WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting

Philippe Thomas **Mariana Neves** **Tim Rocktäschel** **Ulf Leser**
Humboldt-Universität zu Berlin
Knowledge Management in Bioinformatics
Unter den Linden 6
Berlin, 10099, Germany
{thomas,neves,trocktae,leser}@informatik.hu-berlin.de

Abstract

This work describes the participation of the WBI-DDI team on the SemEval 2013 – Task 9.2 DDI extraction challenge. The task consisted of extracting interactions between pairs of drugs from two collections of documents (DrugBank and MEDLINE) and their classification into four subtypes: advise, effect, mechanism, and int. We developed a two-step approach in which pairs are initially extracted using ensembles of up to five different classifiers and then relabeled to one of the four categories. Our approach achieved the second rank in the DDI competition. For interaction detection we achieved F_1 measures ranging from 73 % to almost 76 % depending on the run. These results are on par or even higher than the performance estimation on the training dataset. When considering the four interaction subtypes we achieved an F_1 measure of 60.9 %.

1 Introduction

A drug-drug interaction (DDI) can be described as interplay between drugs taken during joint administration. DDIs usually lead to an increase or decrease in drug effects when compared to isolated treatment. For instance, sildenafil (Viagra) in combination with nitrates can cause a potentially live-threatening decrease in blood pressure (Cheitlin et al., 1999). It is therefore crucial to consider potential DDI effects when co-administering drugs to patients. As the level of medication generally is raising all over the world, the potential risk of unwanted side effects, such as DDIs, is constantly increasing (Haider et al., 2007).

Only a fraction of knowledge about DDIs is contained in specialized databases such as DrugBank (Knox et al., 2011). These structured knowledge bases are often the primary resource of information for researchers. However, the majority of new DDI findings are still initially reported in scientific publications, which results in the situation that structured knowledge bases lag behind recently published research results. Thus, there is an urgent need for researchers and database curators to cope with the fast growth of biomedical literature (Hunter and Cohen, 2006).

The SemEval 2013 – Task 9.2 (Extraction of Drug-Drug Interactions from BioMedical Texts) is a competitive evaluation of methods for extracting mentions of drug-drug interactions from texts (Segura-Bedmar et al., 2013). For training, the organizers provide a corpus annotated with drugnames and interactions between them. This corpus is composed of 572 articles collected from DrugBank and 142 PubMed abstracts. Interactions are binary (always between two drugs) and undirected, as target and agent roles are not annotated. Furthermore, the two interacting drugs are always mentioned within the same sentence. In contrast to the previous DDI-challenge 2011 (Segura-Bedmar et al., 2011), four different DDI-subtypes (advise, effect, mechanism, and int) have been introduced. Details about the four subclasses can be found in the task's annotation guideline.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 628–635, Atlanta, Georgia, June 14-15, 2013.

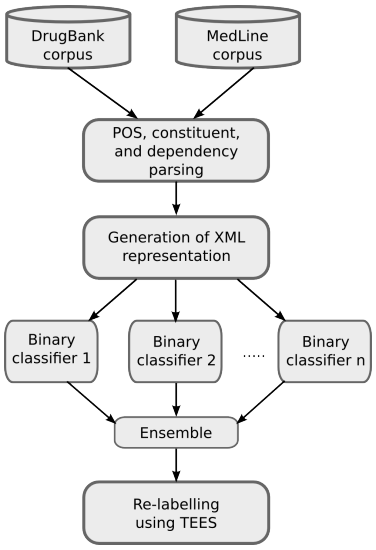

Figure 1: Workflow developed for the SemEval 2013 Task 9.2 challenge.

2 Methods

Binary relationship extraction is often tackled as a pair-wise classification problem, where all $\binom{n}{2}$ co-occurring entities in a sentence are classified as interacting or not. To account for the four different subtypes of DDIs, the problem definition could be translated into a multiclass classification problem between all co-occurring entities.

Contrary to that, we propose a two step strategy: First, we detect general drug-drug interactions regardless of subtype using a multitude of different machine-learning methods. The output of these methods is aggregated using a majority voting approach. Second, detected interactions are re-classified into one of the four possible DDI categories. The latter is referred to as DDI relabeling throughout this paper. A detailed view on the proposed workflow is depicted in Figure 1.

