# Enhancing Drug-Drug Interaction Classification with Corpus-level Feature and Classifier Ensemble

Jing-Cyun Tu
Yuan-Ze University, Taoyuan, Taiwan
st410634@gmail.com

Po-Ting Lai
National Tsing-Hua University, HsinChu, Taiwan
potinglai@gmail.com

Richard Tzong-Han Tsai*
National Central University, Taoyuan, Taiwan
thtsai@csie.ncu.edu.tw

**Abstract**

The study of drug-drug interaction (DDI) is important in the drug discovering. Both PubMed and DrugBank are rich resources to retrieve DDI information which is usually represented in plain text. Automatically extracting DDI pairs from text improves the quality of drug discovering. In this paper, we presented a study that focuses on the DDI classification. We normalized the drug names, and developed both sentence-level and corpus-level features for DDI classification. A classifier ensemble approach is used for the unbalance DDI labels problem. Our approach achieved an F-score of 65.4% on SemEval 2013 DDI test set. The experimental results also show the effects of proposed corpus-level features in the DDI task.

## 1 Introduction

Drug-drug interaction (DDI) is a situation that a drug modifies the effect of another drug, and the modified effect may be increased, decreased or new. For examples, if a patient takes two drugs and one increases the effect of another, an overdose may occur. In contrary, an under dosage may occur if the effect is decreased. Furthermore, the DDI may also cause the side effects. Therefore, the survey of DDI studies is important for improving the quality of drug discovering. Many drug-drug interactions are publicly available through PubMed or DrugBank (Law, et al., 2014). However, only a fraction of them is in a structured format such as DrugBank (Law, et al., 2014). Most DDIs are represented in unstructured plane text. Therefore, automatically extracting DDI form these texts is an important issue.

In 2013, SemEval (Segura-Bedmar, et al., 2013) sets this task as the one of its shared task challenge. Extracting DDI consists of two tasks: (1) drug name recognition (DNR) and (2) DDI classification. The named entity recognition (NER) is usually formulated as the sequence label problem and resolved by the Conditional Random Fields model (Campos, et al., 2013; Leaman, et al., 2015). The most important thing is to design and select proper features which capture the boundaries of the named entities. For DNR, several approaches (Björne, et al., 2013; Liu, et al., 2015; Rocktäschel, et al., 2013) have been proposed. For instance, Liu et.al (Liu, et al., 2015) considered selecting DNR features as a feature engineering problem, and their experiments combined several features. Their approach achieved an F-score of 79.36% in the DDIExtraction 2013 dataset (Segura-Bedmar, et al., 2013).

The second task is to classify the drug-drug pair in the sentence into one of advice, effect, mechanism, int (interaction) or negative labels. The advice label indicates that the drug-drug pair is recommended or advised to have the interaction. The effect label indicates that the DDI effect is described in the sentence. The mechanism label indicates that the DDI is described about Pharmacology which includes both pharmacodynamics and pharmacokinetics. The int label indicates that the physical interaction is stated without any other information. The negative label indicates that there is no interaction. For the second task, several Machine Learning (ML)-based approaches have been proposed. For an example, WBI-DDI (Thomas, et al., 2013) proposed a two-step strategy. They detect general drug-drug interactions regardless of subtype using the different machine-learning methods, and proposed an ensemble voting approach to ensemble these methods. Their approach achieved an F-score of

* Corresponding author



*Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 52–56,
Taipei, Taiwan, November 27, 2017. 

60.9%. The ML-based approaches usually suffered from bias number of the positive and negative DDI pairs. Only few of them have true interactions. To resolve this problem, FBK-irst (Chowdhury and Lavelli, 2013) proposed a multi-phase kernel-based approach. They consider the negative DDI sentence and pair as less informative sentence (LLS) and less informative instance (LLI). They design the rules and classifier to discard LLS and LLI. They proposed a hybrid kernel approach and several syntactic dependent features for DDI detection and classification. Their approach achieves an F-score of 65% on SemEval DDI 2013 test set (Segura-Bedmar, et al., 2013).

