The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts

Isabel Segura-Bedmar, Paloma Martínez, and Daniel Sánchez-Cisneros

Universidad Carlos III de Madrid, Computer Science Department,
Avd. Universiad, 30, 28911 Leganés, Madrid, Spain
{isegura,pmf,dscisner}@springer.com
http://labda.inf.uc3m.es/

Abstract. We present an evaluation task designed to provide a frame-work for comparing different approaches to extracting drug-drug interac-tions from biomedical texts. We define the task, describe the training/test data, list the participating systems and discuss their results. There were 10 teams who submitted a total of 40 runs.

Keywords: Biomedical Text Mining, Drug-Drug Interaction Extraction

1 Task Description and Related Work

A drug-drug interaction (DDI) occurs when one drug influences the level or activity of another drug. Since negative DDIs can be very dangerous, DDI detection is the subject of an important field of research that is crucial for both patient safety and health care cost control. Although health care professionals are supported in DDI detection by different databases, those being used currently are rarely complete, since their update periods can be as long as three years [12]. Drug interactions are frequently reported in journals of clinical pharmacology and technical reports, making medical literature the most effective source for the detection of DDIs. The management of DDIs is a critical issue, therefore, due to the overwhelming amount of information available [8].

Information extraction (IE) can be of great benefit for both the pharmaceutical industry by facilitating the identification and extraction of relevant information on DDIs, as well as health care professionals by reducing the time spent reviewing the relevant literature. Moreover, the development of tools for automatically extracting DDIs is essential for improving and updating the drug knowledge databases.

Different systems have been developed for the extraction of biomedical relations, particularly PPIs, from texts. Nevertheless, few approaches have been proposed to the problem of extracting DDIs in biomedical texts. We developed two different approaches for DDI extraction. Since no benchmark corpus was available to evaluate our approaches to DDI extraction, we created the DrugDDI corpus annotated with 3,160 DDIs. Our first approach is a hybrid linguistic

approach [13] that combines shallow parsing and syntactic simplification with pattern matching. This system yielded a precision of 48.69%, a recall of 25.70% and an F-measure of 33.64%. Our second approach [14] is based on a supervised machine learning technique, more specifically, the shallow linguistic kernel proposed in Giuliano et al. (2006) [7]. It achieved a precision of 51.03%, a recall of 72.82% and an F-measure of 60.01%.

In order to stimulate research in this direction, we have organized the challenge task DDIExtraction2011. Likewise the BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge evaluation has devoted to provide a common frameworks for evaluation of text mining driving progress in text mining techniques applied to the biological domain, our purpose is to create a benchmark dataset and evaluation task that will enable researchers to compare their algorithms applied to the extraction of drug-drug interactions.

2 The DrugDDI corpus

While Natural Language Processing(NLP) techniques are relatively domain-portable, corpora are not. For this reason, we created the first annotated corpus, the DrugDDI corpus, studying the phenomenon of interactions among drugs. We hope that the corpus serves to encourage the NLP community to conduct further research in the field of pharmacology.

As source of unstructured textual information on drugs and their interactions, we used the DrugBank database[17]. This database is a rich resource combining chemical and pharmaceutical information of approximately 4,900 pharmacological substances. For each drug, DrugBank contains more than 100 data fields including drug synonyms, brand names, chemical formula and structure, drug categories, ATC and AHFS codes (i.e., codes of standard drug families), mechanism of action, indication, dosage forms, toxicity, etc. Of particular interest to this study, DrugBank offers the field 'Interactions' (it is no longer available) that contained a link to a document describing DDIs in unstructured texts. DrugBank provides a file with the names of approved drugs[1], approximately 1,450. We randomly chose 1,000 drug names and used the RobotMaker[2], a screen-scrapper application, to download the interaction documents for these drugs. We only retrieved a total of 930 documents since some drugs did not have any linked document. Due to the cost-intensive and time consuming nature of the annotation process, we decided to reduce the number of documents to be annotated and only considered 579 documents. We believe that these texts are a reliable and representative source of data for expressing DDI since the language used is mostly devoted to descriptions of DDIs. Additionally, the highly specialized pharmacological language is very similar to that found in the Medline pharmacology abstracts.