2.1 Preprocessing

Sentences have been parsed using Charniak-Johnson PCFG reranking-parser (Charniak and Johnson, 2005) with a self-trained re-ranking model augmented for biomedical texts (McClosky, 2010). Resulting constituent parse trees have been converted into dependency graphs using the Stanford converter (De Marneffe et al., 2006). In the last step, we created an augmented XML using the open source framework from Tikk et al. (2010). This XML file encompasses tokens with respective part-of-speech tags, constituent parse tree, and dependency parse tree information. This format has been subsequently transformed into a related XML format[1] used by two of the utilized classifiers. Properties of the training corpus are shown for DrugBank and MEDLINE in Table 1.

Corpus	Sentences	Pairs		
		Positive	Negative	Total
DrugBank	5,675	3,788	22,217	26,005
MEDLINE	1,301	232	1,555	1,787

Table 1: Basic statistics of the DDI training corpus shown for DrugBank and MEDLINE separately.

2.2 Machine Learning Methods

Tikk et al. (2010) systematically analyzed nine different machine learning approaches for the extraction of undirected binary protein-protein interactions. This framework has been successfully applied to other domains, such as the I2B2 relation extraction challenge (Solt et al., 2010), the previous DDI extraction challenge (Thomas et al., 2011), and to the extraction of neuroanatomical connectivity statements (French et al., 2012).

Drug entities are blinded by replacing the entity name with a generic string to ensure the generality of the approach. Without entity blinding drug names are incorporated as features, which clearly affects generalization capabilities of a classifier on unseen entity mentions (Pyysalo et al., 2008).

We decided to use the following methods provided by the framework: All-paths graph (APG) (Airola et al., 2008), shallow linguistic (SL) (Giuliano et al., 2006), subtree (ST) (Vishwanathan and Smola, 2002), subset tree (SST) (Collins and Duffy, 2001), and spectrum tree (SpT) (Kuboyama et al., 2007) method. The SL method uses only shallow linguistic features, *i.e.*, token, stem, part-of-speech tag and morphologic properties of the surrounding words. APG builds a classifier using surface features and a weighting

[1] https://github.com/jbjorne/TEES/wiki/Interaction-XML

scheme for dependency parse tree features. The remaining three classifier (ST, SST, and SpT) build kernel functions based on different subtree representations on the constituent parse tree. To calculate the constituent–tree kernels ST and SST we used the SVM-LIGHT-TK toolkit (Moschitti, 2006). Before applying these methods, constituent parse trees have been reduced to the shortest-enclosed parse following the recommendations from Zhang et al. (2006). For a more detailed description of the different methods we refer to the original publications.

In addition to the PPI framework, we also employed the general purpose relationship extraction tool "Turku Event Extraction System" (TEES) (Björne et al., 2011), a customized version of the case-based reasoning system Moara (Neves et al., 2009), and a self-developed feature based classifier which is referred to as SLW. Regarding TEES, we have used the edge extraction functionality for performing relationship extraction. TEES considers features related to the tokens (*e.g.*, part-of-speech tags), dependency chains, dependency path N-grams, entities (*e.g.*, entity types) and external resources, such as hypernyms in WordNet.

Moara is a case-based reasoning system for the extraction of relationships and events. During training, interaction pairs are converted into cases and saved into a HyperSQL database which are retrieved through case similarity during the classification. Cases are composed by the following features: the type of the entities (*e.g.* `Brand` and `Group`), the part-of-speech tag of the tokens between the two drugs (inclusive), the tags of the shortest dependency path between the two drugs, and the lemma of the non-entity tokens of the shortest dependency path using BioLemmatizer (Liu et al., 2012). We also consider the PHARE ontology (Coulet et al., 2011) in the lemma feature: When a lemma matches any of the synonyms contained in this ontology, the category of the respective term is considered instead. Case similarity is calculated by exact feature matching, except for the part-of-speech tags whose comparison is based on global alignment using insertion, deletion, and substitution costs as proposed by Spasic et al. (2005).