Despise many DDI classification approaches had been proposed, the approaches for DDI still can be further improved. For an example, since a drug can be represented in its branded name or generic name, the interaction describe in other sentence could not be directly used in other sentences without linking its different names. Therefore, the features used in current approaches usually focus on sentence-level and the information above the target DDI sentences, such as corpus-level features, were not referred. Iyer et.al. (Iyer, et al., 2013) proposed an annotation-based approach to learn DDI from the electronic medical records (EMR). Since the EMR has rich temporal information such as section times, they annotate temporal relationship between the drugs and event and calculate the odds ratio (OR) of the drug-drug pair with events to only one drug with the event. Their experimental results show the effect of odds ratio in learning DDI pairs.

In this paper, we present a study that focuses on the second task. First, we proposed a ML-based approach which includes the basic words, Part-of-speech, syntactic and template features as our baseline. Second, we deal with the drug various names and proposed the corpus-level features by calculating the odds ratios of the drugs matched our automatically generated DDI template which is inspired by Iyer et.al.'s approach (Iyer, et al., 2013). Third, to tackle the bias labels in DDI corpus, we used a classifier ensemble approach with voting strategy. The experiments are in the SemEval DDI 2013 dataset (Segura-Bedmar, et al., 2013). Our proposed approach achieved an F-score of 65.4%, which outperforms both WBI-DDI and FBK-irst's approaches.

## 2 Method

The proposed DDI classification approach consists of four main steps. The first is *Drug Name Normalization*, which used RxNorm (Nelson, et al., 2011) to normalize drug synonyms in order calculate odds ratio more accurately. Following is the *Odds Ratio* step, in which we calculate the odds ratio of drug-drug pair matched DDI templates to only one drug matched. Next, *Features for Classification* presents the DDI classification features. Lastly, the *Classifier Ensemble* divides positive and negative training data into equal size, and training five classifiers in the different sets, then used a voting strategy to ensemble classification results.

### 2.1 Drug Name Normalization

A drug might be represented as its generic name or branded name in the text. Here we refer them as drug synonyms. To normalize the synonyms can make the calculation of odds ratio more accurately. RxNorm is a tool developed by National Library of Medicine (NIH). It contains the normalized drug names and links them to many drug vocabularies which are commonly used in pharmacy management and drug interaction software. RxNorm can links the drug names between different systems which do not use the same software and vocabulary. Before calculating the odds ratio, we will use RxNorm to normalize the drug name $d$ into its generic name $g$. If the $d$ cannot be normalized to any generic name, then we will use $d$ as its normalized name.

### 2.2 Odds Ratio

The odds is the ratio $r$ of the probability $p_1$ that the event of interest occurs to the probability $p_2$ that it does not occur. This is often estimated by the ratio of the number of times $t_1$ that the event of interest occurs to the number of times $t_2$ that it does not. In this paper, the odds ratio refers to the ratio *or* of the odds $r_1$ that the drug $d_1$ interacts with the drug $d_2$ to the odds $r_2$ that $d_1$ or $d_2$ interacts with the other drugs. For example, the odds $r_1$ that a drug $d_1$ interacts with the drug $d_2$ is 4 and the odds $r_2$ that $d_1$ or $d_2$ interacts with the other drugs is 2. The odds ratio *or* of $d_1$ and $d_2$ will be 4/2 = 2. The higher *or* indicates the higher odds that $d_1$ and $d_2$ have interaction than they interact with the other drugs. While calculating *or*, whether $d_1$ interacts with $d_2$ is obtained by the DDI templates which we will introduce in section 2.3.2.

### 2.3 Features for Classification

Our classifier uses basic, template and odds ratio

features. The basic and template features utilized the immediate context of the drugs pair as features, whereas odds ratio features used the corpus-level information.

