Relation Extraction for Drug-Drug Interactions using Ensemble Learning

Philippe Thomas[1], Mariana Neves[1], Illés Solt[2], Domonkos Tikk[2], and Ulf Leser[1]

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

[2] Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Magyar tudósok körútja 2, 1117 Budapest, Hungary {solt, tikk}@tmit.bme.hu

Abstract. We describe our approach for the extraction of drug-drug interactions from literature. The proposed method builds majority voting ensembles of contrasting machine learning methods, which exploit different linguistic feature spaces. We evaluated our approach in the context of the DDI Extraction 2011 challenge, where using document-wise cross-validation, the best single classifier achieved an F_1 of 57.3 % and the best ensemble achieved 60.6 %. On the held out test set, our best run achieved an F_1 of 65.7 %.

Keywords: Text mining; Relation extraction; Machine learning; Ensemble learning

1 Introduction

Most biomedical knowledge appears first as research results in scientific publications before it is distilled into structured knowledge bases. For researchers and database curators there is an urgent need to cope with the fast increase of biomedical literature [6]. Biomedical text mining currently achieves good results for named entity recognition (NER), *e.g.* gene/protein-names and recognition of single nucleotide polymorphisms [3, 11]. A recent trend is the extraction of simple or complex relations between entities [7].

In this work, we describe our approach for the extraction of drug-drug interactions (DDI) from text that was also the core task of the DDI Extraction 2011 challenge[1]. DDIs describe the interference of one drug with another drug and usually lead to an enhanced, reduced, neutralized, or even toxic drug effect. For example: "Aspirin administered in combination with Warfarin can lead to bleeding and has to be avoided." DDI effects are thus crucial to decide when (not) to administer specific drugs to patients.

[1] http://labda.inf.uc3m.es/DDIExtraction2011/

Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), pages 11–18, Huelva, Spain, September 2011.

1.1 Problem definition

The DDI challenge[1] consisted of one task, namely the identification of interactions between two drugs. This interaction is binary and undirected, as target and agent roles are not labeled. In the challenge setting, recognition of drug names was readily available.

2 Methods

Binary relation extraction is often tackled as a pair-wise classification problem between all entities mentioned within one sentence. Thus a sentence with n entities contains at most $\binom{n}{2}$ interacting pairs.

Corpus annotations have been made available in two different formats. (1) contained only the documents with respective drug annotations in a format previously used for protein-protein interactions (PPIs) [13]. (2) additionally contained linguistic information such as part-of-speech tags and shallow parses. Further phrases were annotated with corresponding UMLS concepts. This information has been automatically derived using MetaMap and incorporated by the organizers. We exclusively used (1) and extended it with linguistic information as described in the following subsection.

2.1 Preprocessing

Sentences have been parsed using Charniak–Lease parser [8] with a self-trained re-ranking model augmented for biomedical texts [10]. Resulting constituent parse trees have been converted into dependency graphs using the Stanford converter [4]. In the last step we created an augmented XML following the recommendations of [2]. This XML encompasses tokens with respective part-of-speech tags, constituent parse tree, and dependency parse tree information. Properties of the training and test corpora are shown in Table 1. Please note that the number of positive and negative instances in the test set has been made available after the end of the challenge. A more detailed description of the DDI corpus can be found in [14].

Corpus	Sentences	Pairs		
		Positive	Negative	Total
Training	4,267	2,402	21,425	23,827
Test	1,539	755	6,271	7,026

Table 1: Basic statistics of the DDI corpus training and test sets.

2.2 Kernel based approaches

Tikk *et al.* [17] systematically analyzed 9 different machine learning approaches for the extraction of undirected binary protein-protein interactions. In their analysis, three kernel have been identified of being superior to the remaining six approaches, namely all-paths graph (APG) [2], k-band shortest path spectrum (kBSPS) [17], and the shallow linguistic (SL) [5] kernel. The SL kernel uses only shallow linguistic features, *i.e.* word, stem, part-of-speech tag and morphologic properties of the surrounding words. kBSPS builds a classifier on the shortest dependency path connecting the two entities. It further allows for variable mismatches and also incorporates all nodes within distance k from the shortest path. APG builds a classifier using surface features and a weighting scheme for dependency parse tree features. For a more detailed description of the kernel we refer to the original publications. The advantage of these three methods has been replicated and validated in a follow up experiment during the i2b2 relation extraction challenge [15]. In the current work we also focus on these three methods.

