UC3M: A kernel-based approach to identify and classify DDIs in biomedical texts.

Daniel Sanchez-Cisneros
Universidad Carlos III de Madrid
Avda. de la Universidad, 30
28911 Leganés - Madrid - Spain
dscisner@inf.uc3m.es

Abstract

The domain of DDI identification is constantly showing a rise of interest from scientific community since it represents a decrease of time and healthcare cost. In this paper we purpose a new approach based on shallow linguistic kernel methods to identify DDIs in biomedical manuscripts. The approach outlines a first step in the usage of semantic information for DDI identification. The system obtained an F1 measure of 0.534.

1 Introduction

In recent years a new discipline appeared in the biomedical domain for processing pharmacological manuscripts related to drug substances. This discipline is the so called *Pharmacovigilance*, and takes care of the management and control of Drug-Drug interactions (DDI) among other faculties. A DDI occurs when one drug influences the effect level or activity of another drug.

Some events such as *BioCreative*[1] and *BioNLP*[2] establish a benchmark of comparison in the field of natural language processing applied to biomedical domain. This is the case of *Semeval 2013: Extraction of Drug-Drug Interactions from BioMedical Texts*[3], where our system has been evaluated.

The field of DDI extraction from biomedical text has been faced from different perspectives such as rule-based approaches, SVM approaches and kernel-methods approaches, among others.

Segura-Bedmar et al. (2010) proposed an approach to extract DDI from biomedical texts based on Shallow Linguistic (SL) Kernel (Giuliano et al., 2006) methods obtaining an F1 measure of 60,01%. The system was evaluated over a DrugDDI dataset created in 2010 that contains 579 biomedical documents collected from the pharmacological database *DrugBank*[4]. The dataset contains a total of 3,160 DDIs.

Recently, the DDIExtraction2011 task[5] compared the latest advances in Information Extraction techniques applied to the DDI identification. The event provided a benchmark forum of 10 different approaches. The evaluation of the systems was made over the DrugDDI dataset. We now describe the most relevant works.

Thomas et al. (2011) developed a system by combining a preprocessing phase based on Charniak-Lease (Lease, Charniak, 2005) and Stanford (Marneffe et al., 2006) parsers, with a classification phase based on SL kernel (Giuliano et al., 2006), k-Band Shortest Path Spectrum (kBSPS) kernel (Airola et al., 2008), All Path Graphic (APG) kernel (Tikk et al., 2010) and case-based reasoning (CBR) (Aamodt, Plaza, 1994) techniques. The system obtained a F1 measure of 65.7%.

Chowdhury et al. (2011) presented a system combining a preprocessing phase based on Stanford parser and SPECIALIST (Browne, 2000) lexicon tool, with a classification phase based on Featured-Based kernel such as SL kernel and Tree-Based kernel such as Dependency tree (DT) kernel (Culotta and Sorensen, 2004) and Phrase Structure Tree (PST) kernel (Moschitti, 2004). The system achieved an F1 of 63.7%.

[1] http://www.biocreative.org/
[2] http://2013.bionlp-st.org/
[3] http://www.cs.york.ac.uk/semeval-2013/task9/
[4] http://www.drugbank.ca/
[5] http://labda.inf.uc3m.es/DDIExtraction2011/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 617–621, Atlanta, Georgia, June 14-15, 2013.

Björne et al. (2011) proposed a different approach by combining a preprocessing phase based on a collection of features and n-grams; with a classification based on support vector machine (SVM) (Vapnik, 1995). The SVM methods perform classification tasks by building hyperplanes in a multidimensional space that divide cases of different classes (binary classification). The system yielded an F1 measure of 62.99%.

Kernel methods seem to be the best choice for extracting DDI since they obtained the highest results. Thus, we decided to use kernel methods to identify and classify DDI in our system. Furthermore, we hypothesize that using semantic features of pharmacological substances, can provide valuable knowledge in the classification phase. Therefore, we decide to integrate semantic information in the classification process of kernel methods.

In this paper we present a kernel-based approach to identify and classify DDIs in biomedical text by using SL kernels. In section 2 we describe the system used for identifying DDIs. Section 3 present the results obtained by the system and a little comparison with other approaches. In section 4 we expose some conclusions obtained and ideas for future work.