These documents were then analyzed by the UMLS MetaMap Transfer (MMTx) [2] tool performing sentence splitting, tokenization, POS-tagging, shal-

[1] http://www.drugbank.ca/downloads
[2] http://openkapow.com/

low syntactic parsing (see Figure 1) and linking of phrases with UMLS Metathesaurus concepts. Drugs are automatically identified by MMTx since the tool allows for the recognition and annotation of biomedical entities occurring in texts according to the UMLS semantic types. An experienced pharmacist reviewed the UMLS Semantic Network as well as the semantic annotation provided by MMTx and recommended us the inclusion of the following UMLS semantic types as possible types of interacting drugs: Clinical Drug (clnd), Pharmacological Substance (phsu), Antibiotic (antb), Biologically Active Substance (bacs), Chemical Viewed Structurally (chvs) and Amino Acid, Peptide, or Protein (aapp).

The principal value of the DrugDDI corpus undoubtedly comes from its DDIs annotations. To obtain these annotations, all documents were marked-up by a researcher with pharmaceutical background. DDIs were annotated at the sentence level and, thus, any interactions spanning over several sentences were not annotated here. Only sentences with two or more drugs were considered and the annotation was made sentence by sentence. Figure 1 shows an example of an annotated sentence that contains three interactions. Each interaction is represented as a *DDI* node in which the names of the interacting drugs are registered in its *NAME_DRUG_1* and *NAME_DRUG_2* attributes. The identifiers of the phrases containing these interacting drugs are also annotated, providing an easily access to the related concepts provided by MMTx. As mentioned, Figure 1 shows three DDIs: the first DDI represents an interaction between *Aspirin* and *probenecid*, the second one an interaction between *aspirin* and *sulfinpyrazone*, and the last one a DDI between *aspirin* and *phenylbutazone*.

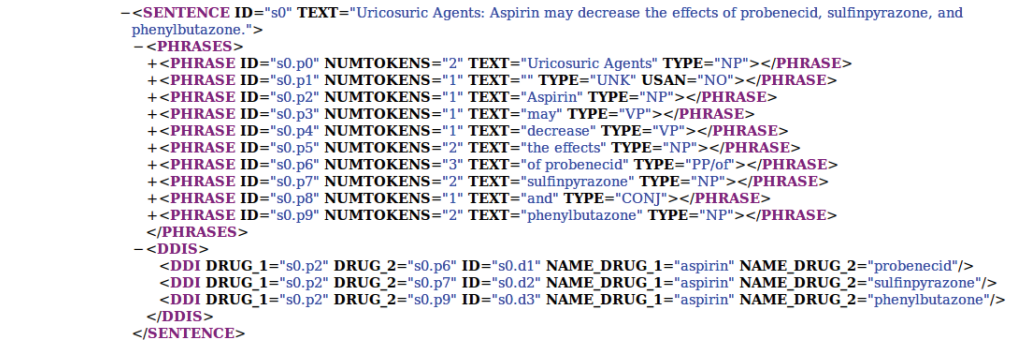

Fig. 1. Example of DDI annotations.

