Committee-based Selection of Weakly Labeled Instances for Learning Relation Extraction

Tamara Bobić[1,2] and Roman Klinger[1,3]

[1] Fraunhofer SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2] B-IT, University of Bonn, Dahlmannstraße 2, 53113 Bonn, Germany
[3] Semantic Computing, CITEC, Bielefeld University, 33615 Bielefeld, Germany**
tamara.bobic@scai.fraunhofer.de
rklinger@cit-ec.uni-bielefeld.de

Abstract. Manual annotation is a tedious and time consuming process, usually needed for generating training corpora to be used in a machine learning scenario. The distant supervision paradigm aims at automatically generating such corpora from structured data. The active learning paradigm aims at reducing the effort needed for manual annotation. We explore active and distant learning approaches jointly to limit the amount of automatically generated data needed for the use case of relation extraction by increasing the quality of the annotations.
The main idea of using distantly labeled corpora is that they can simplify and speed-up the generation of models, *e. g.* for extracting relationships between entities of interest, while the selection of instances is typically performed randomly. We propose the use of query-by-committee to select instances instead. This approach is similar to the active learning paradigm, with a difference that unlabeled instances are weakly annotated, rather than by human experts. Different strategies using low or high confidence are compared to random selection. Experiments on publicly available data sets for detection of protein-protein interactions show a statistically significant improvement in F_1 measure when adding instances with a high agreement of the committee.

1 Introduction

Developing manually annotated training corpora for information extraction tasks like named entity recognition or relation extraction is tedious, time-consuming and therefore expensive work. One approach to overcome these issues is to build weakly supervised information extraction models, *e. g.* by using distantly labeled text, as proposed by [1]. This paradigm has shown to achieve reasonable, competitive results [2–4].

Unfortunately, the assumption that co-occurring entities in a sentence are related if they are mentioned in a source of distant supervision (for instance a database) does not hold in general. Therefore, such automatically annotated data sets are typically noisy. Methods addressing this issue include filtering approaches by formulating heuristics [5, 6] or classifying if the instance is actually representing a positive example [7]. In addition, though there is a huge amount of data available, the instances used for training may be uninformative and redundant.

** Present address of Roman Klinger. This work was performed at Fraunhofer SCAI.

In this paper, we explore and discuss the idea of making use of the active learning paradigm [8–10] to select meaningful distantly labeled instances from a large pool. Active learning is a strategy for reducing the overall annotation effort without diminishing the system's performance. It is a semi-automated approach where only data points that are considered to be most informative are presented to the "oracle" (usually a human expert) for manual annotation. We focus on the use case of classifying pairs of named entities as interacting or non-interacting. Objectives are to avoid using non-informative or misleading instances and to reduce the amount of data needed to train a model which leads to less complex models, as a lower number of features is generated.

Other approaches to circumvent the need for manual work include unsupervised machine learning approaches relying on discovering structure in unlabeled data. Although automatic generation of rule sets [11, 12], dictionaries [13, 14], or clusters [15] is effective, unsupervised approaches are often suffering from a limited performance in comparison to supervised approaches. Semi-supervised learning aims at obtaining good performance at a low cost by combining (potentially large) amounts of unlabeled data with human supervision. In the work by [16–18], a relatively small labeled seed set is used for learning initial patterns, while additional prediction rules are generated through further iterations. Such approach has the advantage of considerably reducing the amount of work for human annotators, however, due to its dependency on the initial seed set, the generalizability may be limited. A combined approach including both semi-supervised and active learning by [19] tends to increase the accuracy of label predictions, while keeping the human interference at minimum. In contrast, active learning aims at limiting the amount of work for a manual annotator. The fundamental idea is to make use of an estimator for selecting the instances to be shown to the annotator. That can be based on minimization of expected variance [20], uncertainty sampling [21], or query-by-committee [22], amongst others.

In the following, we shortly introduce interaction classification in Section 2.1 and explain how informative instances could be distinguished from redundant ones in Section 2.2. The results in Section 3 are based on evaluations of the proposed method on a publicly available data set for protein-protein and drug-drug interaction detection. We end with a discussion and summary.

2 Methods

2.1 Interaction Classification

As common, 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. A sentence with n entities contains at most $\binom{n}{2}$ interacting pairs. We are using a linear support vector machine classifier [23] with lexical features, *i. e.*, bag-of-words and n-grams, 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 [24]. Additionally, dictionary-based domain specific trigger words are taken into account. For details of the configuration, we refer to [5].

