

NIL_UCM: Extracting Drug-Drug interactions from text through combination of sequence and tree kernels

Behrouz Bokharaeian, Alberto Díaz

Natural Interaction Based on Language Group

Universidad Complutense de Madrid

Madrid 28011, Spain

{bokharaeian, albertodiaz}@fdi.ucm.es

Abstract

A drug-drug interaction (DDI) occurs when one drug affects the level or activity of another drug. Semeval 2013 DDI Extraction challenge is going to be held with the aim of identifying the state of the art relation extraction algorithms. In this paper we firstly review some of the existing approaches in relation extraction generally and biomedical relations especially. And secondly we will explain our SVM based approaches that use lexical, morphosyntactic and parse tree features. Our combination of sequence and tree kernels have shown promising performance with a best result of 0.54 F1 macroaverage on the test dataset.

1 Introduction

A drug-drug interaction occurs when one drug affects the level or activity of another drug, for instance, drug concentrations. This interaction can result on reducing its effectiveness or possibly increasing its side effects (Stockley, 2007). There are some helpful DDIs but most of them are dangerous (Aronson, 2007), for example, patients that take *clarithromycin* and *glibenclamide* concurrently may experiment *hypoglycaemia*.

There is a great amount of information about DDI described in papers that health experts have to consult in order to be updated. The development of tools for extracting this type of information from biomedical texts would produce a clear benefit for these professionals reducing the time necessary to review the literature. Semeval 2013 DDI Extraction challenge decided to being held with the aim of identifying the

state of the art algorithms for automatically extracting DDI from biomedical articles. This challenge has two tasks: recognition and classification of drug names and extraction of drug-drug interactions. For the second task, the input corpus contains annotations with the drug names.

A previous Workshop on Drug-Drug Interaction Extraction (Segura-Bedmar et al., 2011) was held in 2011 in Huelva, Spain. The main difference is that the new challenge includes the classification of the drug-drug interactions in four types depending on the information that is described in the sentence making the task much more complicated than before. Additionally the current task involves DDIs from two different corpora with different characteristics (Segura-Bedmar et al., 2013).

We participated in the task of extracting drug-drug interactions with two approaches that exploit a rich set of tree and sequence features. Our implemented methods utilize a SVM classifier with a linear kernel and a rich set of lexical, morphosyntactic and semantic features (e.g. trigger words) extracted from texts. In addition some tree features such as shortest path and subtree features are used.

2 Related work

Due to the importance of detecting biological and medical relations several methods have been applied for extracting biological relation information from text. In (Song et al., 2010) is presented a method for extracting protein-protein interaction (PPI) through combination of an active learning technique and a semi-supervised SVM.

Another motivating work can be found in (Chen et

al., 2011) in which a PPI Pair Extractor was developed that consists of a SVM for binary classification which exploits a linear kernel with a rich set of features based on linguistic analysis, contextual words, interaction words, interaction patterns and specific domain information.

Another PPI extraction method have been developed in (Li et al., 2010). They have applied an ensemble kernel composed of a feature-based kernel and a structure-based kernel. A more recent research on tree kernels has been carried out by (Guodong et al., 2010). They have introduced a context-sensitive convolution tree kernel, which specifies both context-free and context-sensitive sub-trees by taking into account the paths of their ancestor nodes as their contexts to capture structural information in the tree structure. A recent work (Simões et al., 2013) has introduced an approach for Relationship Extraction (RE) based on labeled graph kernels. The proposed kernel is a specification of a random walk kernel that exploits two properties: the words between the candidate entities and the combination of information from distinct sources. A comparative survey regarding different kernel based approaches and their performance can be found in (Frunza and Inkpen, 2008).

Using external knowledge and resources to the target sentence is another research direction in the relation extraction task that Chan and Roth have investigated in (Chan and Roth, 2010). They have reported some improvements by using external sources such as Wikipedia, comparing to basic supervised learning systems. Thomas and his colleagues in (Thomas et al., 2011) have developed a majority voting ensemble of contrasting machine learning methods using different linguistic feature spaces.

A more systematic and high quality investigation about feature selection in kernel based relation expression can be found in (Jiang and Zhai, 2011). They have explored a large space of features for relation extraction and assess the effectiveness of sequences, syntactic parse trees and dependency parse trees as feature subspaces and sentence representation. They conclude that, by means of a set of basic unit features from each subspace, a reasonably good performance can be achieved. But when the three subspaces are combined, the performance can

slightly improve, which shows sequence, syntactic and dependency relations have much overlap for the task of relation extraction.