The DrugDDI corpus is also provided in the unified format for PPI corpora proposed in Pyysalo et al. [11] (see Figure 2). This shared format could attract attention of groups studying PPI extraction because they could easily adapt their systems to the problem of DDI extraction. The unified XML format does not contain any linguistic information provided by MMTx. The unified format only

```
−<sentence id="DrugDDI.d346.s0" origId="s0" text="Uricosuric Agents: Aspirin may decrease the effects of probenecid,
 sulfinpyrazone, and phenylbutazone.">
   <entity id="DrugDDI.d346.s0.e0" origId="s0.p0" charOffset="0-17" type="drug" text="Uricosuric Agents"/>
   <entity id="DrugDDI.d346.s0.e1" origId="s0.p2" charOffset="19-26" type="drug" text="Aspirin"/>
   <entity id="DrugDDI.d346.s0.e2" origId="s0.p6" charOffset="55-65" type="drug" text="probenecid"/>
   <entity id="DrugDDI.d346.s0.e3" origId="s0.p7" charOffset="67-81" type="drug" text="sulfinpyrazone"/>
   <entity id="DrugDDI.d346.s0.e4" origId="s0.p9" charOffset="87-101" type="drug" text="phenylbutazone"/>
   <pair id="DrugDDI.d346.s0.p0" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e1" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p1" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e2" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p2" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e3" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p3" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e4" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p4" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e2" interaction="true"/>
   <pair id="DrugDDI.d346.s0.p5" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e3" interaction="true"/>
   <pair id="DrugDDI.d346.s0.p6" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e4" interaction="true"/>
   <pair id="DrugDDI.d346.s0.p7" e1="DrugDDI.d346.s0.e2" e2="DrugDDI.d346.s0.e3" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p8" e1="DrugDDI.d346.s0.e2" e2="DrugDDI.d346.s0.e4" interaction="false"/>
   <pair id="DrugDDI.d346.s0.p9" e1="DrugDDI.d346.s0.e3" e2="DrugDDI.d346.s0.e4" interaction="false"/>
</sentence>
```

Fig. 2. The unified XML format.

Table 1. Basic statistics on the DrugDDI corpus.

	Number	Avg. per document
Documents	579	
Sentences	5,806	10.03
Phrases	66,021	114.02
Tokens	127,653	220.47
Sentences with at least one DDI	2,044	3.53
Sentences with no DDI	3,762	6.50
DDIs	3,160	5.46 (0.54 per sentence)

provides the sentences, their drugs and their interactions. Each entity (drug) includes reference (origId) to its id phrase in the MMTX format corpus text in which the corresponding drug appears. For each sentence from the DrugDDI corpus represented in the unified XML format, its DDI candidate pairs should be generated from the different drugs appearing therein. Each DDI candidate pair is represented as a *pair* node in which the ids of the interacting drugs are registered in its *e1* and *e2* attributes. If the pair is a DDI, the *interaction* attribute must be set to *true*, and *false* value otherwise.

Table 1 shows basic statistics of the DrugDDI corpus. In general, the size of biomedical corpora is quite small and usually does not exceed 1,000 sentences. The average number of sentences per MedLine abstract was estimated at 7.2 ± 1.9 [18]. Our corpus contains 5,806 sentences with 10.3 sentences per document on average. MMTx identified a total of 66,021 phrases of which 12.5% (8,260) are drugs. The average number of drug mentions per document was 24.9, and the average number of drug mentions per sentence was 2.4. The corpus contains a total of 3,775 sentences with two or more drug mentions, although only 2,044 sentences contain at least one interaction. With the assistance of a pharmacist, a total of 3,160 DDIs were with an average of 5.46 DDIs per document and 0.54 per sentence.

DDI extraction can be formulated as a supervised learning problem, more particularly, as a drug pair classification task. Therefore, a crucial step is to

generate suitable datasets to train and test a classifier from the DrugDDI corpus. The simplest way to generate examples to train a classifier for a specific relation R is to enumerate all possible ordered pairs of sentence entities. We proceeded in a similar way. Given a sentence S with at least two drugs, we defined D as the set of drugs in S and N as the number of drugs. The set of examples generated for S, therefore, was defined as follows: $\{(D_i, D_j) : D_i, D_j \epsilon D, 1 <= i, j <= N, i \neq j, i < j\}$. If the interaction existed between the two DDI candidate drugs, then the example was labeled 1. Otherwise, it was labeled 0. Although some DDIs may be asymmetrical, the roles of the interacting drugs were not included in the corpus annotation and are not specifically addressed in this task. As a result, we enumerate candidate pairs here without taking their order into account, such that (D_i, D_j) and (D_j, D_i) are considered as a single candidate pair. Since the order of the drugs in the sentence was not taken into account, each example is the copy of the original sentence S where the candidates were assigned the tag, 'DRUG', and remaining drugs were assigned the tag, 'OTHER'. The set of possible candidate pairs was the set of $2-$combinations from the whole set of drugs appearing in S. Thus, the number of examples was $C_{N,2} = \binom{N}{2}$.

