AN EXPERIMENTAL EXPLORATION OF DRUG-DRUG INTERACTION EXTRACTION FROM BIOMEDICAL TEXTS

Man Lan, Jiang Zhao, Kezun Zhang, Honglei Shi, and Jingli Cai

East China Normal University, Shanghai, P.R.China

Abstract. The First Challenge of Drug-Drug Interaction Extraction (DDIExtraction 2011) involves doing a binary DDI detection to determine whether a drug pair in a given sentence (with annotated drug names) has interaction information. This may be the first attempt at extraction of drug interaction information in wide community. In this paper we compare and evaluate the effectiveness of different strategies of example generation from texts and different feature types for drug relation extraction. The comparative results show that (1) drug interaction classification at drug entity pair level performs better than that at sentence level; (2) simple NLP output does not improve performance and more advanced way of incorporating NLP output need to be explored.

1 Introduction

In pharmacology domain, one drug may influence the level or activity of another drug if there is a drug-drug interaction (DDI) between them. Typically, the detection of DDIs between drug pair is an important research area for health care professionals to find dangerous drug interactions and possible side effects, which helps to decrease health care costs.

Like other entity (e.g., gene or protein) relation extraction tasks (i.e., BioCreAtIvE) from biomedical literature, information extraction (IE) techniques can provide an interesting way of reducing the time spent by health care professionals on reviewing the literature. Recently, DDIExtraction Challenge 2011 has played a key role in comparing various IE techniques applied to the pharmacological domain by providing a common benchmark for evaluating these techniques. Specifically, they create the first annotated Drug DDI corpus that studies the phenomena of interactions among drugs. Meanwhile, the organizers have devoted to several comparative experimental assessments of different exploration strategies on this corpus, e.g., Segura-Bedmar et al. (2010a), (2010b), (2011a) and (2011b). For example, they manually created linguistic rules (i.e. pattern) using shallow parsing and syntactic and lexical information with the aid of domain expert in Segura-Bedmar et al. (2010a) and (2011b). Moreover, they adopted shallow linguistic kernel-based supervised machine learning (SVM) method to build relation classifier for DDI extraction. Their experimental results showed

Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), pages 75–82
Huelva, Spain, September 2011.

that the sequence kernel-based method performs significantly better than the construction of linguistic rules.

The basic idea of our system is to make use of feature-based supervised machine learning approach for DDI extraction. Our work consists of two explorations, i.e., comparison of different strategies of example generation from texts and comparison of different feature types. The purpose of this work is twofold: (1) compares the performance of different strategies of example generation, different feature types for drug interaction extraction; (2) provides an overview of our practical and effective process for this challenge.

The rest of the paper is structured as follows. Section 2 describes the overview of DDIExtraction Challenge 2011. Section 3 presents the methods adopted in our participation. Section 4 describes the system configurations and results on the test data. Finally, Section 5 summarizes the concluding remarks and suggests the future work.

2 Overview of DDIExtraction Challenge 2011

In recent years, most biomedical relation extraction study and corpora have focused on describing genetic or protein entity interactions, e.g., BioInfer (2007), BioCreative II (2008) and II.5 (2009), or AIMed (2005), rather than drug-drug interaction. The First Challenge of Drug-Drug Interaction Extraction (i.e., DDIExtraction Challenge 2011) provides a new standard benchmark and creates the first annotated corpus for drug interaction extraction to a wider community. The DDI corpus is created by Segura-Bedmar et al.(2011a). The Drug DDI corpus consists of 579 documents describing DDI, which are randomly selected from the DrugBank database (2008). In DDIExtraction Challenge 2011, this corpus is split into 435 training documents (4267 sentences) and 144 test documents (1539 sentences) for evaluation. Table 1 lists the detailed various statistical information of training and test data set. From this table, we can see that the data distribution in training data set is quite close to that in test data set.

This corpus is provided in two different formats: (1) the unified XML format as the PPI Extraction format proposed in Pyysalo et al. (2008) and (2) a Metamap format based on the information provided by the UMLS MetaMap Transfer (MMTx) tool (2001). In MMTx format, the documents were analyzed by the MMTx tool that performs sentence splitting, tokenization, POS-tagging, shallow syntactic parsing, and linking of phrases with Unified Medical Language System (UMLS) Metathesaurus concepts. Besides, the MMTx format documents annotate a variety of biomedical entities occurring in texts according to the UMLS semantic types. An experienced pharmacist recommended the inclusion of the following UMLS semantic types as possible types of interacting drugs: (1) Clinical Drug (clnd), (2) Pharmacological Substance (phsu), (3) Antibiotic (antb), (4) Biologically Active Substance (bacs), (5) Chemical Viewed Structurally (chvs) and (6) Amino Acid, Peptide, or Protein (aapp).

