Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach

Sun Kim [1], Haibin Liu *,[1], Lana Yeganova [1], W. John Wilbur

National Center for Biotechnology Information (NCBI), Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 21 October 2014
Revised 21 January 2015
Accepted 11 March 2015
Available online 19 March 2015

Keywords:
Drug–drug interaction
Biomedical literature
Linear kernel approach

ABSTRACT

Identifying unknown drug interactions is of great benefit in the early detection of adverse drug reactions. Despite existence of several resources for drug–drug interaction (DDI) information, the wealth of such information is buried in a body of unstructured medical text which is growing exponentially. This calls for developing text mining techniques for identifying DDIs. The state-of-the-art DDI extraction methods use Support Vector Machines (SVMs) with non-linear composite kernels to explore diverse contexts in literature. While computationally less expensive, linear kernel-based systems have not achieved a comparable performance in DDI extraction tasks. In this work, we propose an efficient and scalable system using a linear kernel to identify DDI information. The proposed approach consists of two steps: identifying DDIs and assigning one of four different DDI types to the predicted drug pairs. We demonstrate that when equipped with a rich set of lexical and syntactic features, a linear SVM classifier is able to achieve a competitive performance in detecting DDIs. In addition, the *one-against-one* strategy proves vital for addressing an imbalance issue in DDI type classification. Applied to the DDIExtraction 2013 corpus, our system achieves an *F*1 score of 0.670, as compared to 0.651 and 0.609 reported by the top two participating teams in the DDIExtraction 2013 challenge, both based on non-linear kernel methods.

1. Introduction

New drugs are generally studied on relatively small and homogeneous patient populations. As a result, pharmaceuticals often have side effects that remain unnoticed until they are already available to the public. This is especially true of side effects that emerge when two drugs are co-administered. A change in the effect of one drug in the presence of another drug is known as a drug–drug interaction (DDI) [1]. It is characterized as an increase or decrease in the action of either substance, or it may be an adverse effect that is not normally associated with either drug. Understanding these drug–drug interactions and their downstream effects is of significant importance, leading to reduced number of drug-safety incidents and reduced healthcare costs.

To address the DDI problem, a number of drug databases such as DrugBank [2] and Stockley's Drug Interactions [1] have been created. Yet, they cover only a fraction of knowledge available. A large amount of up-to-date information is still hidden in the text of journal articles, technical reports and adverse event reporting systems, and this body of unstructured published literature is growing rapidly. MEDLINE®, for example, has doubled in size within the last decade and currently contains about 23 million documents. This creates an urgent need for text mining techniques to extract DDI information.

Using text mining techniques for DDI extraction has received less attention compared to other biomedical relation extraction tasks (e.g., protein–protein interactions), possibly due to the lack of gold standard sets [3–6]. The DDIExtraction challenges are the first community-wide competition addressing the DDI extraction problem [7,8] and a series of studies have been reported at the 2011 and 2013 challenge workshops [9–11].

Top performing systems in the DDIExtraction challenges use Support Vector Machines (SVMs) with non-linear kernels [12,13]. To handle structural representations of input instances, such as dependency graphs, non-linear kernels directly calculate similarities between two graphs by comparing embedded subgraphs [14]. While non-linear kernels are theoretically capable of implicitly searching a high dimensional feature space of subgraphs, existing methods generally exploit only a partial feature space because of the exponential number of subgraphs [15]. In addition, non-linear kernels are frequently combined into composite kernels [12,11]. Composite kernels, however, incur more computational cost because the complexity of the underlying kernels accumulates

* Corresponding author.
E-mail addresses: sun.kim@nih.gov (S. Kim), haibin.liu@nih.gov (H. Liu), lana.yeganova@nih.gov (L. Yeganova), wilbur@ncbi.nlm.nih.gov (W.J. Wilbur).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2015.03.002
1532-0464/Published by Elsevier Inc.

and additional learning is required to optimize the weights for individual kernels.

Despite the popularity of non-linear kernel methods, linear kernels are a good alternative for relation extraction tasks [16–18]. Linear kernels with word-level features alone provide a strong baseline performance [11,12]. Moreover, they can explicitly include nodes, edges and path structures of the dependency graphs [17]. Also, the straightforward representation of linear kernels enables the intuitive interpretation of obtained results. Most importantly, when training large-scale datasets, it has been demonstrated that often linear kernels are the only practical choice [19,20]. However, the performance of linear kernel systems in DDI extraction tasks has a noticeable gap from that of the top systems using non-linear kernels [7,8,21].

We conjecture that linear kernel-based systems may benefit from a rich set of lexical and syntactic features. With the goal to build a simple and scalable system, we develop a DDI extraction system based on a single linear SVM classifier. We define five types of features to capture the complexity of data: word features with position information, pairs of non-adjacent words, dependency relations, parse tree structures and tags for differentiating DDI pairs within the same noun phrase. Unlike other state-of-the-art systems [13,21] which incorporate external, domain-specific resources, our features originate exclusively from training data.