Table 2 shows the total number of relation examples or instances generated from the DrugDDI corpus. Among the 30,757 candidate drug pairs, only 3,160 (10.27%) were marked as positive interactions (i.e., DDIs) while 27,597 (89.73%) were marked as negative interactions (i.e., non-DDIs).

Table 2. Distribution of positive and negative examples in training and testing datasets.

Set	Documents	Examples	Positives	Negatives
Train	437 (75.5%)	25,209	2,421 (9.6%)	22,788 (90.4%)
Final Test	142 (24.5%)	5,548	739 (13.3%)	4,809 (86.7%)
Total	579	30,757	3,160 (10.27%)	27,597 (89.73%)

Once we generated the set of relation instances from the DrugDDI corpus, the set was then split in order to build the datasets for the training and evaluation of the different DDI extraction systems. In order to build the training dataset used for development tests, 75% of the DrugDDI corpus files (435 files) were randomly selected for the training dataset and the remaining 25% (144 files) is used in the final evaluation to determine which model was superior. Table 3 shows the distribution of the documents, sentences, drugs and DDIs in each set. Approximately 90% of the instances in the training dataset were negative examples (i.e., non-DDIs). The distribution between positive and negative examples in the final test dataset was also quite similar (see Table 2).

3 The participants

The task of extracting drug-drug interactions from biomedical texts has attracted the participation of 10 teams who submitted 40 runs. Table 4 lists the teams,

Table 3. Training and testing datasets.

Set	Documents	Sentences	Drugs	DDIs
Training	435	4,267	11,260	2,402
Final Test	144	1,539	3,689	758
Total	579	5,806	14,949	3,160

their affiliations, the number of runs submitted and the description of their systems.

The runs' performance information in terms of precision, recall, F-measure and accuracy, appears in Table 5.

Table 4. Short description of the teams.

Team	Institution	Runs	Description
WBI	Humboldt-Universitat Berlin	5	combination of several kernels and a case-based reasoning (CBR) system using a voting approach
FBK-HLT	Fondazione Bruno Kessler - HLT	5	composite kernels using the MEDT, PST and SL kernels
LIMSI-FBK	LIMSI - Fondazione Bruno Kessler	1	a feature-based method using SVM and a composite kernel-based method.
UTurku	University of Turku	4	machine learning classifiers such as SVM and RLS; DrugBank and MetaMap
LIMSI-CNRS	LIMSI-CNRS	5	a feature-based method using libSVM and SVMPerf
bnb_nlel	Universidad Politécnica de Valencia	1	a feature-based method using Random Forests
laberinto-uhu	Universidad de Huelva	5	a feature-based method using classical classifiers such as SVM, Nave Bayes, Decision Trees, Adaboost
DrIF	University of Pavia (Department Mario Stefanelli)	4	two machine learning-based (CFFs and SVMs) and one hybrid approach which combines CRFs and a rule-based technique.
ENCU	East China Normal University	5	a feature-based method using SVM.
IUPUITMGroup	Indiana University-Purdue University Indianapolis	5	all paths graph (APG) kernel

Table 5. Precision, recall, F-measure and accuracy over each run's performance.