Clearly, the MMTx format contains not only shallow NLP information but also domain-specific annotations. Therefore it is expected to provide more useful

Table 1. Statistical information of training and test data set.

Category	Training set	Test set
#-Documents	435	144
#-Sentences	4267	1539
#-Drug entities	11260	3689
#-Drug pairs	23827	7026
#-DDIs	2402	755
#-Documents containing, at least, one drug pair	399	134
#-Sentences with, at least, one drug pair	2812	965
#-Sentences with, at least, one DDI	1530	503
#-Total entities that participate in a pair	10374	3398
Avg drug per doc (considering only docs with drug pairs)	26.02	25.36
Avg drug per sentence (considering only sentences with drug pairs)	3.69	3.52
Avg DDI per doc (considering only docs with drug pairs)	6.02	5.63
Avg DDI per sentence (considering only sentences with drug pairs)	0.85	0.80

information than unified XML format for DDI extraction. Consequently, participants are required to indicate the document format their methods involved. Another thing need to note is that this challenge only considers the interactions between drugs within the same sentence.

Participants are allowed to submit a maximum of 5 runs. For each drug pair within one sentence, the participated algorithm is expected to generate label "0" for non-authentic DDI and label "1" for predicted DDI. For performance evaluation, this challenge adopted the most widely-used text classification evaluation measures, i.e., precision (P), recall (R) and their combination F_1 score.

3 Methods

In our work we cast drug relation extraction as a classification problem, in which each example is generated from texts and formed as a feature vector for classification. Specifically, we generate examples from all sentences containing at least two drug entities. That is, the sentences which have none or only one drug should be removed first before they come into the pipeline of text processing.

Here we need to take into account the following special considerations. One is the issue of example generation from texts. Another is the issue of feature types extracted from texts. Next we will discuss these two special considerations.

3.1 Example Generation

The training and test examples from texts can be generated at different levels, e.g., sentence level or drug pair level.

At sentence level, each example corresponds to one sentence. That is, each sentence is represented as a feature vector, no matter how many DDIs this sentence has. Typically, a sentence having n drugs ($n \geq 2$) generates C_n^2 drug

pairs but not all drug pairs are DDIs. Thus, in order to assign the DDI label to each sentence, we have the following two assumptions and they serve as baselines in our work.

Assumption 1: In training step, if there is at least one DDI annotated in the sentence, we assign the DDI label of this sentence 1. That is, this sentence is assumed to be a DDI sentence. In test step, if one sentence is predicted by classifier to be a DDI sentence, then all drug pairs within this sentence are predicted to be DDIs as well.

Assumption 2: In training step, if the number of DDIs is equal to or larger than the number of non-DDIs in the sentence, we label this sentence as DDI sentence. That is, for a sentence having n drugs, if it has at least $C_n^2/2$ DDIs, it is regarded as DDI sentence. In test step, if one sentence is predicted by classifier to be a DDI sentence, all drug pairs within this sentence are predicted to be DDIs as well.

Clearly, the built-in flaw of the above two assumptions is that they consider all drug-pairs in one sentence have one common taxonomy label. This is not true in real world case. We use the two assumptions as baseline systems in our work.

At drug pair level, each example corresponds to each drug pair in a sentence. That is, the number of examples generated for each sentence is given by the combinations of distinct drug entities (n) selected two at a time, i.e. C_n^2. For example, if one sentence contains three drug entities, the total number of examples generated from this sentence is $C_3^2 = 3$. In training step, for each example, we use its annotated DDI label as the label of this example. If a DDI relation holds between a drug pair, the example is labeled 1; otherwise 0. In test step, for each drug pair, the classification system predicts its DDI label based on the classifier constructed on training examples.

3.2 Features Extraction

No matter which level examples are generated from texts, the examples are represented as feature vectors for classifier construction and prediction. Here we describe the feature sets adopted by above two example generation approaches.

As for sentence level feature representation, we adopt a feature set consisting of all words in texts. Specifically, we remove stop words (504 stop words), punctuation, special characters and numbers from sentences.

