Extraction of Drug-Drug Interactions Using Convolutional Neural Networks

Puneet Souda*

Department of Pharmaceutical and Drug Discovery, Faculty of Science, King Mongkut's University of Technology Thonburi, Thailand

Corresponding author: Puneet Souda

✉ puneet.sou@kmutt.ac.th

Department of Pharmaceutical and Drug Discovery, Faculty of Science, King Mongkut's University of Technology Thonburi, Thailand

Abstract

Drug-drug interaction (DDI) extraction has long been a popular relation extraction task in natural language processing (NLP). Modern support vector machines (SVM) with a high number of manually set features are the foundation of most DDI extraction methods. Convolutional neural networks (CNN), a reliable machine learning technique that nearly never requires manually generated features, have recently shown significant promise for a variety of NLP tasks. CNN should be used for DDI extraction, which has never been looked at. A CNN-based technique for DDI extraction was put forth. CNN is a good option for DDI extraction, as shown by experiments done on the 2013 DDI Extraction challenge corpus. The CNN-based DDI extraction approach outperforms the currently highest performing method by 69.75%, achieving a score of 69.75%.

Keywords: Drug-drug interaction (DDI); Convolutional neural networks (CNN); Support vector machines (SVM); Extraction

Introduction

Drug-drug interactions (DDIs) happen when two or more medications are taken together; this changes how one or more medications behave in the body and may have unanticipated side effects. Unexpected side effects brought on by DDIs are almost usually extremely risky (and even fatal) and substantially raise medical expenses. Medical errors are reduced as healthcare practitioners become more knowledgeable about DDIs. Therefore, medication safety and healthcare management have traditionally given DDIs a lot of attention. Multiple publicly accessible databases help healthcare workers locate DDIs. As an illustration, the online drug database Drug Bank has 8311 drug listings. More than 200 fields, including a DDI field, are included in each drug entry [1]. The databases do contain certain restrictions, though. First off, the majority of DDI databases are dictionaries with a DDI field like "Drug Bank" specified in text. Healthcare practitioners cannot immediately access the DDIs in these databases like they do with relational databases. Second, new DDIs are frequently discovered by medical experts and reported in literature such as books, journals, and technical reports. Healthcare practitioners cannot manually sift through the voluminous literature to identify DDIs or stay current on the most recent DDI results. As a result, medical text mining is becoming more and more interested in DDI extraction, which finds DDIs in unstructured text and automatically groups them into specified categories.

A common relation extraction problem in natural language processing is DDI extraction (NLP). Rule-based and machine learning-based methods are two of the many DDI extraction techniques that have been developed. While machine learning-based methods handle DDI extraction as a conventional supervised learning issue over annotated corpora, rule-based methods use manually set rules to extract DDIs. Machine learning-based approaches typically outperform rule-based approaches in terms of performance and portability [2]. The majority of early DDI extraction approaches are rule-based because there aren't any annotated corpora. With the establishment of DDI Extraction challenges in 2011 and 2013, a set of domain-specific criteria were defined to extract DDIs in the Drug Bank. Machine learning-based approaches have since been developed for DDI extraction on the public corpora of the challenges. Drug-drug interactions can be extracted from biological publications using DDI Extraction 2011, as well as DDI Extraction 2013. The DDI Extraction 2011 corpus labels DDIs without type information, but the DDI Extraction 2013 challenge categorises DDIs into four types: "mechanism," "affect," "advice," and "int." Support vector machines (SVM) are used in the best-performing algorithms on these corpora, and

they have several manually designed features. For instance, the winning system for the 2013 DDI Extraction challenge is based on SVM and uses a hybrid kernel with characteristics like trigger words, dependency trees, and parse trees [3].