2.2 Committee-based Selection of Instances

We assume a small manually annotated training set to be available, the generation of which would require only a moderate amount of work. This set should provide the information for distinction between helpful and unhelpful or even misleading pairs of entities, *i. e.*, relation instances.

In active learning, a human annotator is asked to provide classification for an instance. The number of instances presented to the annotator is to be minimized, while the annotation is assumed to be perfect. Therefore, instances of highest expected information content (given the existing data) are normally chosen [25]. Here, we replace the human annotator by a predictor of limited knowledge, *i. e.*, the distantly labeled data. Retrieving high quality annotation for highly informative instances is possible from a human annotator, however, labels coming from the database are not always correct. Therefore, for the distantly labeled data the relation between the annotation quality and the information gain, given the seed set which is used for selecting instances is of importance. The hypothesis is, that a higher quality of the data annotation is positively correlated to a lower information gain. Therefore, in the distant supervision setting, there may be a trade-off between quality and information gain.

We follow several strategies to rank the instances and select the preferred ones. All strategies are based on a query-by-committee approach [26]. The training set for each committee member $c \in \mathcal{C}$ is selected by sampling n times with replacement, leading to approximate use of 63% of the available instances for each committee member [27] (where n is the number of available instances).

The agreement of the committee $\mathcal{C}$ concerning an instance i is measured as

$$u_{\mathcal{C}}(i) = \frac{|\mathbf{1}^i_{\mathcal{C}} - \mathbf{0}^i_{\mathcal{C}}|}{|\mathcal{C}|} ,$$

where $\mathbf{0}^i_{\mathcal{C}}$ denotes the number of committee members predicting "no interaction" and $\mathbf{1}^i_{\mathcal{C}}$ accordingly for predicting "interaction" for instance i.

High agreement of the committee is interpreted as high confidence regarding the label of an instance [22]. Let x_i be a random value from the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with variance σ^2 and mean $\mu = 0$.

1. Rank descending by $u_{\mathcal{C}}(i)$ (prefer instances of high confidence).
2. Rank ascending by $u_{\mathcal{C}}(i)$ (prefer instances of low confidence).
3. Rank descending by $u_{\mathcal{C}}(i) + x_i$.

The idea of the first strategy is to select instances which are most similar (and therefore have a high quality) to the manually annotated training data, but may not lead to useful information. The second strategy pertains to the common approach used in active learning, where instances that are dissimilar to known ones may bring high information gain. The motivation of the third strategy is to take instances into account which are similar to the manually annotated data, but allowing the chance of having additional "novel" aspects.

3 Results

The silver standard corpora[4] of [29] consisting of 200,000 protein entity pair and 200,000 drug entity pair mentions are used as a source of weakly labeled data to draw training instances from. The text source are abstracts from MEDLINE[5]. They are labeled making use of the databases IntAct [30] and KUPS [31]. An overview of these corpora is given in Table 1.

Table 1: Weakly labeled PPI and DDI corpora.

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

Table 2: PPI and DDI corpora.

