Information Sciences 415–416 (2017) 100–109

Contents lists available at ScienceDirect

Information Sciences

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

Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks

Degen Huang[a], Zhenchao Jiang[b], Li Zou[c,*], Lishuang Li[a,*]

[a] School of Computer Science and Technology Dalian University of Technology, Dalian, Liaoning, China
[b] Sangfor Technologies, Shenzhen, Guangdong, China
[c] School of Computer Science and Information Technology, Liaoning Normal University, Dalian,Liaoning, China

ARTICLE INFO

Article history:
Received 7 January 2017
Revised 28 April 2017
Accepted 17 June 2017
Available online 19 June 2017

MSC:
00-01
99-00

Keywords:
Drug-Drug interaction extraction
Long short term memory
Two-stage approach

ABSTRACT

Since Drug-drug interactions (DDIs) can cause adverse effects when patients take two or more drugs and therefore increase health care costs, the extraction of DDIs is an important research area in patient safety. To improve the performance of Drug–drug interaction extraction (DDIE), we present a novel two-stage method in this paper. It first identifies the positive instances using a feature based binary classifier, and then a Long Short Term Memory (LSTM) based classifier is used to classify the positive instances into specific category. The experimental results show that the two-stage method has many advantages over one-stage ones, and among the factors related to LSTM, we find that the two layer bidirectional LSTM embedded with word, distance and Part-of-Speech obtains the highest F-score of 69.0%, which is state-of-the-art.

1. Introduction

Drug–drug interaction extraction (DDIE) is an important task in Biomedical Natural Language Processing (BioNLP) domain. The DDIExtraction 2013 challenge [1] is the second edition of the DDIExtraction Shared Task series, a community-wide effort to promote the implementation and comparative assessment of natural language processing (NLP) techniques in the field of the pharmacovigilance domain. In the challenge the DDIs need to be classified into four predefined DDI types (“advise”, “effect”, “mechanism” and “int”). “Advise” is assigned when a recommendation or advice regarding concomitant use of two drugs involved is described; “Effect” is assigned when the effect of the DDI is described; “Mechanism” is assigned when a DDI is described by its pharmacokinetic mechanism; “Int” is assigned when a DDI appears in the text without any additional information provided. The corpus was annotated manually consisting of 792 texts selected from the DrugBank database and other 233 Medline abstracts. This fined-grained corpus has been annotated with a total of 18,502 pharmacological substances and 5028 DDIs. The full DDIExtraction 2013 corpus which consists of 1017 documents (784 DrugBank documents and 233 MedLine documents, totally 27,792 instances for training and 5716 instances for testing) which was manually annotated with a total of 18,491 drug names and 5021 postive DDIs (4673 for DrugBank and 327 for MedLine).

[*] Corresponding authors.
E-mail addresses: huangdg@dlut.edu.cn (D. Huang), jzc@sangfor.com.cn (Z. Jiang), zoulicn@163.com (L. Zou), lilishuang314@163.com (L. Li).

http://dx.doi.org/10.1016/j.ins.2017.06.021
0020-0255/© 2017 Elsevier Inc. All rights reserved.

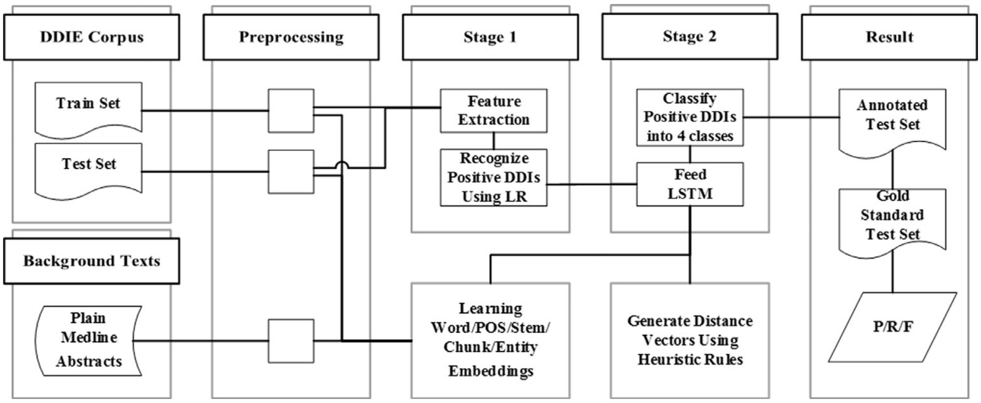

Fig. 1. The two-stage drug–drug interaction extraction model.