2 Description of the systems

The system (see figure 1) is divided in three phases: (i) in the first phase the system makes a preprocessing of the documents in order to extract grammatical and semantic information about each word of the text. (ii) The second phase makes the classification of whether a pair of drugs is a DDI or not by using SL kernel methods. (iii) In the third phase, the system classifies all DDIs into the purpose type (*advice, effect, mechanism, int*) using SL kernel methods.

The corpus is processed sentence by sentence, using the identification tag provided for each sentence.

2.1 Preprocessing

In this phase we make a preprocessing of the documents to obtain linguistic and semantic information about the words and entities contained in the text. Since linguistic and semantic

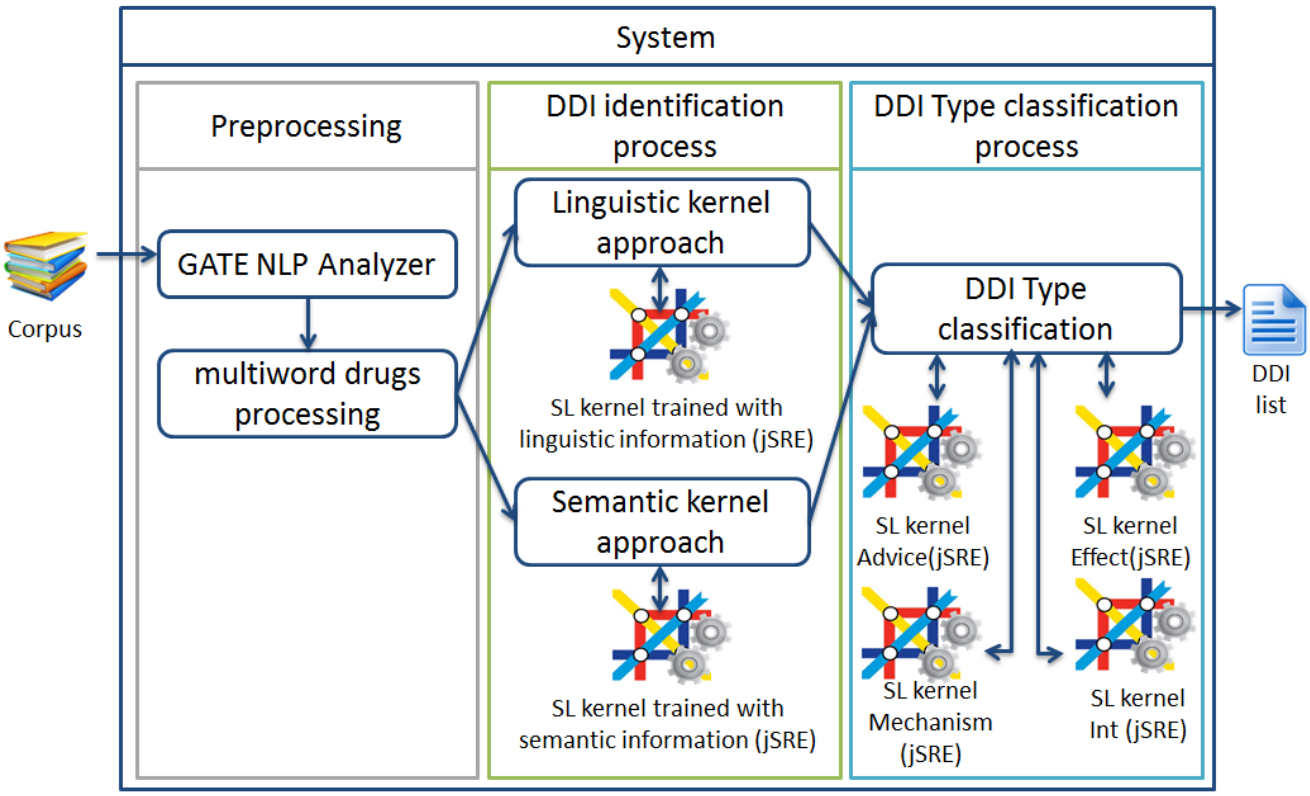

Figure 1: Architecture of the system.

approaches are based on different types of information, our participation in the task will be separated in two runs: first run will be based on linguistic information and second run will be based on semantic information.