Although most of the previous researches in biomedical domain has been carried out with respect to protein-protein interaction extraction, and more recently on drug-drug interaction extraction, other types of biomedical relations are being studied: e.g. gene-disease (Airola et al., 2008), disease-treatment (Jung et al., 2012) and drug-disease.

3 Dataset

The dataset for the DDIEExtraction 2013 task contains documents from two sources. It includes MedLine abstracts and documents from the DrugBank database describing drug-drug interactions (Segura-Bedmar et al., 2013). These documents are annotated with drug entities and with information about drug pair interactions: true or false.

In the training corpus the interaction type is also annotated. There are 4 types of interactions: *effect*, *mechanism*, *int*, *advice*.

The challenge corpus is divided into training and evaluation datasets (Table 1). The DrugBank training data consists of 572 documents with 5675 sentences. This subset contains 12929 entities and 26005 drug pair interactions. On the other hand, the MedLine training data consists of 142 abstracts with 1301 sentences, 1836 entities and 1787 pairs.

The distribution of positive and negative examples are similar in both subsets, 12.98% of positives instances on MedLine and 14.57% on DrugBank. With respect to the distribution of categories, the figures show that there is a small number of positive instances for categories *int* and *advice* on the MedLine subset. The *effect* type is the most frequent, outmatching itself on the MedLine subset.

The evaluation corpus contains 158 abstracts with 973 sentences and 5265 drug pair interactions from Drugbank, and 33 abstracts with 326 sentences and 451 drug pair interactions from Medline. It is worth to emphasize that the distribution of positive and negative examples is a bit greater (2.22%) in the DrugBank subset compared to the training data, but is almost doubled with respect to MedLine (12,98% to 21,06%). The categories *advice* and *int* have very few positive instances in the MedLine subset.

Training	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	26005	22217	3788	1535	1257	818	178
MedLine	1787	1555	232	152	62	8	10
Test	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	5265	4381	884	298	278	214	94
MedLine	451	356	95	62	24	7	2

Table 1: Basic statistics of the training and test datasets.

4 Method

Initially several experiments have been developed to explore the performance of shallow linguistic (SL) and parse tree based methods on a subset of the training corpus. Although the SL kernel achieved considerably good results we have found that the best option was the combination of different kernels using linguistic and tree features.

Our implemented kernel based approach consists of four different processes that have been applied sequentially: preprocessing, feature extraction, feature selection and classification (Figure 1). Our two submitted results were obtained by two different strategies. In the first outcome, all the DDIs and type of interactions were extracted in one step, as a 5-class categorization problem. The second run was carried out in two steps, initially the DDIs were detected and then the positively predicted DDIs were used to determine the type of the interaction. In the next subsection the four different processes are described.

4.1 Preprocessing

In this phase we have carried out two types of text preprocessing steps before training the classifier.

We have removed some stop words in special places in the sentences that clearly were a matter of concern and caused some inaccuracy, for example, removing question marks at the beginning of a sentence. We also carried out a normalization task for some tokens because of usage of different used encodings and processing methods, mainly html tags.

4.2 Feature extraction

Initially 49 feature classes were extracted for each instance that correspond to a drug pair interaction between Drug1 and Drug2:

- **Word Features:** Include Words of Drug1, words of Drug2, words between Drug1 and Drug2,

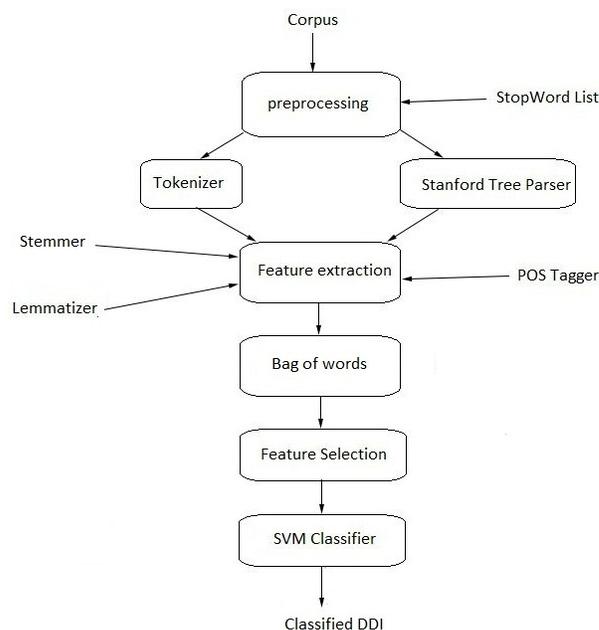


Figure 1: The different processes followed for our two submitted results.