As for drug pair level feature representation, instead of using all words in texts, we explore different feature types, i.e., lexical, morpho-syntactic, semantic and heuristic features (from annotated biomedical information), with the purpose of capturing information between drug pairs. The features consist of the following 6 types. The first two feature types are generated from unified XML text format. The following four feature types are obtained from MMTx text format.

F1: Token between drug pair. This feature includes the tokens (words) between two target drug entities. Given two annotated target drug entities, first all the words between them are extracted and then the Porter's stemming (1980) is performed to reduce words to their base forms.

F2: Lemma of target entities. This feature consists of the lemma of the target drug entities annotated in the given sentence. That is, this feature records the words of the target drug names.

F3: UMLS semantic types of target entities. This feature is to record the six UMLS semantic types of the drug entities annotated in the given sentence.

F4: Information of other drug entities. This feature is to indicate whether there is other drugs between the current target drug pair and the number of other drug entities.

F5: Relative position between verbs and target drug entities. This feature is to record if there is verb before, between or after the target drug pair.

Except for the above two approaches, we also explore experiment using only the position information of verbs and target drug entities as follows.

F6: Position of verbs and target drug entities. This feature is different from above 5 feature types, which only records the position information of verbs and drug entities. To do so, for the first drug entity, we record the relative positions of three closest verbs before it and after it. For example, if the position of the two verbs offset is 10 and 11, and the position of the first drug is 15, the first three feature values is 5, 4 (relative position) and 0 (since no third verb before the first drug). For the second drug entity, we record the relative positions of three closest verbs after it. In addition, we also assign one label for each verb to record if there is a negation before it, yes for 1 and no for 0. We manually created list of 16 negation words including: *little*, *few*, *hardly*, *never*, *none*, *neither*, *seldom*, *scarcely*, *rarely*, *cannot*, *can't*, *isn't*, *hasn't*, *couldn't*, *unlike*, *without*.

3.3 Learning Algorithms

Generally, according to the different kernel functions from computational learning theory, SVMs are classified into two categories, i.e., linear and nonlinear (such as polynomial, radial-based function (RBF), etc). Specifically, in this study, we adopt the radial-based nonlinear SVM because in our preliminary study the nonlinear SVM performs better than linear SVM models. The SVM software we used in all experiments is LIBSVM-2.9 (2001).

4 Results And Discussion

4.1 Text Preprocessing

In text processing step, the stop words (504 stop words), punctuation and numbers were removed. The Porter's stemming (1980) was performed to reduce words to their base forms. The resulting vocabulary has 3715 words (terms).

4.2 System configuration and Results

In this work, we config five different classification systems with different example generation strategies and different feature types. The classifiers for all systems

were optimized independently in a number of 5-fold cross-validation (CV) experiments on the provided training sets. First we consider two baseline systems at sentence level described in section 3.1. We create a global feature set consisting of all words in texts. The resulting vocabulary of the two systems has 3715 and 3224 words (terms) respectively. Table 2 shows the results of the first two systems at sentence level.

Table 2. Two system configurations at sentence level with two assumptions and results on the test data.

System	Description (sentence level)	P (%)	R (%)	F_1(%)
1	assumption 1, all words in texts	14.37	76.82	24.21
2	assumption 2, all words in texts	39.63	16.95	23.75

In the third system, we conducted several comparative experiments at drug pair level using different combination of features described in section 3.2. In addition, in the fourth system, we evaluated the system with only relative position information between drugs and verbs in one sentence. Finally, in the fifth system, we performed majority voting to combine the best results of the first four systems to further improve performance. Table 3 shows the results of these three systems at drug pair level using different feature sets.

Table 3. System configurations at drug pair level with different feature types and results on the test data.

System	Description (drug pair level)	P (%)	R (%)	F_1(%)
3	F1	31.49	68.48	43.14
	F1, F2	28.08	42.91	33.94
	F1, F2, F3	32.70	31.92	32.31
	F1, F2, F3, F4	37.96	31.13	34.21
	F1, F2, F3, F4, F5	41.71	35.63	38.43
4	F6	32.70	27.28	29.75
5	Majority voting	29.57	46.49	36.15

4.3 Discussion

Based on the above series of experiments and results shown in Table 2 and Table 3, some interesting observations can be found as follows.