SLW is inspired by SL (Giuliano et al., 2006; Bunescu and Mooney, 2006) and uses the Breeze[2] library. We generate n-grams over sequences of arbitrary features (*e.g.* POS-tags, morphological and syntactical features) to describe the global context of an entity pair. Furthermore, we calculate features from the local context of entities, but in addition to SL, we include domain-specific features used for identifying and classifying pharmacological substances (see our paper for DDI Task 9.1 (Rocktäschel et al., 2013)). In addition, we take the name of the classes of a pair's two entities as feature to capture that entities of some class (*e.g.* `Brand` and `Group`) are more likely to interact than others (*e.g.* `Brand` and `Brand`).

2.3 Ensemble learning

Several community competitions previously noted that combinations of predictions from different tools help to achieve better results than one method alone (Kim et al., 2009; Leitner et al., 2010). More importantly, it is well known that ensembles increase robustness by decreasing the risk of selecting a bad classifier (Polikar, 2006). In this work we combined the output of several classifiers by using majority voting. The ensemble is used to predict DDIs regardless of the four different subtypes. This complies with the partial match evaluation criterion defined by the competition organizers.

2.4 Relabeling

To account for DDI subtypes, we compared two approaches: (a) using the subtype prediction of TEES; (b) training a multi-class classifier (SLW) on the available training data for DDI subtypes. We decided on using TEES, as it generated superior results over SLW (data not shown). Thus, previously identified DDIs are relabeled into one of the four possible subtypes using the most likely interaction subtype from TEES.

3 Results

3.1 Cross validation

In order to compare the different approaches, we performed document-wise 10-fold cross validation (CV) on the training set. It has been shown that such

[2] `http://www.scalanlp.org/`

Type	Pairs	Precision	Recall	F_1
total	3,119	78.6	78.6	78.6
effect	1,633	79.8	79.1	79.4
mechanism	1,319	79.8	79.2	79.4
advise	826	77.3	76.4	76.9
int	188	68.5	80.9	74.1

Table 4: Performance estimation for relabeling DDIs. Pairs denotes the number of instances of this type in the training corpus.

a setting provides more realistic performance estimates than instance-wise CV (Sætre et al., 2008). All approaches have been tested using the same splits to ensure comparability. For APG, ST, SST, and SpT we followed the parameter optimization strategy defined by Tikk et al. (2010). For TEES and Moara, we used the cost parameter C (50000) and best performing features, respectively, based on the CV results. For SL and SLW, we used the default parameters.

We performed several different CV experiments: First, we performed CV on the two corpora (DrugBank and MEDLINE) separately. Second, data from the other corpus has been additionally used during the training phase. This allows us to estimate the impact of additional, but potentially different text. CV results for DrugBank and MEDLINE are shown in Table 2 and 3 respectively.

3.2 Relabeling

Performance of relabeling is evaluated by performing 10-fold CV on the training set using the same splits as in previous analysis. Note that this experiment is solely performed on positive DDI instances to estimate separability of the four different DDI-subtypes. Results for relabeling are shown in Table 4.

3.3 Test dataset

For the test set we submitted results using the following three majority voting ensembles. For Run 1 we used Moara+SL+TEES, for Run 2 we used APG+Moara+SL+SLW+TEES and for Run 3 we used SL+SLW+TEES. Due to time constraints we did not use different ensembles for the two corpora. We rather decided to use ensembles which achieved generally good results for both training corpora. All classifiers, except APG, have been retrained on the combination of MEDLINE and DrugBank using the parameter setting yielding the highest F_1 in the training phase. For APG, we trained two different models: One model is trained on MEDLINE and DrugBank and one model is trained on DrugBank solely. The first model is applied on the MEDLINE test set and the latter on the DrugBank test set. Estimated results on the training corpus and official results on the test corpus are shown in Table 5.

4 Discussion

4.1 Training dataset

Document-wise CV results for the DrugBank corpus show no clear effect when using MEDLINE as additional training data. By using MEDLINE during the training phase we observe an average decrease of 0.3 percentage points (pp) in F_1 and an average increase of 0.7 pp in area under the receiver operating characteristic curve (AUC). The strongest impact can be observed for APG with a decrease of 2.3 pp in F_1. We therefore decided to train APG models for DrugBank without additional MEDLINE data. For almost all ensembles (with the exception of APG+SpT+SL) we observe superior results when using only DrugBank as training data. Interestingly, this effect can mostly be attributed to an average increase of 3.3 pp in recall, whereas precision remains fairly stable between ensembles using DrugBank solely and those with additional training data.