three words before Drug1 and three words after Drug2. Lemmas and stems of all these words. We have used TreeTagger to obtain lemmas and Paice/Husk Stemmer (Paice, 1990) to obtain stems.

- **Morphosyntactic Features:** Include Part-of-speech (POS) tags of each drug words (Drug1 and Drug2), POS of the previous 3 and next 3 words. We have used TreeTagger.
- **Constituency parse tree features:** Include shortest path between Drug1 and Drug2 in the constituency parse tree, shortest path between first token in the sentence and Drug1, and shortest path between Drug2 and last token in the sentence in the parse tree, and all subtrees gener-

ated from the constituency parse tree. We have used Stanford parser ¹ for producing tree features.

- **Conjunction features:** We have produced some new conjunction features by combination of different word features and morphosyntactic features such as POSLEMMA and POSSTEM for all the words before Drug1, words between Drug1 and Drug2 and words after Drug2.
- **verbs features:** Include verbs between Drug1 and Drug2, first verb before Drug1 and first verb after Drug2. Their stem, lemma and their conjunction features are also included.
- **negation features:** Only if the sentence contains negation statements. The features extracted include the left side tokens of the negation scope, the right side tokens of the negation scope and the tokens inside the negation scope. We have used NegEx² as negation detection algorithm.

Finally we have deployed a bag of words approach (BoW) for each feature class in order to obtain the final representation for each instance. We have limited the size of the vocabulary in the BoW representation with a different number depending on the data subset. We carried out several experiments to fix these numbers and at the end we have used 1000 words/feature class for MedLine and 6000 words/feature class for DrugBank.

4.3 Feature selection

We have conducted some feature selection experiments to select the best features for improving the results and reducing running time. We have finally used Information Gain ranker to eliminate the less effective features. We have computed the information gain for each feature class as the linear combination of the information gain of each corresponding word. Empirically we have selected the best 42 feature classes.

On the other hand, we have done a preliminary study of the effect of the negation related features. We have found more than 3000 sentences containing negation, most of them corresponds to sentences

¹<http://nlp.stanford.edu/software/lex-parser.shtml>

²<http://code.google.com/p/negex/>

associated with negative examples of interactions. However, these features have been eliminated because we have not obtained a clear improvement when we combined them with the other features.

4.4 Classification

First we have performed several experiments with different supervised machine learning approaches such as SVM, Naivebayes, Randomtree, Random forest, Multilayer perceptron in addition to combination of methods. Finally we decided to use a SVM approach, the Weka Sequential Minimal Optimization (SMO) algorithm. We used the inner product of the BoW vectors as similarity function.

We have submitted two approaches:

- **approach 1:** SVM (Weka SMO) with 5 categories (effect, mechanism, int, advice and null).
- **approach 2:** We have extracted final results in two stages. In the first step we have used a SVM (Weka SMO) with 2 categories (positive and negative) and then we have used a second SVM classifier with 4 classes on positive extracted DDIs to train and extract the type of interaction in the test dataset.

The classifiers have been applied separately with each data subset, that is, a classifier per approach has been developed using the DrugBank training subset and has been evaluated using the DrugBank test subset, and the same process has been applied with the MedLine training and test subset.

5 Results

In this section we first show the evaluation results with our two approaches. Secondly an error analysis was carried out with a development set extracted from the training corpus.

5.1 Test data results

We have submitted two runs that corresponds with the approaches described in the previous section. Table 2 shows the results obtained with the first approach (one step) and Table 3 shows the results with the second approach (two steps).