Team	run	TP	FP	FN	TN	P	R	F	Acc
WBI	5	543	354	212	5917	0.6054	0.7192	0.6574	0.9194
WBI	4	529	332	226	5939	0.6144	0.7007	0.6547	0.9206
WBI	2	568	465	187	5806	0.5499	0.7523	0.6353	0.9072
WBI	1	575	585	180	5686	0.4957	0.7616	0.6005	0.8911
WBI	3	319	362	436	5909	0.4684	0.4225	0.4443	0.8864
LIMSI-FBK	1	532	376	223	5895	0.5859	0.7046	0.6398	0.9147
FBK-HLT	4	529	377	226	5894	0.5839	0.7007	0.6370	0.9142
FBK-HLT	1	513	344	242	5927	0.5986	0.6795	0.6365	0.9166
FBK-HLT	2	560	458	195	5813	0.5501	0.7417	0.6317	0.9071
FBK-HLT	3	534	423	221	5848	0.5580	0.7073	0.6238	0.9083
FBK-HLT	5	544	674	211	5597	0.4466	0.7205	0.5514	0.8740
Uturku	3	520	376	235	5895	0.5804	0.6887	0.6299	0.9130
Uturku	4	370	179	385	6092	0.6740	0.4901	0.5675	0.9197
Uturku	2	368	197	387	6074	0.6513	0.4874	0.5576	0.9169
Uturku	1	350	172	405	6099	0.6705	0.4636	0.5482	0.9179
LIMSI-CNRS	1	490	398	265	5873	0.5518	0.6490	0.5965	0.9056
LIMSI-CNRS	2	491	402	264	5869	0.5498	0.6503	0.5959	0.9052
LIMSI-CNRS	4	462	380	293	5891	0.5487	0.6119	0.5786	0.9042
LIMSI-CNRS	5	373	264	382	6007	0.5856	0.4940	0.5359	0.9081
LIMSI-CNRS	3	388	470	367	5801	0.4522	0.5139	0.4811	0.8809
BNBNLEL	1	420	266	335	6005	0.6122	0.5563	0.5829	0.9145
laberinto-uhu	1	335	335	420	5936	0.5000	0.4437	0.4702	0.8925
laberinto-uhu	2	324	371	431	5900	0.4662	0.4291	0.4469	0.8859
laberinto-uhu	3	368	551	387	5720	0.4004	0.4874	0.4397	0.8665
laberinto-uhu	4	238	153	517	6118	0.6087	0.3152	0.4154	0.9046
laberinto-uhu	5	193	107	562	6164	0.6433	0.2556	0.3659	0.9048
DrIF	1	369	545	386	5725	0.4037	0.4887	0.4422	0.8675
DrIF	4	369	545	386	5726	0.4037	0.4887	0.4422	0.8675
DrIF	3	317	456	438	5815	0.4101	0.4199	0.4149	0.8728
DrIF	2	196	110	559	6161	0.6405	0.2596	0.3695	0.9048
ENCU	5	351	836	404	5435	0.2957	0.4649	0.3615	0.8235
ENCU	3	324	830	431	5441	0.2808	0.4291	0.3394	0.8205
ENCU	1	580	3456	175	2815	0.1437	0.7682	0.2421	0.4832
ENCU	2	713	4781	42	1490	0.1298	0.9444	0.2282	0.3135
ENCU	4	206	424	549	5847	0.3270	0.2728	0.2975	0.8615
IUPUITMGroup	4	193	1457	562	4814	0.1170	0.2556	0.1605	0.7126
IUPUITMGroup	1	237	2005	518	4266	0.1057	0.3139	0.1582	0.6409
IUPUITMGroup	2	127	943	628	5328	0.1187	0.1682	0.1392	0.7764
IUPUITMGroup	3	125	937	630	5334	0.1177	0.1656	0.1376	0.7770
IUPUITMGroup	5	110	770	645	5501	0.1250	0.1457	0.1346	0.7986

4 Discussion

The best performance is achieved by the team WBI [15]. Its system combines several kernels (APG [1], SL [7], kBSPS [16]) and a case-based reasoning (CBR) (called MOARA [10]) using a voting approach. In particular, the combination

of the kernels APG, SL and the MOARA system yields the best F-measure (0.6574).

The team FBK-HLT [5] proposes new composite kernels using well-known kernels such as MEDT [6], PST [9] and SL [7]. Similarly, the team LIMSI-FBK [4] combines the same kernels (MEDT, PST and SL) and a feature-based method using SVM. This system achieves an F-measure of 0.6398.