Experiments have been done using an open-source relation extraction framework.[2] Entities were blinded by replacing the entity name with a generic string to ensure the generality of the approach. Without entity blinding a classifier uses drug names as features, which clearly affects its generalization abilities on unseen entity pairs.

2.3 Case-based reasoning

In addition to kernel classifiers, we also used a customized version of Moara, an improvement of the system that participated in the BioNLP'09 Event Extraction Challenge [12]. It uses case-based reasoning (CBR) for classifying the drug pairs. CBR [1] is a machine learning approach that represents data with a set of features. In the training step, first the cases from the training data are learned and then saved in a knowledge base. During the testing step, the same representation of cases is used for the input data, the documents are converted to cases and the system searches the base for cases most similar to the case-problem.

Each drug pair corresponds to one case. This case is represented by the local context, *i.e.*, the tokens between a drug pair. We have limited the size of the context to 20 tokens (pairs separated by more tokens are treated as false). The features may be related to the context as a whole or to each of the tokens that is part of the context. Features may be set as mandatory or optional, here no feature was defined as mandatory. As features we considered part-of-speech tag, role and lemma.

The part-of-speech tag is the one obtained during the pre-processing of the corpus. The role of the token is set to *DRUG* in case that the token is annotated as drug that takes part in the interaction. No role is set to drugs which are part of the context and are not part of the interaction pair, as well as the remaining

[2] `http://informatik.hu-berlin.de/forschung/gebiete/wbi/ppi-benchmark`

tokens. The lemma feature is only assigned for the non-role tokens using the Dragon toolkit [18], otherwise the feature is not set. See Table 2 for an example.

Context	Features		
	Lemma	POS	Role
Buprenorphine	*drug*	NN	DRUG
is	be	VBZ	–
metabolized	metabolized	VBN	–
to	to	TO	–
norbuprenorphine	norbuprenorphine	NN	–
by	by	IN	–
cytochrome	*drug*	NN	DRUG

Table 2: Example of features for the two interacting drugs described in the sentence "Buprenorphine is metabolized to norbuprenorphine by cytochrome." The lemma *drug* is the result of entity blinding.

During the searching step, Moara uses a filtering strategy in which it looks for a case with exactly the same values for the features, *i.e.*, it tries to find cases with exactly the same values for the mandatory features and matching as many optional features as possible. For the case retrieved in this step, a similarity between those and the original case is calculated by comparing the values of the corresponding features using a global alignment. This methodology was proposed as part of the CBR algorithm for biomedical term classification in the MaSTerClass system [16]. By default, for any feature, the insertion and deletion costs are 1 (one) and the substitution cost is 0 (zero) for equal features with equal values, and 1 (one) otherwise. However, we have also defined specific costs for the part-of-speech tag feature which were based on the ones used in the MaSTerClass system. We decided to select those cases whose global alignment score is below a certain threshold, automatically defined as proposed in [16]. The final solution, *i.e.*, whether the predicted category is "positive" or "negative", is given by a voting scheme among the similar cases. When no similar case if found for a determined pair, or if the pair was not analyzed at all due to its length (larger than 20), the "negative" category is assigned by default.

2.4 Ensemble learning

Previous extraction challenges showed that combinations of classifiers may achieve better results than any single classifier itself [7, 9]. Thus we experimented with different combinations of classifiers by using a majority voting scheme.

3 Results

3.1 Cross validation

In order to compare the different approaches, we performed document-wise 10-fold cross validation on the training set (see Table 3). It has been shown that such a setting provides more realistic performance estimates than instance-wise cross validation [2]. All approaches have been tested using the same splits to ensure comparability. For APG, kBSPS, and SL; we followed the parameter optimization strategy described in [17].

Method		Performance		
Type	Name	P	R	F_1
Kernel	APG	53.4	63.1	57.3
	kBSPS	35.9	53.5	42.7
	SL	45.4	**71.6**	55.3
Case-based reasoning	Moara	43.3	40.7	41.6
Ensemble	APG/Moara/SL	**59.0**	63.0	**60.6**
	APG/kBSPS/SL	53.2	65.2	58.3

Table 3: Document-wise cross-validation results on the training set for selected methods.

