Drug-Drug Interactions Discovery Based on CRFs, SVMs and Rule-Based Methods

Stefania Rubrichi, Matteo Gabetta, Riccardo Bellazzi, Cristiana Larizza, and Silvana Quaglini

Laboratory for Biomedical Informatics "Mario Stefanelli",
Department of Computers and Systems Science,
University of Pavia, Pavia, Italy

Abstract. Information about medications is critical in improving the patients' safety and quality of care. Most adverse drug events are predictable from the known pharmacology of the drugs and many represent known interactions and are, therefore, likely to be preventable. However, most of this information is locked in free-text and, as such, cannot be actively accessed and elaborated by computerized applications. In this work, we propose three different approaches to the problem of automatic recognition of drug-drug interactions that we have developed within the "First Challenge Task: Drug-Drug Interaction Extraction" competition. Our approaches learn to discriminate between semantically interesting and uninteresting content in a structured prediction framework as well as a rule-based one. The systems are trained using the DrugDDI corpus provided by the challenge organizers. An empirical analysis of the three approaches on this dataset shows that the inclusion of rule-based methods is indeed advantageous.

Keywords: Drug-Drug Interactions, Information Extraction, Conditional Random Fields, Support Vector Machines, Adverse Drug Events

1 Background

The use of medications has a central role in health care provision, yet on occasion it may endanger patients' safety and account for increased health care costs, as result of adverse drug events (ADEs). Many of these injuries are inevitable, but at least a quarter may be secondary to medication errors [7] that can be avoidable. That is the case of ADEs due to drug-drug interactions (DDIs), since many of them are due to disregarded known interactions and are therefore likely to be preventable. Over the 6.5% of drug-related hospital admissions are a consequence of DDIs.

DDIs are a common problem during drug treatment. Widely, a drug interaction represents the situation in which a substance affects the activity of an active ingredient, resulting in various effects such as alterations in absorption,

Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), pages 67–74,
Huelva, Spain, September 2011.

metabolism, excretion, and pharmacodynamics (i.e. the drug effects are decreased or increased, or the drug produces a new effect that neither produces on its own). Safe medication use requires that prescribers receive clear information on the medication itself including information about any potential interactions. This information is constantly changing, and while most of the necessary updated knowledge is available somewhere, it is not always readily accessible. In particular, most of this information is locked in free-text, then cannot be actively used by health information systems. Reliable access to this comprehensive information, by Natural Language Processing (NLP) systems, can represent a useful tool for preventing medication errors and, more specifically, DDIs. Over the last two decades there has been an increase of interest in applying NLP, in particular information extraction (IE) techniques, to biomedical text. Excellent efforts have been documented in the medication domain literature on IE from textual clinical documents [4,5,9,11,12,14,15,18], and its subsequent application in summarization, case finding, decision-support, or statistical analysis tasks.
In this context, we accepted the challenge presented within the "First Challenge Task: Drug-Drug Interaction Extraction" competition and developed a system for the automatic extraction of DDIs from a corpus [13] of documents, collected from the DrugBank database [8], describing, for each drug, the relating DDIs.

2 Methods

On the following section we present the proposed system and its components.

2.1 System Outline

We exploit three different approaches, which rely upon different methods for the extraction of such information. The first approach (henceforth referred as hybrid approach) is twofold: it combines a supervised learning technique based on Conditional Random Fields (CRFs) [16] with a rule-based method. We modeled the problem as follows: in a first step we employed the CRFs classifier in order to assign the correct semantic category to each word, or segment of sentence, of the text. We considered the following three semantic categories:

1. *DrugNotInteracting*: describes a drug entity, which is not involved in an interaction;
2. *DrugInteracting*: describes a drug entity, which is involved in an interaction;
3. *None*: indicates elements that are not relevant for this task.