In contrast for MEDLINE, all methods largely benefit from additional training data with an average increase of 9.8 pp and 3.6 pp for F_1 and AUC respectively. For the ensemble based approaches, we observe an average increase of 13.8 pp for $F_1$when using DrugBank data in addition.

When ranking the different methods by F_1 and calculating correlation between the two different corpora, we observe only a weak correlation (Kendall's $\tau = 0.286$, $p< 1$). In other words, machine learning methods show varying performance-ranks between the two corpora. This difference is most pronounced for SL and SpT, with four ranks difference between DrugBank and MEDLINE. It is noteworthy that the two corpora are not directly

Method	Regular CV				Combined CV			
	P	R	F_1	AUC	P	R	F_1	AUC
SL	61.5	79.0	69.1	92.8	62.1	78.4	**69.2**	93.0
APG	77.2	62.6	**69.0**	91.5	75.9	59.8	66.7	91.6
TEES	77.2	62.0	**68.6**	87.3	75.5	60.9	67.3	86.9
SLW	73.7	60.0	65.9	91.3	73.4	61.2	**66.6**	91.3
Moara	72.1	55.2	**62.5**	—	72.0	54.7	62.1	—
SpT	51.4	73.4	60.3	87.3	52.7	71.4	**60.6**	87.7
SST	51.9	61.2	**56.0**	85.4	55.1	57.1	**56.0**	86.1
ST	47.3	64.2	54.2	82.3	48.3	64.3	**54.9**	82.7
SL+SLW+TEES	76.1	69.9	**72.7**	—	75.9	65.3	70.1	—
APG+SL+TEES	79.3	69.9	**74.2**	—	79.2	65.4	71.5	—
Moara+SL+TEES	79.9	69.6	**74.2**	—	79.6	65.1	71.6	—
Moara+SL+APG	81.4	70.6	**75.5**	—	81.3	70.3	75.3	—
APG+Moara+SL+SLW+TEES	84.0	68.1	**75.1**	—	83.7	64.2	72.6	—
APG+SpT+TEES	76.8	68.0	**72.1**	—	77.1	63.4	69.6	—
APG+SpT+SL	68.7	74.8	71.5	—	69.7	73.8	**71.6**	—

Table 2: Cross validation results on DrugBank corpus. Regular CV is training and evaluation on DrugBank only. Combined CV is training on DrugBank and MEDLINE and testing on DrugBank. Higher F_1 between these two settings are indicated in boldface for each method. Single methods are ranked by F_1.

Method	Regular CV				Combined CV			
	P	R	F_1	AUC	P	R	F_1	AUC
TEES	70.7	36.0	44.5	82.2	59.6	46.5	**51.4**	84.9
SpT	37.8	38.6	34.6	78.6	42.3	55.3	**47.1**	80.4
APG	46.5	44.3	42.4	82.3	38.1	62.2	**46.4**	82.8
SST	31.3	37.7	31.8	74.1	36.7	61.7	**44.9**	79.5
SL	43.7	40.1	38.7	78.9	34.7	67.1	**44.7**	81.1
SLW	58.0	14.3	20.4	73.4	50.1	38.0	**42.0**	82.4
Moara	49.8	31.9	37.6	—	45.6	43.2	**41.9**	—
ST	25.2	43.8	30.1	70.5	36.1	48.3	**39.8**	74.2
SL+SLW+TEES	73.6	29.0	37.6	—	55.2	52.7	**53.1**	—
APG+SL+TEES	60.7	37.9	43.4	—	49.9	62.4	**54.3**	—
Moara+SL+TEES	68.0	33.0	42.2	—	62.1	55.5	**57.4**	—
Moara+SL+APG	57.7	36.7	42.4	—	48.3	60.9	**52.8**	—
APG+Moara+SL+SLW+TEES	73.3	28.3	36.8	—	60.6	54.4	**56.5**	—
APG+SpT+TEES	58.5	37.4	41.7	—	57.5	59.2	**57.1**	—
APG+SpT+SL	48.3	39.9	40.0	—	43.6	64.3	**51.0**	—

Table 3: Cross validation results on MEDLINE corpus. Regular CV is training and evaluation on MEDLINE only. Combined CV is training on DrugBank and MEDLINE and testing on MEDLINE. Higher F_1 between these two settings are indicated in boldface for each method. Single methods are ranked by F_1.