Firstly, we process each sentence and obtain linguistic information about *part-of-speech* (PoS) tagging and lemmatization for each word contained in the text. To do so we use the Stanford parser[6] by using the GATE analyzer[7]. The result of this step is a list of words and PoS tags, but entity concepts are missing. Therefore, we make a multiword entities processing to keep the words related to the same concept together. For example, the entity *beta-adrenergic receptor blocker* is processed by Stanford parser as three different annotations nodes: *beta-adrenergic* as type JJ; *receptor* as type NN; and *blocker* as type NNS. Thus we unify the three words into an only one concept *beta-adrenergic_receptor_blocker* as type NNP. This information corresponds to the linguistic approach of our participation in the task (see figure 2b).

On the other hand, we process the text and collect semantic information about Anatomical Therapeutic Chemical (ATC) identification for each drug found in the text. The ATC code is a widely used classification system provided from WHO collaborating centre for Drug statistics methodology. The classification divides drugs in groups at five different levels according to the organ or system on which they act, and their therapeutic, pharmacological and chemical properties. The system obtains the ATC code of the drugs by searching the drug entities in the ATC Index resource[8]. Then, we associate the ATC code results with the drug entity. This information corresponds to the semantic approach of our participation in the task.

2.2 Identification of DDI

In this phase the system will predict whether a pair of drugs is a DDI or not by the use of Shallow linguistic Kernel methods. To do so we use the jSRE tool[9].

In one hand, the linguistic approach is based on shallow linguistic information such as PoS tagging and lemmatization. Therefore, the information introduced into the SL kernel model consists of: *token_identifier*, *ATC_code*, *token_lemmatization*, *POS_tag*, *entity_type* and *entity_label*; as show in figure 2b.

On the other hand, the semantic approach uses the semantic information of drugs (ATC codes) to increase the available knowledge in the kernel classification process. To do so, we trained a SL kernel model by replacing the token value with the ATC code value. In case of a non-drug token, we replace the token value with 0. This way the information introduced to the SL kernel model consists of: *token_identifier*, *ATC_code*, *token_lemmatization*, *POS_tag*, *entity_type* and *entity_label*; as show in figure 2c.

```
<Feature>
  <Name className="java.lang.String">string</Name><Value className="java.lang.String">beta-adrenergic</Value>
  <Name className="java.lang.String">category</Name><Value className="java.lang.String">JJ</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">string</Name><Value className="java.lang.String">receptor</Value>
  <Name className="java.lang.String">category</Name><Value className="java.lang.String">NN</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">string</Name><Value className="java.lang.String">blockers</Value>
  <Name className="java.lang.String">category</Name><Value className="java.lang.String">NNS</Value>
</Feature>
```

Figure 2a: Example of separated multiword entity.

```
8&&beta-adrenergic_receptor_blocker&&beta-adrenergic_receptor_blocker&&NNP&&group&&A
```

Figure 2b: Example of linguistic input token into the SL kernel.

```
0&&B01AB&&heparin&&NNP&&drug&&A
```

Figure 2c: Example of semantic input token into the SL kernel.

[6] http://nlp.stanford.edu/software/lex-parser.shtml.
[7] http://gate.ac.uk.
[8] http://www.whocc.no/atc_ddd_index/
[9] http://hlt.fbk.eu/en/technology/jSRE

2.3 Type classification of DDI

In the third phase, the system makes a classification of DDIs to determine the type of the interaction. To do so, the system face the classification task as a machine learning task, and use SL kernel methods. Hence, we train one SL kernel model for each possible values of DDI type: *advice, effect, mechanism, int*. To train the kernel models we separate by type each DDI of the training dataset. The result is four groups of training dataset, where the correspondent type class value are set to 1, and 0 otherwise. Once we trained the kernel models, each DDI go through four different prediction processes. The conflictive cases are solved by frequency of appearance. This step is the same for both linguistic and semantic approach. Finally, we collect the results and generate the task output format.