### 2.3.1 Basic Features

The basic features comprised words, Part-of-speech (POS) and syntactic features. There are two sets of word features used in our system, each with a different feature label. Inter-Drugs *n*-grams set includes all word unigrams and bigrams located between drugs. If none is present, the feature is given a *"NULL"* value. Surrounding Words set includes the two words before the first drug and the two after the second drug. If there are no words before or after both NEs, a *"NULL"* value is set. All words are treated as bag-of-words. That is, the order of these words is not considered. Similarly, the unigrams of POS tags between drugs are also used as POS features. We also parse each sentence with a full-sentence syntactic parser (Roark, et al., 2006) to generate its full parse tree. We use the syntactic path through the parse tree from the drug $d_1$ to the drug $d_2$ as a feature.

### 2.3.2 Template Features

Our template generation (TG) algorithm, which extracts word patterns for drugs pairs using Smith and Waterman's local alignment algorithm (Smith and Waterman, 1981). Firstly, we pair all sentences containing positive relations. The sentence pairs are then aligned word-by-word and a pattern satisfying the alignment result is created. Each slot in the template is given by the corresponding constraint information expressed in the form of a word (e.g. *"associated"*). If two aligned sentences have nothing in common for a given slot, the TG algorithm puts a wildcard in the position. The complete TG algorithm is described with pseudo code in the Algorithm. The similarity function used to compare the similarity of two tokens in local algorithm is defined as:

$$Sim(x, y) = \max \begin{cases} \mathbf{1,\ if}\ x = y \\ \mathbf{0,\ otherwise} \end{cases}$$

where $x$ and $y$ are tokens in sentences $s_i$ and $s_j$, respectively. The similarity of two sentences is calculated by the local algorithm on the basis of this token-level similarity function.

**Template Generation Algorithm**

**INPUT:** A set of sentences $S = \{s_1, \ldots, s_k\}$
1 : $T = \{\}$;
2 : **for** $s_i$ in $s_1$ to $s_{k-1}$
3 : **for** $s_j$ in $s_{i+1}$ to $s_k$
4 : **if** the similarity of $s_i$ and $s_j$ above the threshold
5 : **then** generate template $t$ from $s_i$ and $s_j$
6 : $T \leftarrow t$ ;
7 : **end**;
8 : **end**;
9 : **return** $T$
**OUTPUT:** A set of templates $T = \{t_1, \ldots, t_k\}$

### 2.3.3 Odds Ratio Features

The odds ratio is the ratio of one odds to another, and it is larger than zero. In our experiment, we use different thresholds as odd ratio features include 1.0, 1.5, 2.0 and 2.5. The real number of odds ratio is also used as one of the odds ratio features.

## 2.4 Classifier Ensemble

The amount of negative DDI pairs is higher than the positive ones in both DDI corpus and real world. The support vector machine model (Chang and Lin, 2011) used in our experiment is suffered from this problem. To tackle this problem, we proposed a classifier ensemble approach to training our classifiers. Firstly, we randomly divide the negative data into five unique subsets, since the ratio of the positive pairs to the negative pairs is approximate 5 in the experimental training corpus. Secondly, we construct five training datasets that each contains all positive data and one negative subset. Thirdly, we train five base classifiers with SVM. Here we use the Gaussian kernel. Once the classifiers are constructed, new DDI pairs are classified by the classifiers, and their results are aggregated to form the final ensemble decision output. The vote method is used in this paper. Given classifiers $C_i$, $i$ = 1, 2, ... , $N_C$, and DDI labels $L_j$ , $j$ = 1, 2, ... , $N_L$, where $N_C$ is the ensemble size and $N_L$ is the number of DDI labels. The final aggregated decision is the winning classifier that has the highest votes across all classifiers. If any tie situation existed, the label with the highest predicted value will be assigned.