Once every potential interacting entity has been identified by the CRFs classifier, we defined a set of rules for the construction of the actual pairs of interacting entities, and match them with the sentences.
The second (henceforth referred as pair-centered CRFs approach) and third (henceforth referred as pair-centered SVMs approach) approaches are very similar: they are both based on supervised learning methods, CRFs and Support

Vector Machines (SVMs) [2,17], respectively. In this case we focused on the single pair of drug entities: for any given pair in a sentence, such techniques predict the presence or absence of interaction relation, relying on a set of hundreds of engineered features, which take into account the properties of the text, by learning the correspondence between semantic categories and features. We considered only two semantic categories:

1. *Interaction*: describes a pair of drug entities which interact;
2. *NotInteraction*: describes a pair of drug entities which don't interact;

All these three methodologies have been developed through different steps. We began with a pre-processing pass over the corpus in order to prepare the dataset for the use by the extraction module. Then, we defined a set of binary features that express some descriptive characteristics of the data, and we converted the data in a set of corresponding features. Finally, we processed the data through the three methodologies described above.

2.2 Supervised Learning Methods: CRFs and SVMs

Supervised learning approaches have been widely applied to the domain of IE from free text. A typical application of supervised learning works to classify a novel instance x as belonging to a particular category y. Given a predefined set of categories, such methods use a set of training examples to take decision in front of new examples. They automatically tune their own parameters to maximize their performance on the training set and then generalize from the new samples. We processed the data through the two linear classifiers, CRFs and SVMs: both algorithms iterate the tokens in the sentence, and label proper tokens with semantic categories. These classifiers discriminate between semantically interesting and uninteresting content through the automatic adaptation of a large number of interdependent descriptive characteristics (features) taking into account the properties of the input text. Each token is represented by a set of features, then the classifiers learn a correspondence between semantic categories and features, and assign real-valued weight to such features.

2.3 Pre-processing

The first step of our DDIs detection system has been a pre-processing over the data provided within the challenge contest.
We designed two different pre-processing strategies, one for the hybrid approach, the other one for the pair-centered CRFs and the pair-centered SVMs approach. The first pre-processing strategy analyzes sentence-by-sentence the training corpus, using a quite classical NLP system developed using Gate [3], an open source framework for language processing. This system includes:

- Tokenizer: splits the atomic parts of the sentence (tokens) according to a specific language (English in our case);

– Part of Speech (POS) Tagger [6]: assigns to the tokens their grammatical class (e.g. noun, verb, adjective . . .);
– Morphological Analyzer: assigns the lexical roots to the tokens;
– UMLS concept finder: a module we developed, in order to discover concepts referable to the Unified Medical Language System (UMLS) [10] within the text.

The pre-processing system returns as output a line for each token; such line contains the token itself together with additional information necessary for the features generation task. In particular:

– the semantic category of the token itself;
– the "entity tag" that is the entity's code (e.g. DrugDDI.d385.s4.e0) when the token is an entity and null otherwise;
– the "main drug tag" that is *true* if the token matches the standard name of the referential drug[1] and *false* otherwise;
– the "brand name tag" that is *true* if the token matches one of the brand names of the referential drug and *false* otherwise. Brand names come from the DrugBank;
– the "POS tag" that is the grammatical class provided by the POS Tagger (entities are automatically tagged as proper nouns - NNP);
– the "root tag" which is the root of the token provided by the Morphological Analyzer (the entity itself for the entities);
– the "semantic group tag" that, when the token belongs to a UMLS concept, is the semantic group of the concept itself (e.g. "DISO" for concepts belonging to the "Disorders" group); it is "ENT" when the token is an entity and null otherwise.

As an example, given the input sentence:

```
<sentence id="DrugDDI.d368.s0" origId="s0" text="Itraconazole
decreases busulfan clearance by up to 25%, and may produce AUCs >
1500 muMolmin in some patients.">
    <entity id="DrugDDI.d368.s0.e0" origId="s0.p0" charOffset="0-12"
type="drug" text="Itraconazole" />
    <entity id="DrugDDI.d368.s0.e1" origId="s0.p2" charOffset="23-31"
type="drug" text="busulfan" />
    <pair id="DrugDDI.d368.s0.p0" e1="DrugDDI.d368.s0.e0"
e2="DrugDDI.d368.s0.e1" interaction="true" />
</sentence>
```

the first pre-processing strategy will generate the following lines:

```
itraconazole-DrugInteracting-DrugDDI.d368.s0.e0-false-false-NNP-
    itraconazole-ENT
decreases-None-null-false-false-NNS-decrease-CONC
busulfan-DrugInteracting-DrugDDI.d368.s0.e1-true-false-NNP-busulfan-
```

[1] We indicate by "referential drug" the drug described in the specific document under examination.

```
    ENT
clearance-None-null-false-false-NN-clearance-PHEN
...
```

and so on.

