakenberg, C. Plake, L.Royer, H. Strobelt, U. Leser, and M. Schroeder. 2008. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol*, 9 Suppl 2:S14.

J. Hakenberg, M. Gerner, M. Haeussler, I. Solt, C. Plake, M. Schroeder, G. Gonzalez, G. Nenadic, and C.M. Bergman. 2011. The GNAT library for local and remote gene mention normalization. *Bioinformatics*, Aug.

R. Hoffmann, C. Zhang, and D. Weld. 2010. Learning 5000 relational extractors. In *Proc. of ACL'10*, pages 286–295.

J. Kim, Y. Wang, T. Takagi, and A. Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proc. of BioNLP-ST 2011*, pages 7–15.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. of AFNLP*, volume 2 of *ACL '09*, pages 1003–1011.

T.V. Nguyen and A. Moschitti. 2011. End-to-End Relation Extraction Using Distant Supervision from External Semantic Repositories. In *ACL'2011*, pages 277–282.

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.

B. Raskutti and A. Kowalczyk. 2004. Extreme rebalancing for SVMs: a case study. *ACM SIGKDD Explorations Newsletter*, 6(1):60–69.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Comput*, 13(7):1443–1471, Jul.

P. Smialowski, P. Pagel, P. Wong, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, T. Rattei, D. Frishman, and A. Ruepp. 2010. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, 38:D540–D544, Jan.

P. Thomas, S. Pietschmann, I. Solt, D. Tikk, and U. Leser. 2011. Not all links are equal: Exploiting dependency types for the extraction of protein-protein interactions from text. In *Proc. of BioNLP'11*.

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.

L. Yao, S. Riedel, and A. McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proc. of EMNLP'10*, pages 1013–1023.