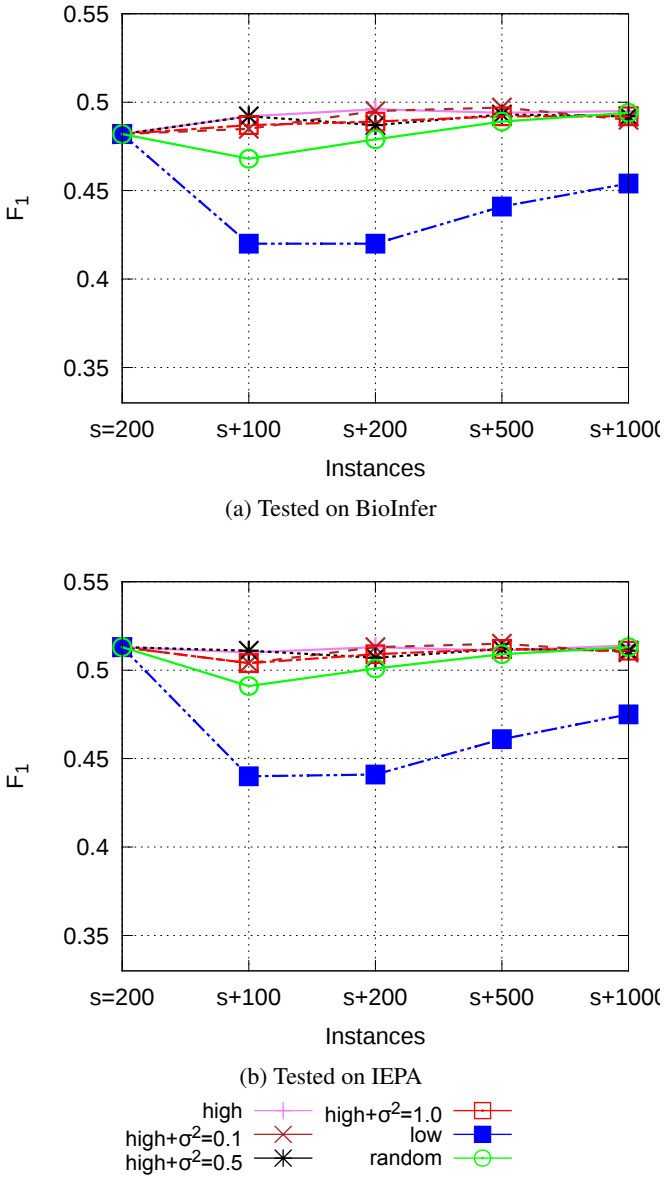
Corpus	Pos. pairs	Neg. pairs	Total
BioInfer	2,534	7,132	9,666
HPRD50	163	270	433
IEPA	335	482	817
LLL	164	166	330
DDI train	2,400	21,411	23,811
DDI test	755	6,275	7,030

The publicly available manually annotated corpora for protein-protein interaction HPRD50 [32], LLL [33], BioInfer [34], and IEPA [35] are used for training and testing. In case of drug-drug-interaction, the corpus published by [36] is used (being divided into train and test set). Table 2 shows an overview of the manually annotated PPI and DDI corpora.

The experimental setting is as follows. For each of the corpora, 200 pairs (instances) are randomly sampled from manually annotated data, corresponding to the seed set in an active learning setting. Based on these, classifiers are trained on sub-samples to predict if

[4] The term "silver standard" refers to an automatically annotated resource, contrary to a gold standard with (by definition) perfect annotation [28].

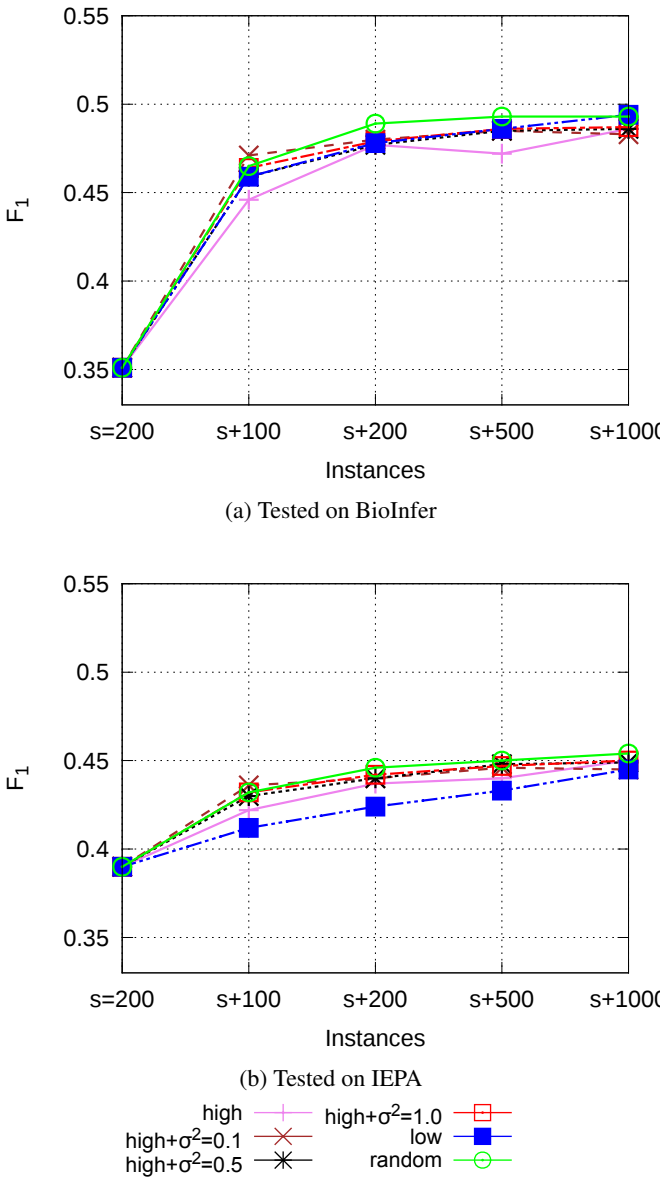
[5] `http://www.ncbi.nlm.nih.gov/pubmed/`