The sub sequential best system is based on a linear SVM with a variety of rich characteristics, such as word, word pair, and dependency graph, parse tree, and coordination features that are confined to noun phrases. The fussy feature engineering of these systems must be endured. The majority of the characteristics employed in these systems are typically produced using outdated and ineffective NLP toolkits. The DDI extraction systems necessarily experience errors brought on by NLP toolkits. Convolutional neural networks (CNN), a recently developed robust machine learning technique, have shown significant promise for a variety of NLP tasks, including sentiment analysis, semantic parsing, and search query retrieval. However, DDI extraction has never been carried out using it. In this paper, CNN is used to complete this assignment. Sentences with annotations for drugs are used as inputs in the CNN-based technique [4]. The look-up table layer, the convolutional layer, the max pooling layer, and the softmax layer are the four layers that make up the CNN-based approach. In the look-up table layer, given a sentence containing two drugs of interest, each word is represented by word embeddings and position embeddings, and the sentence is then represented by a matrix that concatenates the word embeddings and position embeddings of its words in the order of their occurrence. The phrase matrix is convolved with filters of various sizes in the convolutional layer to create a collection of feature vectors. The size of each feature vector depends on the context window being taken into consideration, and the number of feature vectors is equal to the number of filters. Each vector in the group is reduced to a feature in the max pooling layer, which creates a new vector from the group of vectors. The fully connected softmax layer is then fed the vector that was obtained in the max pooling layer for classification. The "Order" approach was employed in 2013 to initialise the word embeddings used in the look-up table layer, while the position embeddings are initialised at random [5].

Materials and Method

For all potential drug-drug interactions that could occur in a single sentence, DDI extraction is acknowledged as a multiclass classification challenge. Each drug pair is categorised as either a no-interacting pair or one of the present forms of DDIs. The complete workflow of our CNN-based method for DDI extraction is illustrated by the requirement to classify all of the DDI candidates given a sentence including pharmaceuticals. The pre-processing programme first tokenizes sentences, blinds drugs, normalises tokens, and eliminates DDI possibilities with no interacting pairings. The DDI extraction is then performed using the CNN module. DDI candidates that were annotated in the training set are positive samples of various sorts, while the other candidates are negative samples during the training phase. Using these samples as training data, a CNN model must be created. All DDI candidates are divided into distinct categories of DDIs or non-DDIs throughout the test process [6].

We follow prior studies to blind the drugs in a sentence in the following manner to assure generalisation of machine learning-based methods: the two drugs of interest are replaced by "drug1" and "drug2" in the order of their occurrence, respectively, and all other drugs are replaced by "drug0." For instance, DDI candidates with context (referred to as instances) are created for the statement "When ALFENTA is provided in combination with other CNS depressants such as barbiturates, or tranquillizers," which contains four medicines. The four-layer CNN model, which is a version of the model for sentence classification in this work, is suggested for DDI extraction [7].

We then proceed to randomly remove units (along with their connections) from the networks during training in order to avoid neural networks from over fitting. The fully linked softmax layer does not receive the feature vector created by max pooling directly for classification. In order to create a new feature vector, we first set each element of at random and with a chance of zero (according to the Bernoulli distribution). The completely connected softmax layer is then fed the vector. Without dropping out, the feature vector is sent directly into the softmax layer at test time for classification. The word embeddings matrix, the position embeddings matrix, the filters, and the weight matrix of the softmax layer must all be modified during training for the CNN model. To learn the parameters, we employ stochastic gradient descent with shuffled minibatches and the AdaDelta update rule. We rescale the weight vectors of the softmax layer at each gradient descent step when their -norms are greater than a predetermined limit [8].

Discussion

We suggest a CNN-based approach in this study to extract DDIs from biomedical texts. To the best of our knowledge, CNN is being used for DDI extraction for the first time. Table 4 demonstrates how our CNN-based system beats all other systems now in use the majority of which is based on SVM and includes features like syntactic and features derived from outside lexical resources. Our CNN-based system has an advantage over modern SVM-based systems in that it does not use any manually defined characteristics produced by already available NLP toolkits [9]. Beyond the explicitly defined features, the features employed in the CNN-based system (i.e., word embeddings and position embeddings automatically learned during training) may also contain other important information. On the Drug Bank subset, position embedding enhances our system's performance, but not on the MEDLINE subset. The primary factor is that the Drug Bank subset's position distribution of words resembles that of the training set more closely than the MEDLINE subset does. To demonstrate this, we compare the average distance between two drugs of interest in the training set with that in the two test subsets, and discover that the difference between the training set and the MEDLINE subset is much smaller than that between the training set and the average distance in the training set (18.06) and the Drug Bank subset (15.07). (8.55). [10]