Support Vector Machine (SVM) based methods have achieved successful and promising results in the past five years, e.g., Chowdhury et al. [2] applied a two-stage hybrid kernel based relation extraction approach, taking advantage of different kernels of SVM, Thomas et al. [3] also combined several kernel methods. These two methods were the top two ranked teams in DDIExtraction 2013, and other teams such as [4,5] also used SVM as classifier. Afterwards, Kim et al. [6] used a feature based method, which is also a two-stage system, achieving an F-score of 0.670. It is clear that SVM is effective on this task, however, the one-stage methods generally cannot perform better than the two-stage ones [7]. On the other hand, one of limitations of SVM is the incapability of dealing with text of arbitrary length.

In recent years, the machine learning community has witnessed significant advances of deep learning, and deep learning based methods have been applied on related tasks. For example, Zeng et al. [8] exploited a convolutional deep neural network to extract lexical and sentence level features and further these features were fed into a softmax classifier to predict the relationship between two marked nouns. Sahu et al. [9] proposed a Joint AB-LSTM model that utilized word and position embedding as latent features on DDIExtracion 2013 and achieved competitive results against traditional feature based methods.

Compared to Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) can deal with input of arbitrary length, which is a good property for text data. Previous successes in real world applications with RNNs were limited due to practical problems when long time lags between relevant events make learning difficult, i.e., gradient vanishing [10]. The reason for this failure is the rapid decay of back-propagated error. The LSTM algorithm, an improved version of RNN, overcomes this by enforcing constant error flow. Using gradient descent, LSTM explicitly learns when to store information and when to access it. For many tasks, LSTMs are better than the standard RNNs. Almost all results based on RNNs are achieved with LSTMs, and many studies have attempted to solve text mining problems using LSTM, for example, Limsopatham et al. [11] investigated an approach for named entity recognition by enabling bidirectional LSTM to automatically learn orthographic features, Sutskever et al. [12] used a multilayered LSTM to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector sequence to sequence learning. Due to effectiveness of LSTM on other related text mining tasks, LSTM is expected to help improving the DDIE performance.

Considering that on the one hand, LSTM is a suitable machine learning model for text mining that can deal with input of arbitrary length and have enough layers to overcome gradient vanishing, and on the other hand, the distribution of DDIExtraction is highly skewed (the numbers of instances of “advise”, “effect”, “mechanism”, “int” and “negative” are 1047, 2047, 1621, 285 and 28,508, respectively), in this work, to improve the performance of DDIE, we present a novel two-stage method. In the first stage, we identify the four kinds of positive DDI instances from the negative ones using feature based binary classifier, and in the second stage, an LSTM based classifier is used to classify the positive instances into each of the four drug interaction types. We study many factors that possibly influence the LSTM model, such as part-of-speech (POS) tag embedding, distance information, dropout, etc. By conducting experiments, we show that word embedding, part-of-speech (POS) tag embedding, distance information and multi-layer bidirectional LSTM help to improve the performance of DDIE.

2. Method

A two-stage classifier offers a distinct advantage over a one-stage classifier for DDIE, not only because it is highly skewed towards one class (the negative class) but also because this majority class is clearly semantically distinct from the other positive classes. Two-stage classifer comprises two classifiers in separate stages. In the first stage, a binary classifer is trained to classify drug pairs into positive and negative classes. Then in the second stage, only the instances that are classified as positive by the first classifer are considered, and classified into one of four types within the positive class (“advise”, “effect”, “mechanism” and “int”) using a multi-class classifier.