Specifically, the first two baseline systems at sentence level yield quite similar F-measures of 24.21 and 23.75 but different recall and precision. The first system has high recall but low precision. Conversely, the second system has high

precision but quite low recall. This difference comes from the different principle of the two assumptions. This F-measure is similar to the result reported in Segura-Bedmar et al. (2011b) using only linguistical patterns with the aid of domain expert.

Generally, the systems at drug pair level (Table 3) perform better than those at sentence level (Table 2). This result is consistent with our preliminary surmise that it is too rough for example generation at sentence level and it did not take the relation between drug pair into consideration. Certainly many previous work on entity relation extraction generated example using this representation.

Moreover, the comparative result of the third serial of systems, i.e., the systems at drug pair level with different feature sets, is beyond our preliminary expectation. Surprisingly, the system with only words between two drug entities performs the best among the serial of the third systems. Although we extracted and constructed more features which are supposed to hold more useful information, such as drug names, drug types and the position information between drug and verb, these features did not improve the performance. One possible explanation is that the number of F1 feature is much larger than other features, and thus F1 feature dominates the performance of classifier. Another possible reason is that these manually constructed or NLP features may not be appropriate for representation and thus more advanced NLP techniques and advanced ways of incorporating NLP output is necessary for future exploration.

Another surprise is that the fourth system performs better than the two baseline systems at sentence level but still worse than the third system. Since the fourth system only considers relative position information rather than words and other features, this result is quite interesting. However, we do not expect more improvement on this simple feature set and we have no further explorations.

As an ensemble system, the fifth system combines the best results of the previous four systems. However, this majority voting strategy has not shown significant improvements. The possible reason may be that these classifiers come from a family of SVM classifiers and thus the random errors are not significantly different.

5 Summary

Based on the comparative experimental results, we summarized that, first, example generated at drug pair level performs better than sentence level; second, using only words between drug pair entities performs better than adding more constructed NLP and domain-specific features. It indicates that NLP output has not yet succeeded in improving classification performance over the simple bag-of-words approach and more advanced way of incorporating NLP output need to be explored.

We have to mention that although the best performance on the test set yields a final score of no more than 45% (F-measure), which is quite lower than the best performance 60.01% reported in Segura-Bedmar et al. (2011a), it is still quite promising since we do not involve domain expert, domain knowledge and

complicated NLP outputs neither. In other words, this suggests that there may be ample room for improving the performance.

ACKNOWLEDGMENTS

This research is supported by grants from National Natural Science Foundation of China (No.60903093), Shanghai Pujiang Talent Program (No.09PJ1404500) and Doctoral Fund of Ministry of Education of China (No.20090076120029).

References

Pyysalo, F., Airola, A., Heimonen, J., Bjorne, J., Ginter, F., and Salakoski, T. Comparative analysis of protein-protein interaction corpora. BMC Bioinformatics 9 (Suppl 3) : S6 (2008).

Isabel Segura-Bedmar, Paloma Martinez, Cesar de Pablo-Sanchez. Extracting drug-drug interactions from biomedical texts. BMC Bioinformatics 2010, 11(Suppl 5):P9.

Isabel Segura-Bedmar, Paloma Martinez, Cesar de Pablo-Sanchez. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics 2011, 12(Suppl 2):S1.

Giuliano C, Lavelli A, Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006) 2006, 5-7.

Isabel Segura-Bedmar, Paloma Martinez, Cesar de Pablo-Sanchez. Using a shallow linguistic kernel for drug-drug interaction extraction. Journal of Biomedical Informatics, April 2011.

Isabel Segura-Bedmar, Paloma Martinez, Cesar de Pablo-Sanchez. Combining syntactic information and domain-specific lexical patterns to extract drugCdrug interactions from biomedical texts. In: Proceedings of the ACM fourth international workshop on data and text mining in biomedical informatics (DTMBIO10); 2010. p. 49-56.

Sampo Pyysalo , Filip Ginter , Juho Heimonen , Jari Bjorne , Jorma Boberg , Jouni Jarvinen and Tapio Salakoski. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 2007, 8:50

Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biology 2008, 9(Suppl 2):S4.

Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, Wong YW. Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine 2005, 33(2):139-155.

Wishart D, Knox C, Guo A, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 2008, 36(Database issue):D901-6.

Aronson A: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium 2001, 17-22.

M. Porter. An algorithm for suffix stripping. Program, vol. 14, no. 3, pp.130-137, 1980.

C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.