Evaluation	Training									Test								
	Run 1			Run 2			Run 3			Run 1			Run 2			Run 3		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Partial	78.7	67.3	72.6	82.9	66.4	73.7	75.2	67.6	71.2	84.1	65.4	73.6	86.1	65.7	74.5	80.1	72.2	75.9
Strict	65.7	56.1	60.5	70.0	56.0	62.2	63.0	56.7	59.7	68.5	53.2	59.9	69.5	53.0	60.1	64.2	57.9	60.9
-mechanism	61.8	49.7	55.1	68.1	50.0	57.7	59.2	50.3	54.4	72.2	51.7	60.2	74.9	52.3	61.6	65.3	58.6	61.8
-effect	68.8	57.9	62.9	71.8	57.6	63.9	66.1	57.4	61.5	63.7	57.5	60.4	63.6	55.8	59.5	60.7	61.4	61.0
-advise	64.6	60.5	62.5	68.2	59.7	63.6	61.1	61.5	61.3	73.3	53.4	61.8	74.5	55.7	63.7	69.0	58.4	63.2
-int	68.6	50.0	57.8	75.4	52.1	61.6	70.9	56.9	63.1	67.8	41.7	51.6	67.3	38.5	49.0	67.8	41.7	51.6

Table 5: Relation extraction results on the training and test set. Run 1 builds a majority voting on Moara+SL+TEES, Run 2 on APG+Moara+SL+SLW+TEES, and Run 3 on SL+SLW+TEES. Partial characterizes only DDI detection without classification of subtypes, whereas strict requires correct identification of subtypes as well.

comparable, as DrugBank is one order of magnitude larger in terms of instances than the MEDLINE corpus. Additionally, documents come from different sources and it is tempting to speculate that there might be a certain amount of domain specificity between DrugBank and MEDLINE sentences.

We tested for domain specificity by performing cross-corpus experiments, *i.e.*, we trained a classifier on DrugBank, applied it on MEDLINE and *vice versa*. When training on MEDLINE and testing on DrugBank, we observe an average decrease of about 15 pp in F_1 in comparison to DrugBank in-domain CV results. For the other setting, we observe a lower decrease of approximately 5 pp in comparison to MEDLINE in-domain CV results.

From the current results, it seems that the documents from DrugBank and MEDLINE have different syntactic properties. However, this requires a more detailed analysis of different aspects like distribution of sentence length, negations, or passives between the two corpora (Cohen et al., 2010; Tikk et al., 2013). We assume that transfer learning techniques could improve results on both corpora (Pan and Yang, 2010).

The DDI-relabeling capability of TEES is very balanced with F_1 measures ranging from 74.1 % to 79.4 % for all four DDI subclasses. This is unexpected since classes like "effect" occur almost ten times more often than classes like "int" and classifiers often have problems with predicting minority classes.

4.2 Test dataset

On the test set, our best run achieves an F_1 of 76 % using the partial evaluation schema. This is slightly better than the performance for DrugBank training data shown in Table 2 and substantially better than estimations for MEDLINE (see Table 3). With F_1 measures ranging between 74 % to 76 % only minor performance differences can be observed between the three different ensembles.

When switching from partial to strict evaluation scheme an average decrease of 15 pp in F_1 can be observed. As estimated on the training data, relabeling performance is indeed very similar for the different DDI-subtypes. Only for the class with the least instances (*int*), a larger decrease in comparison to the other three classes can be observed for the test set. In general, results for test set are on par or higher than results for the training set.

5 Conclusion

In this paper we presented our approach for the SemEval 2013 – Task 9.2 DDI extraction challenge. Our strategy builds on a cascaded (coarse to fine grained) classification strategy, where a majority voting ensemble of different methods is initially used to find generic DDIs. Predicted interactions are subsequently relabeled into four different subtypes. DDI extraction seems to be a more difficult task for MEDLINE abstracts than for DrugBank articles. In our opinion, this cannot be fully attributed to the slightly higher ratio of positive instances in DrugBank and points towards structural differences between the two corpora.

Acknowledgments

This work was supported by the German Research Foundation (DFG) [LE 1428/3-1] and the Federal

Ministry of Economics and Technology (BMWi) [KF 2205209MS2].