The team Uturku [3] proposes a feature-based method using the classifiers SVM and RLS. Features used by the classifiers include syntactic information (tokens, dependency types, POS tags, text, stems, etc) and semantic knowledge from DrugBank and MetaMap. This system achieves an F-measure of 0.6299.

In general, approaches based on kernels methods achieved better results than the classical feature-based methods. Most systems have used primarily syntactic information, however semantic information has been poorly used.

5 Conclusion

This paper describes a new semantic evaluation task, Extraction of drug-drug interactions from biomedical texts. We have accomplished our goal of providing a framework and a benchmark data set to allow for comparisons of methods for this task. The results that the participating systems have reported show successful approaches to this difficult task, and the advantages of kernel-based methods over classical machine learning classifiers.

The success of the task shows that the framework and the data are useful resources. By making this collection freely accessible, we encourage further research into this domain. Moreover, next SemEval-3 (6th International Workshop on Semantic Evaluations[3]) to be held in summer 2013 has scheduled the "Extraction of drug-drug interactions from biomedical Texts" task [4]. In order to accomplish this new task, the current corpus is being extended to collect new data test.

Acknowledgements

This study was funded by the projects MA2VICMR (S2009/TIC-1542) and MULTIMEDICA (TIN2010-20644-C03-01). The organizers are particularly grateful to all participants who contributed to detect annotation errors in the corpus.

References

1. Airola, A., Pyysalo, S., Bjorne, J., Pahikkala, T., Ginter, F., Salakoski, T.: All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics 9(Suppl 11), S2 (2008)

[3] http://www.cs.york.ac.uk/semeval/
[4] http://www.cs.york.ac.uk/semeval/proposal-16.html

2. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Annual AMIA Symposium pp. 17–21 (Jan 2001)
3. Björne, J., Airola, A., Pahikkala, T., Salakoski, T.: Drug-drug interaction extraction with rls and svm classiffers. In: Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011) (2011)
4. Chowdhury, M., Abacha, A., Lavelli, A., P., Z.: Two different machine learning techniques for drug-drug interaction extraction. In: Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011) (2011)
5. Chowdhury, M., Lavelli, A.: Drug-drug interaction extraction using composite kernels. In: Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011) (2011)
6. Chowdhury, M., Lavelli, A., Moschitti, A.: A study on dependency tree kernels for automatic extraction of protein-protein interaction. ACL HLT 2011 p. 124
7. Giuliano, C., Lavelli, A., Romano, L.: Exploiting shallow linguistic information for relation extraction from biomedical literature. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006). pp. 401–408 (2006)
8. Hansten, P.D.: Drug interaction management. Pharmacy World & Science 25(3), 94–97 (2003)
9. Moschitti, A.: A study on convolution kernels for shallow semantic parsing. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. pp. 335–es. Association for Computational Linguistics (2004)
10. Neves, M., Carazo, J., Pascual-Montano, A.: Extraction of biomedical events using case-based reasoning. In: Proceedings of the Workshop on BioNLP: Shared Task. pp. 68–76. Association for Computational Linguistics (2009)
11. Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F., Salakoski, T.: Comparative analysis of five protein-protein interaction corpora. BMC bioinformatics 9(Suppl 3), S6 (2008)
12. Rodríguez-Terol, A., Camacho, C., Others: Calidad estructural de las bases de datos de interacciones. Farmacia Hospitalaria 33(03), 134 (2009)
13. Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C.: A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics 12(Suppl 2), S1 (2011)
14. Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C.: Using a shallow linguistic kernel for drug-drug interaction extraction. Journal of Biomedical Informatics In Press, Corrected Proof (2011)
15. Thomas, P., Neves, M., Solt, I., Tikk, D., Leser, U.: Relation extraction for drug-drug interactions using ensemble learning. In: Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011) (2011)
16. Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., Leser, U.: A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. PLoS Computational Biology 6(7), e1000837 (2010)
17. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36(Database issue), D901–6 (Jan 2008)
18. Yu, H.: Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. Annual AMIA Symposium proceedings pp. 834–8 (Jan 2006)