It can be observed that the results on detection of DDI are better with the approach 2: 0.656 against 0.588 on F1. This result is a consequence that we

Run	P	R	F1
NILUCM1 (All)	0.632	0.464	0.535
NILUCM2 (All)	0.547	0.507	0.526
NILUCM1 (Drugbank)	0.651	0.498	0.565
NILUCM2 (Drugbank)	0.558	0.542	0.550
NILUCM1 (Medline)	0.333	0.074	0.121
NILUCM2 (Medline)	0.221	0.073	0.110

Table 4: Macroaverage test set results.

have more information to obtain the detection of the interaction if we use the information from all the different types than if we obtain it joining the results obtained per each category. With respect to detection and classification the results are also better with approach 2 for a similar reason: 0.548 against 0.517 on F1.

With respect to the categories, in the more populated ones the general tendency of better results from approach 2 continues, especially in *effect* type: 0.556 against 0.489. With respect to *advice* and *int*, the recall is better in approach 2 but the improvement in precision is greater in approach 1 giving a better result on F1 to approach 1, especially in *int* type: 0.427 against 0.393.

Table 4 shows the macroaverage results separated by subset data. The best results obtained for approach 1 are due to that this type of average gives equal weight to each category, favouring then the categories with less instances.

Other important insight that can be extracted from this table is that our results are much better for DrugBank dataset with both approaches. These results can be justified due to high similarity between sentences in Drugbank training and test corpus. In fact the Medline corpus commonly has more words unrelated to DDI subjects. In addition to this argument, the smaller number of training pairs in the Medline corpus can be other reason to obtain worst results.

5.2 Error analysis

We have extracted a stratified development corpus from the training corpus in order to perform an error analysis. We have used a 10% of the training corpus. It contains 2779 pairs, of which 397 are DDIs. Table 5 shows the results obtained with the two submitted approaches.

The results with our development corpus shows the same tendency, that is, approach 2 is better than approach 1 on detection of DDI and on microaverage classification. On the other hand, results are higher than those on test corpus because the information contained in the development corpus is more similar to the rest of training corpus than information on the test set.

We have performed an analysis of the errors produced for both approaches in the Detection and Classification of DDI subtask. The errors obtained are 112 false positives (Fp) and 116 false negatives (Fn) associated to approach 1, and 111 false positives (Fp) and 112 false negatives (Fn) to approach 2. Apart from the comments explained in the previous section about the small number of instances on the MedLine subset, we think the main problem is related with the management of long sentences with complex syntax. These sentences are more difficult for our approaches because the complexity of the sentence generates more errors in the tokenizing and parsing processes affecting the representation of the instances both in training and test phases. We show below some false positives and false negatives examples.

- The effects of **ERGOMAR** may be potentiated by **triacetyloleandomycin** which inhibits the metabolism of ergotamine. DrugBank. False negative.
- Prior administration of **4-methylpyrazole** (90 mg kg(-1) body weight) was shown to prevent the conversion of **1,3-difluoro-2-propanol** (100 mg kg(-1) body weight) to (-)-erythrofluorocitrate in vivo and to eliminate the fluoride and citrate elevations seen in 1,3-difluoro-2-propanol-intoxicated animals MedLine. False negative.
- Drug Interactions with Antacids Administration of 120 mg of **fexofenadine hydrochloride** (2 x 60 mg capsule) within 15 minutes of an aluminum and magnesium containing antacid (Maalox) decreased **fexofenadine** AUC by 41% and cmax by 43%. DrugBank. False positive.
- **Dexamethasone** at 10(-10) M or retinyl acetate

approach 1	Tp	Fp	Fn	total	P	R	F1
Detection of DDI	557	359	422	979	0.608	0.569	0.588
Detection and classification of DDI	490	426	489	979	0.535	0.501	0.517
Score for type mechanism	147	122	155	302	0.546	0.487	0.515
Score for type effect	200	258	160	360	0.437	0.556	0.489
Score for type advice	115	39	106	221	0.747	0.520	0.613
Score for type int	28	7	68	96	0.800	0.292	0.427

Table 2: Test corpus results (approach1).

approach 2	Tp	Fp	Fn	total	P	R	F1
Detection of DDI	631	315	348	979	0.667	0.645	0.656
Detection and classification of DDI	527	419	452	979	0.557	0.538	0.548
Score for type mechanism	146	102	156	302	0.589	0.483	0.531
Score for type effect	210	186	150	360	0.530	0.583	0.556
Score for type advice	139	96	82	221	0.591	0.629	0.610
Score for type int	32	35	64	96	0.478	0.333	0.393