3.2 Test dataset

For the test set we submitted results for APG, SL, Moara, and the two majority voting ensembles. Results for kBSPS have been excluded, as only 5 submissions were permitted and kBSPS and Moara achieve similar results in F_1. The official results achieved on the test set are shown Table 4.

Run	Method	P	R	F_1
WBI-2	APG	55.0	75.2	63.4
WBI-1	SL	49.6	**76.2**	60.1
WBI-3	Moara	46.8	42.3	44.4
WBI-5	APG/Moara/SL	60.5	71.9	**65.7**
WBI-4	APG/kBSPS/SL	**61.4**	70.1	65.5

Table 4: Relation extraction results on the test set.

4 Discussion

4.1 Cross-validation

The document-wise cross-validation results show that SL and APG outperform the remaining methods. kBSPS and Moara are on a par with each other but F_1 is about 15 percentage points (pp) inferior to SL or APG. Even though the results of kBSPS and Moara are inferior, as ensemble members they are capable of improving F_1 on the training corpus. The combination APG/Moara/SL performs about 2.3 pp better in F_1 than the APG/kBSPS/SL ensemble and yields an overall improvement of 3.3 pp in comparison to the best single classifier (APG). Single method results are in line with previously published results using these kernel for other domains [15, 17]. Again the SL kernel, which uses only shallow linguistic information, achieves considerably good results. This indicates that shallow information is often sufficient for relation extraction.

We estimated the effect of entity blinding by temporarily disabling it. This experiment has been performed for SL exclusively and yielded an increase of 1.7 pp in F_1. This effect was accompanied by an increase of 3.6 pp in precision and a decrease of 3 pp in recall. We did not disable entity blinding for the submitted runs, as such classifiers would be biased towards known DDIs and less capable of finding novel DDIs, the ultimate goal of DDI extraction.

4.2 Test dataset

For the challenge all four classifier have been retrained using the whole training corpus using the parameter setting yielding the highest F_1in the training phase. Our best run achieved 65.7 % in F_1.

Between training and test results we observe a perfect correlation for F_1 (Kendall's tau (τ) of 1.0). Thus the evaluation corpus affirms the general ranking of methods determined on the training corpus. The effect of ensemble learning is less pronounced on the test set but with 2.3 pp still notable.

4.3 Error analysis

To have an impression about the errors generated by these classifiers, we manually analyzed drug mention pairs that were not correctly classified by any method (APG, kBSPS, Moara, and SL). Performing cross-validation, the DDI training corpus contained 442 (1.85 %) such pairs, examples are given in Figure 1.

We identified a few situations that may have caused difficulties: issues with the annotated corpus and linguistic constructs not or incorrectly handled by our methods. Annotation inconsistencies we encountered include dubious drug entity annotations (B1, B6 in Figure 1), and ground truth annotations that were either likely incorrect (B3) or could not be verified without the context (A4, B4). As for linguistic constructs, our methods lack co-reference resolution (A1, B5) and negation detection (A6, B7), and they also fail to recognize complex formulations (A5, B2). As a special case, conditional constructs belong to both

A1 **Probenecid** interferes with renal tubular secretion of *ciprofloxacin* and produces an increase in the level of **ciprofloxacin** in serum.

A2 *Drugs* which may enhance the neuromuscular blocking action of **TRACRIUM** include: **enflurane**;

A3 While not systematically studied, certain *drugs* may induce the metabolism of **bupropion** (e.g., **carbamazepine**, *phenobarbital*, *phenytoin*).

A4 **Auranofin** should not be used together with *penicillamine* (**Depen**, *Cuprimine*), another *arthritis medication*.

A5 These **drugs** in combination with very high doses of **quinolones** have been shown to provoke convulsions

A6 **Diclofenac** interferes minimally or not at all with the protein binding of *salicylic acid* (20% decrease in binding), *tolbutamide*, **prednisolone** (10% decrease in binding), or *warfarin*.

(a) False negatives

B1 **Dofetilide** is eliminated in the kidney by **cation**ic secretion.

B2 Use of **sulfapyridine** with these **medicines** may increase the chance of side effects of these *medicines*.

B3 **Haloperidol** blocks dopamine receptors, thus inhibiting the central stimulant effects of **amphetamines**.