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, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. 2011. Extracting Contextualized Complex Biological Events with Rich Graph-Based Features Sets. *Computational Intelligence*, 27(4):541–557.

R. C. Bunescu and R. J. Mooney. 2006. Subsequence Kernels for Relation Extraction. *Advances in Neural Information Processing Systems*, 18:171.

E. Charniak and M. Johnson. 2005. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proc. of ACL'05*, pages 173–180.

M. D. Cheitlin, A. M. Hutter, R. G. Brindis, P. Ganz, S. Kaul, R. O. Russell, and R. M. Zusman. 1999. Use of sildenafil (viagra) in patients with cardiovascular disease. *J Am Coll Cardiol*, 33(1):273–282.

K. Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11:492.

M. Collins and N. Duffy. 2001. Convolution Kernels for Natural Language. In *Proc. of NIPS'01*, pages 625–632.

A. Coulet, Y. Garten, M. Dumontier, R. Altman, M. Musen, and N. Shah. 2011. Integration and publication of heterogeneous text-mined relationships on the semantic web. *Journal of Biomedical Semantics*, 2(Suppl 2):S10.

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

L. French, S. Lane, L. Xu, C. Siu, C. Kwok, Y. Chen, C. Krebs, and P. Pavlidis. 2012. Application and evaluation of automated methods to extract neuroanatomical connectivity statements from free text. *Bioinformatics*, 28(22):2963–2970.

C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *Proc. of EACL'06*, pages 401–408.

S. I. Haider, K. Johnell, M. Thorslund, and J. Fastbom. 2007. Trends in polypharmacy and potential drug-drug interactions across educational groups in elderly patients in Sweden for the period 1992 - 2002. *Int J Clin Pharmacol Ther*, 45(12):643–653.

L. Hunter and K. Cohen. 2006. Biomedical language processing: what's beyond PubMed? *Mol Cell*, 21(5):589–594.

J.D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proc. of BioNLP'09*, pages 1–9.

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

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

F. Leitner, S.A. Mardis, M. Krallinger, G. Cesareni, L.A. Hirschman, and A. Valencia. 2010. An overview of BioCreative II. 5. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 385–399.

H. Liu, T. Christiansen, W. Baumgartner, and K. Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(1):3.

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

A. Moschitti. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In *Proc. of ECML'06*, pages 318–329.

M. Neves, J.-M. Carazo, and A. Pascual-Montano. 2009. Extraction of biomedical events using case-based reasoning. In *Proc. of BioNLP'09*, pages 68–76.

S. J. Pan and Q. Yang. 2010. A Survey on Transfer

Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

R. Polikar. 2006. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.

S. Pyysalo, R. Sætre, J. Tsujii, and T. Salakoski. 2008. Why Biomedical Relation Extraction Results are Incomparable and What to do about it. In *Proc. of SMBM'08*, pages 149–152.

T. Rocktäschel, T. Huber, M. Weidlich, and U. Leser. 2013. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

R. Sætre, K. Sagae, and J. Tsujii. 2008. Syntactic features for protein-protein interaction extraction. In *Proc. of LBM'07*.

I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

I. Segura-Bedmar, P. Martínez, and D. Sanchez-Cisneros. 2011. The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical text. In *Proc. of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011*, pages 1–9.

I. Solt, F. P. Szidarovszky, and D. Tikk. 2010. Concept, Assertion and Relation Extraction at the 2010 i2b2 Relation Extraction Challenge using parsing information and dictionaries. In *Proc. of i2b2/VA Shared-Task*.

I. Spasic, S. Ananiadou, and J. Tsujii. 2005. MaSTerClass: a case-based reasoning system for the classification of biomedical terms. *Bioinformatics*, 21(11):2748–2758.

P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser. 2011. Relation extraction for drug-drug interactions using ensemble learning. In *Proc. of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011*, pages 11–18.

D. Tikk, I. Solt, P. Thomas, and U. Leser. 2013. A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. *BMC Bioinformatics*, 14(1):12.

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.

S. V. N. Vishwanathan and A. J. Smola. 2002. Fast Kernels for String and Tree Matching. In *Proc. of NIPS'02*, pages 569–576.

M. Zhang, J. Zhang, J. Su, and G. Zhou. 2006. A Composite Kernel to Extract Relations between Entities with Both Flat and Structured Features. In *Proc. of ICML'06*, pages 825–832.