We evaluate our system on the DDIExtraction 2013 corpus [22]. Consistent with other studies [11–13], we adopt a two-phase approach, where DDI pairs are identified first, and then classified into specific DDI types. The proposed method achieves an overall *F*-score of 67% which outperforms the best performing system by 1.9%. We believe that the strength of our method comes from using a diverse set of features. In addition, the *one-against-one* strategy [23] used in the DDI type classification contributes to the higher performance. As the first linear kernel method that achieves the state-of-the-art performance on both DDI detection and classification tasks, we consider it a strong alternative to the nonlinear, composite kernel-based approaches. The inherent simplicity of the method adds transparency to the overall system, which could be especially beneficial if the system is used as a part of a more complicated schema. The source code for generating the features proposed in this article is available at http://www.ncbi.nlm.nih.gov/IRET/DDI.

2. Methods

Fig. 1 illustrates the overall architecture of our DDI extraction system. A binary classifier is trained first to extract interacting drug pairs from all candidate interactions. A DDI type classifier is then built to classify the interacting pairs into predefined relation categories. Our approach focuses on interactions expressed within the boundaries of a single sentence, and also assumes that drug entities involved in the target interactions have been annotated.

In this section, we first elaborate the five types of features used, including two novel features proposed for the DDI problem: word pair and noun phrase-constrained coordination (NPC) features. Then, we briefly introduce the preprocessing steps completed on both training and test data. Next, we describe our linear SVM classifier with a modified Huber loss function [24]. In the end, we compare our method with existing DDI extraction systems.

2.1. Features

2.1.1. Word features

Word-level features such as individual words in a sentence and sequences of words have been demonstrated to provide a strong performance baseline in extracting relational knowledge [11,17,25]. Hence, in our system, we use *n*-gram features of size up to 3, i.e., unigrams, bigrams and trigrams. Including *n*-grams of larger size does not always lead to a performance increase due to the data sparseness problem [25]. Similar to the works of He et al. [11] and Giuliano et al. [26], the position information is appended to each word feature according to positions of words in a sentence relative to an investigated drug pair: *before* (BF), *between* (BE) and *after* (AF). For instance, "Interaction_BF of_BF **ketamine** and_BE **halothane** in_AF rats_AF" where "ketamine" and "halothane" are two drug names.

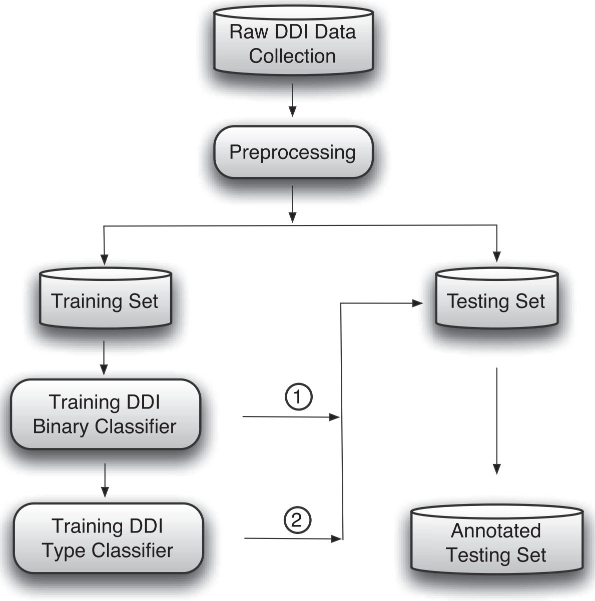

Fig. 1. Two-phase DDI extraction framework. DDI detection (①) decides whether a drug pair interacts. DDI type classification (②) assigns DDI types to interacting pairs.

2.1.2. Word pair features

While word features may capture repetitive expression patterns in neighboring words, they are not able to discover patterns involving distant words in a sentence. A simple solution to capture distant word patterns is to extract all possible word combinations from a training set. However, this approach increases the number of features considerably, and it also may degrade classification performance. To address this issue, we here propose a novel technique for selecting significant word pairs.

First, unigram word features are paired and only those pairs with a minimum frequency *k* are selected. Second, for selected word pairs, *p*-values are calculated using the hypergeometric distribution [27]. The *p*-value reflects how strongly a feature is represented in the positive set as compared to the negative set. It relates to the null hypothesis that the co-occurrence of two words is randomly distributed between positive and negative sets. If the co-occurrence is randomly distributed, the word pair will have a high *p*-value. If the *p*-value is low, this indicates a $1 - p$ probability that the co-occurrence is not random and is likely indicative for positive DDIs.

To obtain the most useful word pairs, we need the least restrictive frequency and the most restrictive *p*-value. In this work, we set $k = 200$ and *p*-value = 0.01 based on *F*1 scores via 10-fold document-level cross validation on the training set. This significant *p*-value helps select 588 word pairs from a total of 449,826 pairs with *k* above 200. This feature set contains certain informative word pairs such as "$drug_1 \cdots drug_2 \cdots$ **increase** $\cdots$ **level**" and