Table 3: Test corpus results (approach2).

approach 1	Tp	Fp	Fn	total	P	R	F1
Detection of DDI:	292	101	105	397	0.743	0.736	0.739
Detection and Classification of DDI:	281	112	116	397	0.715	0.708	0.711
approach 2	Tp	Fp	Fn	total	P	R	F1
Detection of DDI:	296	102	101	397	0.744	0.746	0.745
Detection and Classification of DDI:	285	111	112	397	0.720	0.718	0.719

Table 5: Error analysis with a development corpus.

at about 3×10^{-9} M inhibits **proliferation** stimulated by EGF. MedLine. False positive.

6 Conclusions

In this paper we have shown our approaches for the Semeval 2013 DDI Extraction challenge. We have explored different combinations of tree and sequence features using the Sequential Minimal Optimization algorithm.

The first approach uses a SVM with 5 categories, and the second one extracts the final results in two steps: detection with all the categories, and classification on the positive instances. The results are better for approach 2 mainly due to the improvement on the detection subtask because the information from all the categories is used.

We think some of our errors come from using a general tool (Stanford parser) to obtain the parse tree

of the sentences. In the future we are going to explore other biomedical parsers and tokenizers.

With respect to the data used, we think the MedLine dataset needs to be greater in order to obtain more significant analysis and results. Our approaches are especially affected by this issue because the small number of positive instances on *advice* and *int* categories implies that the algorithm can not learn to classify new instances accurately. On the other hand, although n-fold cross validation is considered as the best model validation technique, it was time consuming for DDI and need powerful processors.

Another interesting future work is related with the application of simplification techniques in order to solve the problems in the processing of complex long sentences (Buyko et al., 2011).

References

- A. Airola, S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter, T. Salakoski. 2008. Allpaths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning *BMC Bioinformatics*, 9(Suppl 11):S2.
- JK. Aronson. 2007. Communicating information about drug interactions. *British Journal of Clinical Pharmacology*, 63(6):637–639.
- E. Buyko, E. Faessler, J. Wermter, U. Hahn. 2011. Syntactic Simplification and Semantic Enrichment - Trimming Dependency Graphs for Event Extraction. *Computational Intelligence*, 27(4):610–644.
- Y. Chen, F. Liu, B. Manderick. 2011. Extract Protein-Protein Interactions from the Literature Using Support Vector Machines with Feature Selection. *Biomedical Engineering, Trends, Researchs and Technologies*, 2011.
- YS. Chan and D. Roth. 2010. Exploiting Background Knowledge for Relation Extraction *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*, pp:152–160.
- O. Frunza and D. Inkpen. 2010. Extraction of disease-treatment semantic relations from biomedical sentences *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp:91–98.
- Z. Guodong, Q. Longhua, F. Jianxi. 2010. Tree kernel-based semantic relation extraction with rich syntactic and semantic information *International Journal on Information Sciences*, 180(8):1313–1325.
- J. Jiang and C. Zhai. 2011. A systematic exploration of the feature space for relation extraction *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACLHLT07)*, pp:113–120.
- H. Jung, S. Choi, S. Lee, S. Song. 2012. Survey on Kernel-Based Relation Extraction.
- L. Li, J. Ping, D. Huang. 2010. Protein-Protein Interaction Extraction from Biomedical Literatures Based on a Combined Kernel *Journal of Information & Computational Science*, 7(5):1065–1073.
- Chris D. Paice. 1990. Another stemmer. *ACM SIGIR Forum*, 24(3):56–61.
- I. Segura-Bedmar, P. Martínez, D. Sánchez-Cisneros. 2011. *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)* CEUR Workshop Proceedings, Vol. 761.
- I. Segura-Bedmar, P. Martinez, M. Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts. *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.
- G. Simões, D. Matos, H. Galhardas. 2013. A Labeled Graph Kernel for Relationship Extraction. *CoRR*, abs/1302.4874.
- M. Song, H. Yu, W. Han. 2010. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. *International Workshop on Data Mining in Bioinformatics*.
- I H Stockley. 2007. *Stockley's Drug Interaction*. Pharmaceutical Press.
- P. Thomas, M. Neves, I. Solt, D. Tikk, U. Leser. 2011. Relation extraction for drug- drug interactions using ensemble learning *Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pp:11–17.