# 3 Experiments

To evaluate our approach, the SemEval 2013 DDI corpus is used. Table 1 shows the number of the DDI categories annotated in the corpus. The most common type was negative pairs in both

training and test set. Here, we first compare the performance achieved by baseline features (basic + template features) to the baseline + odds ratio (OR) features. In Table 2, we can see that OR features improve the baseline's performance by an F-score 11.9%. Our approach performs better than FBK-irst and WBI-DDI, because our OR features are effective. In Table 3, we list the F-score for each category of DDI. We observe that the F-scores for advice and mechanism are comparatively high. This is possibly because they have some specific keywords in both categories. However, although effect is the second most frequent category, it does not have a high F-score. We think this discrepancy is due to the fact that the descriptions of DDI effects are commonly presented more flexible. Int's performance is comparatively higher than the other two systems since the OR features are effective while the training set is very small.

| Type | | Training set | Test set |
|---|---|---|---|
| **#Documents** | | 456 | 116 |
| **#Sentence** | | 2915 | 341 |
| **#Positive pairs** | **Advice** | 658 | 160 |
| | **Effect** | 1243 | 292 |
| | **Mechanism** | 1004 | 253 |
| | **Int** | 168 | 10 |
| | **Total** | 3073 | 715 |
| **#Negative pairs** | | 17905 | 4312 |
| **#Total pair** | | 20978 | 5027 |

Table 1: The statistic of DDI dataset

| Configuration | P(%) | R(%) | F(%) |
|---|---|---|---|
| **Baseline** | 50.4 | 57.0 | 53.4 |
| **WBI-DDI** | 64.2 | 57.9 | 60.9 |
| **FBK-irst** | 65.0 | 66.0 | 65.0 |
| **Baseline + OR** | 64.0 | 66.7 | **65.3** |

Table 2: The DDI classification performances on the test set

| Category | Baseline + OR | FBK-irst | WBI-DDI |
|---|---|---|---|
| **Advice** | **69.5** | 69.2 | 63.2 |
| **Effect** | 62.3 | **62.8** | 61.0 |
| **Mechanism** | **68.2** | 67.9 | 61.8 |
| **INT** | **60.0** | 54.7 | 51.0 |
| **Overall** | **65.3** | 65.0 | 60.9 |

Table 3. The F-scores of individual DDI categories on the test set

## 4 Conclusion

In this paper, we present a classifier ensemble approach for drug-drug interaction classification. We developed the sentence-level features for the classification. To encode corpus-level odds ratio features, we used the RxNorm to normalize the drug names. Our ensemble classifier achieves an F-score of 65.4% on SemEval 2013 DDI test set. The results underscore the effect of corpus-level features in classifying the drug-drug interaction.

## Reference


Björne, J., Kaewphan, S. and Salakoski, T. (2013) UTurku: Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pp. 651-659.

Campos, D., Matos, S. and Oliveira, J.L. (2013) Gimli: open source and high-performance biomedical name recognition, *BMC Bioinformatics*, 14, 54.

Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, 2, 1-27.

Chowdhury, M.M.F. and Lavelli, A. (2013) FBK-irst : A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pp. 351-355.

Iyer, S.V*., et al.* (2013) Learning Signals of Adverse Drug-Drug Interactions from the Unstructured Text of Electronic Health Records, *AMIA Summits on Translational Science Proceedings*, 2013, 98-98.

Law, V*., et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism, *Nucleic*

*Acids Research*, 42, D1091-D1097.

Leaman, R., Wei, C.-H. and Lu, Z. (2015) tmChem: a high performance approach for chemical named entity recognition and normalization, *Journal of Cheminformatics*, 7, S3.

Liu, S.*, et al.* (2015) Feature Engineering for Drug Name Recognition in Biomedical Texts: Feature Conjunction and Feature Selection, *Computational and Mathematical Methods in Medicine*, 2015, 9.

Nelson, S.J.*, et al.* (2011) Normalized names for clinical drugs: RxNorm at 6 years, *Journal of the American Medical Informatics Association : JAMIA*, 18, 441-448.

Roark, B.*, et al.* (2006) SParseval: Evaluation metrics for parsing speech, *Proc. LREC*.

Rocktäschel, T.*, et al.* (2013) WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pp. 356-363.

Segura-Bedmar, I., Martínez, P. and Herrero Zazo, M. (2013) SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pp. 341-350.

Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences, *Journal of Molecular Biology*, 147, 195-197.

Thomas, P.*, et al.* (2013) WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pp. 628-635.