(a) Tested on BioInfer

(b) Tested on IEPA

Fig. 1: Confidence predictor based on LLL and testing on BioInfer and IEPA corpora. The model built on 200 manually annotated instances is compared against training with 100, 200, 500, and 1000 additional weakly labeled instances.



(a) Tested on BioInfer

(b) Tested on IEPA

Fig. 2: Confidence predictor based on HPRD50 and testing on BioInfer and IEPA corpora. The model built on 200 manually annotated instances is compared against training with 100, 200, 500, and 1000 additional weakly labeled instances.

an entity pair is in relation or not. This committee is used to get a score for agreement or disagreement of predicting data from the weakly labeled set. Depending on the selection strategy, 100, 200, 500, and 1000 instances are selected. A classifier is trained on the seed set of 200 instances, as well as on this set unified with the weakly labeled instances. Note that these are not multiple iterations, but separate experiments of active learning. Each of the experiments is repeated 10 times and the average value reported to be able to measure stability as well.

Figures 1 and 2 show the results for training on LLL and HPRD50, while testing on BioInfer and IEPA. The results between same models tested on different corpora are similar (compare 1a with 1b and 2a with 2b). In the case of training on LLL, worst strategy is selecting instances with the lowest confidence, followed by random. Best results are seen for the selection by high confidence, while adding Gaussian noise does not lead to big differences; for adding 100 weakly labeled instances, using $\sigma^2 = 0.5$ works best. All methods based on high confidence are outperforming the random baseline significantly in this step ($\alpha < 0.05$). Comparison of training with LLL and HPRD50 reveals notable differences when adding a low number of instances: for LLL, random and low confidence selection leads to a decrease. For HPRD50, all selection methods have a positive impact. Training on HPRD50 does not provide a clear difference between the selection strategies; low leads to worst results, random and high with some noise to the best. These differences are not significant.

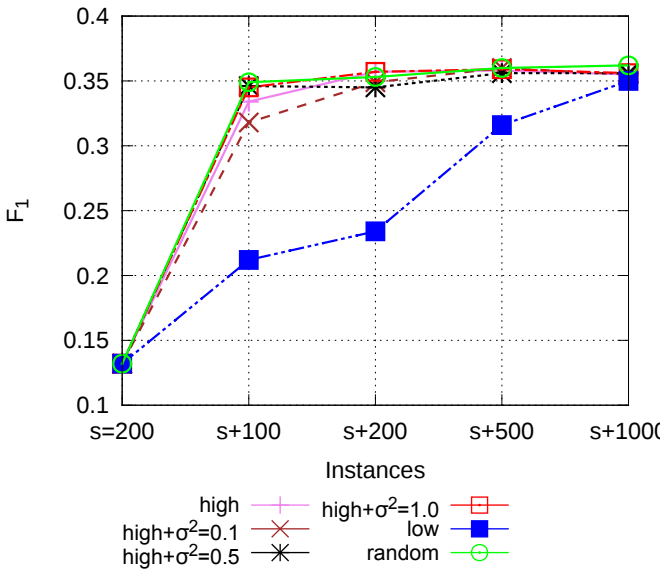


Fig. 3: Results for DDI, starting with 200 manually annotated instances and comparing against training with 100, 200, 500, and 1000 additional weakly labeled instances.

Obviously, adding a higher number of instances leads to lower impact of the selection strategy. Evaluating different strategies on DDI leads to results similar to training on HPRD50, as shown in Figure 3.