B4 This interaction should be given consideration in patients taking **NSAIDs** concomitantly with **ACE inhibitors**.

B5 No dose adjustment of **bosentan** is necessary, but increased effects of **bosentan** should be considered.

B6 **Epirubicin** is extensively metabolized by the **liver**.

B7 **Gabapentin** is not appreciably metabolized nor does it interfere with the metabolism of commonly coadministered **antiepileptic drugs**.

(b) False positives

Fig. 1: Examples of drug mention pairs not classified correctly by any of our methods. The two entities of the pair are typeset in bold, others in italic.

groups, they are nor consistently annotated nor consistently classified by our methods (A2, A3, B2). Furthermore, we found several examples that are not affected by any of the above situations.

5 Conclusion

In this paper we presented our approach for the DDI Extraction 2011 challenge. Primarily, we investigated the re-usability of methods previously proven efficient for relation extraction in other biomedical sub-domains, notably protein-protein interaction (PPI) extraction. In comparison to PPI extraction corpora, the training corpus is substantially larger and also exhibits a higher class imbalance towards negative instances. Furthermore, we experimented with basic ensembles to increase overall performance and conducted a manual error analysis to pinpoint weaknesses in the applied methods. Our best result consisted of a majority voting ensemble of three methodically different classifiers.

Acknowledgments

PT was supported by German Federal Ministry of Education and Research (grant No 0315417B), MN by German Research Foundation, and DT by Alexander von Humboldt Foundation.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 7(1), 39–59 (1994)
2. Airola, A., Pyysalo, S., Björne, 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. Caporaso, J.G., Baumgartner, W.A., Randolph, D.A., Cohen, K.B., Hunter, L.: MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics 23(14), 1862–1865 (Jul 2007)
4. De Marneffe, M., MacCartney, B., Manning, C.: Generating typed dependency parses from phrase structure parses. In: LREC 2006. vol. 6, pp. 449–454 (2006)
5. Giuliano, C., Lavelli, A., Romano, L.: Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In: Proc. of EACL'06. Trento, Italy (2006)
6. Hunter, L., Cohen, K.B.: Biomedical language processing: what's beyond PubMed? Mol Cell 21(5), 589–594 (Mar 2006)
7. Kim, J., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 shared task on event extraction. In: Proc. of BioNLP'09. pp. 1–9 (2009)
8. Lease, M., Charniak, E.: Parsing biomedical literature. In: Proc. of IJCNLP'05. pp. 58–69 (2005)
9. Leitner, F., Mardis, S., Krallinger, M., Cesareni, G., Hirschman, L., Valencia, A.: An overview of BioCreative II. 5. IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics pp. 385–399 (2010)
10. McClosky, D.: Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing. Ph.D. thesis, Brown University (2010)
11. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H., Torres, R., Krauthammer, M., Lau, W.W., Liu, H., Hsu, C.N., Schuemie, M., Cohen, K.B., Hirschman, L.: Overview of BioCreative II gene normalization. Genome Biol 9 Suppl 2, S3 (2008)
12. Neves, M., Carazo, J.M., Pascual-Montano, A.: Extraction of biomedical events using case-based reasoning. In: Proc. of NAACL 2009. pp. 68–76 (2009)
13. Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., Salakoski, T.: Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics 9 Suppl 3, S6 (2008)
14. Segura-Bedmar, I., Martínez, P., de Pablo-Sánchezti, C.: Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform (Apr 2011)
15. Solt, I., Szidarovszky, F.P., Tikk, D.: Concept, Assertion and Relation Extraction at the 2010 i2b2 Relation Extraction Challenge using parsing information and dictionaries. In: Proc. of i2b2/VA Shared-Task. Washington, DC (2010)
16. Spasic, I., Ananiadou, S., Tsujii, J.: MaSTerClass: a case-based reasoning system for the classification of biomedical terms. Bioinformatics 21(11), 2748–2758 (Jun 2005)
17. Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., Leser, U.: A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. PLoS Comput Biol 6 (2010)
18. Zhou, X., Zhang, X., Hu, X.: Dragon Toolkit: Incorporating Auto-Learned Semantic Knowledge into Large-Scale Text Retrieval and Mining. In: Proc. of ICTAI'07. vol. 2, pp. 197–201 (2007)