The second pre-processing strategy evaluates separately all the pairs within a sentence; it uses the same NLP system described for the first strategy, but it formats the output in a different way. For each pair, the output consists of a header line, containing the codes of the involved entities and the semantic category of the pair. The header line is followed by a line for each token standing between the two entities involved in the pair; for each line the elements describing the token are exactly the same as those described for the first strategy (*token, interaction tag, entity tag, etc.*).

Given the input sentence from the previous example, the second pre-processing strategy will generate the following lines:

```
DrugDDI.d368.s0.e0 DrugDDI.d368.s0.e1-Interaction
decreases-None-null-false-false-NNS-decrease-CONC
```

2.4 Feature Definition and Data Conversion

The feature construction process aims at capturing the salient characteristics of each token in order to help the system to predict its semantic label. Feature definition is a critical stage regarding the success of feature-based statistical models such as CRFs and SVMs. A careful inspection of the corpus has resulted in the identification of a set of informative binary features that capture salient aspects of the data with respect to the tagging task. Subsequently, the stream of tokens has been converted to features. In particular, in the pair-centered CRFs and pair-centered SVMs approaches we considered only the tokens between the two entities which form each pair. This means that features for drug entities pair E_1-E_2 contain predicates about the n tokens between E_1 and E_2.

In the following we report on the set of features used in our experiments.

Orthographical Features As a good starting point, this class of features consists of the simplest and most obvious feature set: word identity feature, that is the vocabulary derived from the training data.

Part Of Speech (POS) Features We supposed lexical information might be quite useful for identifying named entities. Thus, we included features that indicate the lexical function of each token.

Punctuation Features Also notable are punctuation features, which contain some special punctuation in sentences. After browsing our corpus we found that colon might prove helpful. Given a medication in fact, colon is usually preceded by the interacting substance and followed by the explanation of the specific interaction effects.

Semantic Features In order to have these models benefit from domain specific knowledge we added semantic features which use external semantic resources. This class of features includes:

1. *root feature*: takes account of the root associated to each word;
2. *UMLS feature*: relies on the UMLS Metathesaurus and for each word returns the corresponding semantic group;
3. *brand name feature*: it recognizes the corresponding brand names occurring in the text. DrugBank database drug entries are provided with the field "Brand Names", which contains a complete list of brand names from different manufacturers. We create a binary feature, which, every time a text token coincides with one of such names, is active, indicating that the token corresponds to a brand name of the specific referential drug;
4. *standard drug name feature*: identifies the standard name of the source drug. For each token this feature tests if it matches such standard name;
5. *drug entity feature*: allows the models to recognize the drug entities annotated by the MetaMap tool: it is active for the tokens which have been annotated as drug entity by the MetaMap tool.

Context Features Finally, we extended all the classes of feature we described above to a token window of [-k,k]. The descriptive characteristics of tokens preceding or following a target token may be useful for modeling the local context. It is clear that the more context words analyzed, the better and more precise the results could become. However, widening the context window quickly leads to an explosion of the computational and statistical complexity. For our experiments, we estimated a suitable window size of [-3,3].

2.5 Rule-based Method

As we have already stated, while both pair-centered CRFs and pair-centered SVMs approaches focus on entities pairs and predict directly the presence or absence of interaction, the first one considers a token at a time, then the semantic category prediction is on a token-by-token basis. Therefore, a further processing pass was necessary in order to build up the interaction pairs, starting from each single entity. For this purpose, we employed a rule-based method which relies upon a set of rules, manually-constructed from the training data analysis. In particular, the rules that we built to find out the interacting pairs are the following:

- if a sentence contains less than two tokens labeled as *DrugInteracting*, then no interacting pair is generated;
- an interacting pair must contain two tokens labeled as *DrugInteracting*;
- one and only one of the token involved in the interacting pair, must be the referential drug or one of its brand names.

3 Experiments

We used the Unified format of the DrugDDI corpus [1] provided by the competition organizers.

For the linear SVMs, we found the regularization parameter $\lambda = 1$ to work well. SVMs results have been produced using 10 passes through the entire training set. For the variance of the Gaussian regularizer of the CRFs we used the value 0.1.

We submitted a total of three runs: the first run includes the predictions generated by the hybrid approach; the second run includes the predictions generated by the pair-centered CRFs approach; the third run includes the predictions generated by the pair-centered SVMs approach.