In comparison to the MEDLINE subset, our algorithm performs significantly better on the Drug Bank subset. There could be two causes: Abstracts for MEDLINE are typically prepared in technical terminology. In MEDLINE abstracts, lengthy and complicated

statements are frequently employed. In contrast, Drug Bank lines are frequently brief and to the point; the training set from MEDLINE samples pale in comparison to Drug Bank. Due to their relative rarity among the four categories of DDI instances, it is clear why our approach performs worse on integer instances. Only 4.7% (189/4021) of the training set's instances are ints. Considering the uneven distributions of various sorts of instances is one approach that could be improved [11].

Although our system performs better than any other systems now in use, it also has a significant amount of faults (listed in Table 5, where the numbers on the two sides of plus signs are negative instances predicted by the CNN model and negative instance filtering module, resp.), The majority of errors happen between positive and negative examples. 979 positive examples out of 277 are mistakenly categorised as negative instances (false negative instances). 134 negative occurrences are mistakenly categorised as good occurrences (false positive instances). In our system, there are only a few faults (69 out of 979) across four different types of DDIs. 39 int instances are incorrectly categorised as effect instances, making up 56.52% of these errors [12].

Conclusions

We suggest a CNN-based approach for DDI extraction in this research. The semantic content of words and the distances between words and two medications of interest, respectively, are captured by word embeddings and location embeddings, which are utilised to encode DDI instances. The 2013 DDI Extraction challenge corpus experiments show that the suggested CNN-based method beats existing cutting-edge DDI extraction methods. The application of CNN to DDI extraction is new. Our CNN-based approach takes location embeddings into account in addition to word embeddings. They neither rely on any NLP toolkits that are already in use.

Conflict of Interest

None

Acknowledgement

None

References

1   Yang Y, Adelstein SJ, Kassis AI (2009) Target discovery from data mining approaches. Drug Discovery Today 14:147-154.

2   Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M, et al. (2007) Drug-target network. Nature Biotechnology 25:1119-1126.

3   Bultinck J, Lievens S, Tavernier J (2012) Protein-protein interactions: network analysis and applications in drug discovery. Current Pharmaceutical Design 18:4619-4629.

4   Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M, et al. (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics24:13:232-240.

5   Sams-Dodd F (2005) Target-based drug discovery: is something wrong? Drug Discovery Today 10:139-147.

6   Zhang X, Li L, Ng MK, Zhang S (2017) Drug–target interaction prediction by integrating multitier network data. Computational Biology and Chemistry 69:185-193.

7   Wang YC, Yang ZX, Wang Y, Deng NY (2010) computationally probing drug-protein interactions via support vector machine. Letters in Drug Design and Discovery 7:370-378.

8   Ezzat A, Zhao P, Wu M, Li X, Kwoh C (2016) Drug-target interaction prediction with graph regularized matrix factorization. IEEE Transactions on Computational Biology and Bioinformatics 14:646-656.

9   Lu L, Yu H (2018) DR2DI: a powerful computational tool for predicting novel drug-disease associations. Journal of Computer-Aided Molecular Design 32:633-642.

10 Wishart DS, Knox C, Guo AC (2008) Drug Bank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research 36:901-906.

11 Bui Q, Sloot PM, Mulligen EM, Kors JA (2014) A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics 30:3365-3371.

12 Thomas P, Neves M, Solt I, Tikk D, Leser U, et al. (2011) Relation extraction for drug-drug interactions using ensemble learning. Drug-Drug Interaction Extraction 43:11-18.