The results, especially for the seed set sampled from LLL (*cf.* Figure 1a), shows the best results using instances similar to the seed set (by means of the committee trained on the seed set having a high agreement). To prove the hypothesis that high quality of annotation is leading to a lower information gain and vice versa, the Pearson correlation coefficient of the committee prediction (based on a seed set of 200 LLL instances) and the labels from the database (the distantly labeled PPI corpus) are reported in Figure 4. For each confidence interval, 500 instances are sub-sampled respectively (1 refers to agreement among the committee; 0 refers to no agreement). A high correlation of the database labels for instances selected to be similar with the seed set can be observed. There is nearly no correlation for instances selected with low confidence.

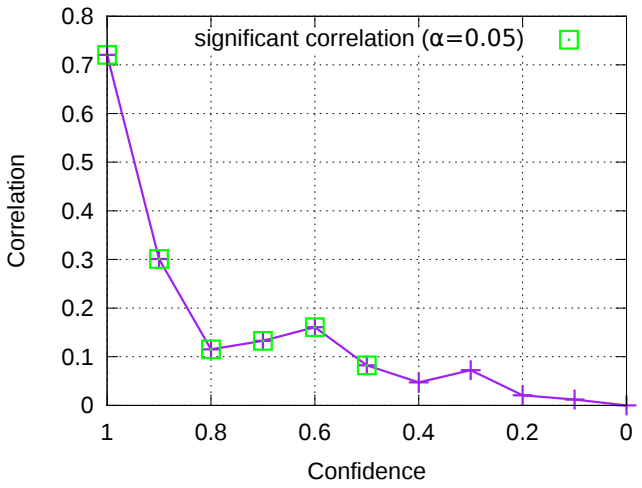


Fig. 4: Correlation of labels predicted by the committee and given from the database (distant supervision) plotted for different confidence threshold values. The correlation is calculated by Pearson's coefficient and the predictions are made using LLL as the training corpus.

4 Discussion and Summary

The results are motivating for a subset of corpora, shown on LLL here. The reason for the difference when training on HPRD50 is presumably the ratio between positive and negative examples; LLL is the most balanced set with a ratio of 1.01. HPRD50 has a ratio of 1.44. Due to the same reason, the initial results are that different (between 0.35

and 0.38 for HPRD50 in comparison to 0.48 to 0.51 for LLL) – the seed sub-sample only includes a low number of positive examples. Committee-based selection increases performance significantly on LLL.

It is notable for this corpus that the committee-based selection of weakly labeled instances leads to comparable results when using 100 additional instances chosen by high strategy and 500–1000 instances chosen randomly. Therefore it needs to be pointed out that, surprisingly, instances being similar to the seed set lead to best results. The reason is a strong correlation of database labels with the committee predictions in cases where the committee fully agrees.

Selecting instances labeled with the highest confidence by the committee appears to be the favourable decision in most cases to deal with the noisy data generated by the distant supervision approach. Such strategy is not common in the active learning paradigm, however, the prevailing in favor of "safe" instances confirms the hypothesis that a higher quality of the data annotation is correlated to a lower information gain.

It needs to be investigated further whether this methodology harms the generalizability of the model. An analysis of the positions of the support vectors from the seed set and from the weakly labeled set may allow insight in this concern.

Future work includes the evaluation of additional parameters. In comparison to active learning with a human annotator, additional knowledge about the weakly labeled data is available. Therefore, the ratio of positive and negative examples needs to be investigated further. Similarly, the characteristics of the seed set need to be analyzed in more detail. Furthermore, correlation of instances chosen to be in the seed set needs to be inspected, as well as the possible correlation between the seed set instances and those that are to be added.

Acknowledgments

T. Bobić was funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School. R. Klinger was partially funded by the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement no. 248726.

References

1. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. Proceedings of the International Conference on Intelligent Systems for Molecular Biology (1999)
2. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Conference of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. (2009)
3. Riedel, S., Yao, L., McCallum, A.: Modeling Relations and Their Mentions without Labeled Text. In: Proceedings of the European Conference on Machine Learning and Principles and Practice in Knowledge Discovery from Databases. (2010)