3 Results

The best result in DDI detection and classification (macro-average score) were obtained by the linguistic approach (run 2), achieving a F1 measure of 0.534.

	DDI Detection			DDI Detection and Classification (micro-average)			DDI Detection and Classification (macro-average)		
Team	P	R	F1	P	R	F1	P	R	F1
Run 1	0,632	0,725	**0,676**	0,495	0,568	**0,529**	0,527	0,541	**0,534**
Run 2	0,404	0,798	0,537	0,222	0,437	0,294	0,275	0,43	0,335

Table 1: Results obtained by the system.

Focusing on DDI detection results, we can see that linguistic approach also overcome the semantic approach, obtaining a F1 score of 0.676 and 0.537 respectively. This can be explained since the SL kernel optimizes linguistic information rather than semantic information. Therefore, ATC code format is not appropriate for SL kernel.

However, the score obtained by the linguistic approach using SL kernel with multiword entities processing seems to be higher than the average results obtained in DDIExtraction 2011 task. This may be due to the great improvement that DrugDDI corpus suffered since the last competition, by enriching the information of each entity.

Finally, we have a word to notice the decrease of the results from DDI detection evaluation to DDI detection and classification evaluation. This could be due to the complexity of the DDI type classification task. However, the final result of macro-average score shows huge margin of improvement.

4 Conclusion and future work

In this paper we present a kernel based approach to identify and classify DDIs by using SL kernel. The result obtained by the system achieves 0.534 F1 measure. From linguistic approach and semantic approach purposed for the participation in the task, the linguistic approach shows better results. However, we can not discard semantic information since we may have not used the appropriate semantic information for a shallow linguistic kernel.

Thus, a possible future work could be the research in semantic information processing to help in the classification process. Therefore, another future work could be the integration of pharmacological ontologies in the classification process since they increase the knowledge available for the classification task.

Acknowledgments

This work has been funded by MA2VICMR project (S2009/TIC-1542) and MULTIMEDICA project[10] (TIN 2010-20644-C03-01).

References

Aamodt A., Plaza E. 1994. *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*. AI Communications 7(1), P 39–59.

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., Salakoski, T. 2008. *Allpaths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning*. BMC Bioinformatics, 9 S11, S2.

Björne J., Airola A., Pahikkala T., Salakoski T. 2011. *Drug-Drug interaction extraction from biomedical*

[10] http://labda.inf.uc3m.es/multimedica/

texts with SVM and RLS classifiers. Proceedings of DDIExtraction2011, SEPLN 2011.

Browne A.C., McCray A.T., Srinivasan S. 2000. *The SPECIALIST Lexicon*. NLM, Bethesda.

Chowdhury MFM, Lavelli A. 2011. *Drug-drug Interaction Extraction Using Composite Kernels*. Proceedings of DDIExtraction2011, SEPLN 2011.

Giuliano C, Lavelli A, Romano L. 2006. *Exploiting shallow linguistic information for relation extraction from biomedical literature*. Proceedings of EACL 2006.

Culotta A., Sorensen J. 2004. *Dependency tree kernels for relation extraction*. Proceedings of the 42nd annual metting of the Association for Computational Linguistics.

Lease, M., Charniak, E. 2005. *Parsing biomedical literature*. Proceedings of IJCNLP'05.

Marneffe M.C., MacCartney B., Manning C.D. 2006. *Generating Typed Dependency Parses from Phrase Structure Parses*. Proceedings of LREC 2006.

Moschitti, A. 2004. *A study on convolution kernels for shallow semantic parsing*. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. ACL '04.

Segura-Bedmar, I., Martínez, P., Pablo-Sánchez, C.d. 2010. *Using a shallow linguistic kernel for drug-drug interaction extraction*. BMC BioInformatics.

Thomas P., Neves M., Solt I., Tikk D., Leser U. 2011. *Relation Extraction for Drug-Drug Interaction using Ensemble Learning*. Proceedings of DDIExtraction2011, SEPLN 2011.

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., Leser, U. 2010. *A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature*. PLoS Comput Biol 6.

Vapnik, V.N. 1995. *The nature of statistical learning theory*. Springer-Verlag New York.