4 Results and Discussion

The evaluation process was performed by the challenge organizers.
The overall results of the three approaches can be found in Table 1. In general, the hybrid approach outperforms the other two. This performance gain can be attributed to the additional contribute of rule-based method, that played an important role in building the interacting pairs. In particular it makes the system benefit from additional knowledge that facilitates the pairs disambiguation process. It specifies, for example, that a pair has to include the referential drug or one of its brand names together with another drug entity different from them.
There is room for improvement, especially for the pair-centered CRFs and pair-centered SVMs approaches. In such approaches we mainly relied on tokens occurring between the two entities which form each pair, however tokens preceding and following the pairs could also be taken into account.

Table 1. Overall experimental results of the different runs

Approach	**Hybrid**	**Pair-centered CRFs**	**Pair-centered SVMs**
True Positive	369	196	317
False Positive	545	110	456
False Negative	386	559	438
True Negative	5726	6161	5815
Precision (%)	40.37	64.05	41.01
Recall (%)	48.87	25.96	41.99
F_1 Score (%)	44.22	36.95	41.49

5 Conclusion and Future Work

In this paper we presented three different approaches for the extraction of DDIs that we have developed within the "First Challenge Task: Drug-Drug Interaction Extraction" competition. We employed three different methodologies: two machine learning-based (CRFs and SVMs) and one which combines a machine learning-based (CRFs) with a rule-based technique. The latter achieved better results with an overall F_1 score of about 44%. This figure doesn't seem encouraging: the comparison with the other systems that face the same problem with the same corpus within this competition probably will allow to understand this result and realize the weakness of our approaches.

References

1. http://labda.inf.uc3m.es/ddiextraction2011/dataset.html
2. Bordes, A., Usunier, N., Bottou, L.: Sequence labelling SVMs trained in one pass. In: ECML PKDD 2008. pp. 146–161. Springer (2008)

3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: Gate: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02) (2002)
4. Evans, D.A., Brownlowt, N.D., Hersh, W.R., Campbell, E.M.: Automating concept identification in the electronic medical record: An experiment in extracting dosage information. In: Proc. AMIA Annu Fall Symp. pp. 388–392 (1996)
5. Gold, S., Elhadad, N., Zhu, X., Cimino, J.J., Hripcsak, G.: Extracting structured medication event information from discharge summaries. In: Proc. AMIA Annu Symp. pp. 237–241 (2008)
6. HeppleIn, M.: Independence and commitment: Assumptions for rapid training and execution of rule-based pos taggers. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000) (2000)
7. Institute of Medicine (ed.): Preventing Medication Errors. The National Academics Press, Washington (2007)
8. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A., Wishart, D.: Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res (2011)
9. Levin, M.A., Krol, M., Doshi, A.M., Reich, D.L.: Extraction and mapping of drug names from free text to a standardized nomenclature. In: Proc AMIA Annu Symp. pp. 438–442 (2007)
10. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. Methods Inf Med (1993)
11. Pereira, S., Plaisantin, B., Korchia, M., Rozanes, N., Serrot, E., Joubert, M., Darmoni, S.J.: Automatic construction of dictionaries, application to product characteristics indexing. In: Proc Workshop on Advances in Bio Text Mining (2010)
12. Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C.: Extracting drug-drug interactions from biomedical texts. In: Workshop on Advances in Bio Text Mining (2010)
13. Segura-Bedmar, I., Martinez, P., de Pablo-Sanchez, C.: Using a shallow linguistic kernel for drug-drug interaction extraction. Journal of Biomedical Informatics, In Press (2011)
14. Shah, A.D., Martinez, C.: An algorithm to derive a numerical daily dose from unstructured text dosage instructions. Pharmacoepidemiology and Drug Safety 15, 161–166 (2006)
15. Sirohi, E., Peissig, P.: Study of effect of drug lexicons on medication extraction from electronic medical records. In: Proc. Pacific Symposium on Biocomputing. vol. 10, pp. 308–318 (2005)
16. Sutton, C.: Grmm: Graphical models in mallet. http://mallet.cs.umass.edu/grmm/ (2006)
17. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, 1453–1484 (2005)
18. Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., Denny, J.C.: Medex: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association 17, 19–24 (2010)