4. Thomas, P., Solt, I., Klinger, R., Leser, U.: Learning Protein Protein Interaction Extraction using Distant Supervision. In: Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing, Recent Advances in Natural Language Processing. (2011)
5. Bobić, T., Klinger, R., Thomas, P., Hofmann-Apitius, M.: Improving distantly supervised extraction of drug-drug and protein-protein interactions. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, European Chapter of the Association for Computational Linguistics. (2012)
6. Buyko, E., Beisswanger, E., Hahn, U.: The extraction of pharmacogenetic and pharmacogenomic relations–a case study using PharmGKB. Pacific Symposium on Biocomputing (2012)
7. Yao, L., Riedel, S., McCallum, A.: Collective Cross-Document Relation Extraction Without Labeled Data. In: Empirical Methods in Natural Language Processing. (2010)
8. Olsson, F.: A literature survey of active machine learning in the context of natural language processing. Technical Report T2009:06, Swedish Institute of Computer Science (SICS), Kista, Sweden (2009)
9. Settles, B.: Active learning literature survey. Technical report, University of Wisconsin-Madison (2010) Computer Sciences Technical Report 1648.
10. Tomanek, K.: Resource-aware annotation through active learning. PhD thesis, TU Dortmund University (2010)
11. Brill, E.: Unsupervised learning of disambiguation rules for part of speech tagging. In: Proceedings of the Third Workshop on Very Large Corpora, Association for Computational Linguistics. (1995)
12. Hassan, H., Hassan, A., Emam, O.: Unsupervised information extraction approach using graph mutual reinforcement. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2006)
13. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999)
14. Nadeau, D., Turney, P.D., Matwin, S.: Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. In: Proceedings of the International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence, Berlin, Heidelberg, Springer-Verlag (2006)
15. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: Proceedings of the Association for Computational Linguistics. (2004)
16. Riloff, E., Jones, R.: Learning dictionaries for information extraction by multi-level bootstrapping. In: Proceedings of the National Conference on Artificial intelligence and the Innovative Applications of Artificial Intelligence Conference. (1999)
17. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Proceedings of the ACM conference on Digital libraries. (2000) 85–94
18. Ravi, S., Baldridge, J., Knight, K.: Minimized models and grammar-informed initialization for supertagging with highly ambiguous lexicons. In: Proceedings of the Association for Computational Linguistics. (2010)
19. Wu, T., Pottenger, W.M.: A semi-supervised active learning algorithm for information extraction from textual data. Journal of the American Society for Information Science and Technology **56** (2005)
20. Cohn, D., Gharahamani, Z., Jordan, M.: Active learning with statistical models. Artifical Intelligence Research (1996)
21. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Empirical Methods on Natural Language Processing. (2008)
22. Freund, Y., Seung, S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning (1997)

23. Fan, E., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A Library for Large Linear Classification. Machine Learning Research **9** (2008) 1871–1874
24. Li, Y., Hu, X., Lin, H., Yang, Z.: Learning an enriched representation from unlabeled data for protein-protein interaction extraction. BMC Bioinformatics **11** (2010) S7
25. Nicholas, R., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: International Conference on Machine Learning. (2001)
26. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the workshop on computational learning theory. (1992)
27. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC (1993)
28. Rebholz-Schuhmann, D., Jimeno-Yepes, A.J., van Mulligen, E.M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Tomanek, K., Beisswanger, E., Hahn, U.: The CALBC Silver Standard Corpus for Biomedical Named Entities – A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. In: Proceedings of the Conference on Language Resources and Evaluation. (2010)
29. Thomas, P., Bobić, T., Hofmann-Apitius, M., Leser, U., Klinger, R.: Weakly labeled corpora as silver standard for drug-drug and protein-protein interaction. In: Workshop on Building and Evaluating Resources for Biomedical Text Mining, Language Resources and Evaluation Conference. (2012)
30. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., Hermjakob, H.: The IntAct molecular interaction database in 2012. Nucleic Acids Research **40** (2012)
31. Chen, X., Jeong, J.C., Dermyer, P.: KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. Nucleic Acids Res **39** (2010)
32. Fundel, K., Kuffner, R., Zimmer, R.: RelEx–Relation extraction using dependency parse trees. Bioinformatics **23** (2007)
33. Nedellec, C.: Learning language in logic-genic interaction extraction challenge. In: Learning Language in Logic, International Conference on Machine Learning. (2005)
34. Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., Salakoski, T.: BioInfer: A Corpus for Information Extraction in the Biomedical Domain. BMC Bioinformatics (2007)
35. Ding, J., Berleant, D., Nettleton, D., Wurtele, E.: Mining MEDLINE: abstracts, sentences, or phrases? Pacific Symposium on Biocomputing (2002)
36. Segura-Bedmar, I., Martínez, P., Sanchez-Cisneros, D.: The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In: Challenge Task on Drug-Drug Interaction Extraction 2011. (2011)