Detecting Adverse Drug Reactions from Biomedical Texts With Neural Networks

Ilseyar Alimova
Kazan Federal University, Kazan, Russia
ISAlimova@kpfu.ru

Elena Tutubalina
Kazan Federal University, Kazan, Russia
Samsung-PDMI Joint AI Center, PDMI RAS, St. Petersburg, Russia
elvtutubalina@kpfu.ru

Abstract

Detection of adverse drug reactions in post-approval periods is a crucial challenge for pharmacology. Social media and electronic clinical reports are becoming increasingly popular as a source for obtaining health-related information. In this work, we focus on extraction information of adverse drug reactions from various sources of biomedical text-based information, including biomedical literature and social media. We formulate the problem as a binary classification task and compare the performance of four state-of-the-art attention-based neural networks in terms of the F-measure. We show the effectiveness of these methods on four different benchmarks.

1 Introduction

Detection of adverse drug reactions (ADRs) in the post-marketing period is becoming increasingly popular, as evidenced by the growth of ADR monitoring systems (Singh et al., 2017; Shareef et al., 2017; Hou et al., 2016). Information about adverse drug reactions can be found in the texts of social media, health-related forums, and electronic health records. We formulated the problem as a binary classification task. The ADR classification task addresses two sub-tasks: (a) detecting the presence of ADRs in a textual message (message-level task) and (b) detecting the class of an entity within a message (entity-level task). In this paper, we focus on the latter task. Different from the message-level classification task, which aims to determine whether a textual fragment such as tweet or an abstract of a paper includes an ADR mention or not, the objective of the entity-level task is to detect whether a given entity (a single word or a multi-word expression) conveys adverse drug effect in the context of a message. For example, in “He was unable to sleep last night because of pain”, the health condition ‘pain’ trigger insomnia. Meanwhile, in “after 3 days on this drug I was unable to sleep due to symptoms like a very bad attack of RLS”, there is an entity ‘unable to sleep’ associated with drug use and can be classified as ADR.

Inspired by recent successful methods, we investigated various deep neural network models for entity-level ADR classification (Alimova and Tutubalina, 2018). Our previous experiments showed that Interactive Attention Neural network (IAN) (Ma et al., 2017) outperforms other models based on LSTM (Hochreiter and Schmidhuber, 1997). In this paper, we continue our study and compare IAN with the following attention-based neural networks for entity-level ADR classification: (i) Attention-over-Attention (AOA) model (Huang et al., 2018); (ii) Attentional Encoder Network (AEN) (Song et al., 2019); (iii) Attention-based LSTM with Aspect Embedding (ATAE-LSTM) (Wang et al., 2016). We conduct extensive experiments on four benchmarks which consist of scientific abstracts and user-generated texts about drug therapy.

2 Related Work

Different approaches are utilized to identify adverse drug reactions (Sarker et al., 2015; Gupta et al., 2018b; Harpaz et al., 2010). First works were limited in the number of study drugs and targeted ADRs due to limitations of traditional lexicon-based approaches (Benton et al., 2011; Liu and Chen, 2013). In order to eliminate these shortcomings, rule-based methods have been proposed (Nikfarjam and Gonzalez, 2011; Na et al., 2012). These methods capture the underlying syntactic and semantic patterns from social media posts. Third group of works utilized popular machine learning models, such as support vec-
tor machine (SVM) (Liu and Chen, 2013; Sarker et al., 2015; Niu et al., 2005; Bian et al., 2012; Alimova and Tutubalina, 2017), conditional random fields (CRF) (Aramaki et al., 2010; Miftahutdinov et al., 2017), and random forest (RF) (Rastegar-Mojarad et al., 2016). The most popular handcrafted features are n-grams, parts of speech tags, semantic types from the Unified Medical Language System (UMLS), the number of negated contexts, the belonging lexicon based features for ADRs, drug names, and word embeddings (Dai et al., 2016). One of the tracks of the shared task SMM4H 2016 was devoted to ADR classification on a tweet level. The two best-performing systems applied machine learning classifier ensembles and obtained 41.95% F-measure for ADR class (Rastegar-Mojarad et al., 2016; Zhang et al., 2016). Two other participants utilized SVM classifiers with different sets of feature and obtained 35.8% and 33% F-measure (Ofoghi et al., 2016; Jonnagaddala et al., 2016). During SMM4H 2017, the best performance was achieved by SVM classifiers with a variety of surface-form, sentiment, and domain-specific features (Kiritchenko et al., 2018). This classifier obtained 43.5% F-measure for ‘ADR’ class. Sarker and Gonzales outperformed these result utilizing SVM with a more rich set of features and the tuning of the model parameters and obtained 53.8% F-measure for ADR class (Sarker and Gonzalez, 2015). However, these results are still behind the current state-of-the-art for general text classification (Lai et al., 2015).

Modern approaches for the extracting of ADRs are based on neural networks. Saldana adopted CNN for the detection of ADR relevant sentences (Miranda, 2018). Huynh T. et al. applied convolutional recurrent neural network (CRNN), obtained by concatenating CNN with a recurrent neural network (RNN) and CNN with the additional weights (Huynh et al., 2016). Gupta S. et al. utilized a semi-supervised method based on co-training (Gupta et al., 2018a). Chowdhury et al. proposed a multi-task neural network framework that in addition to ADR classification learns extract ADR mentions (Chowdhury et al., 2018).

Methods for sentiment analysis are actively adopted in the medical domain as well as in other domains (Serrano-Guerrero et al., 2015; Rusnachenko and Loukachevitch, 2018; Ivanov et al., 2015; Solovyev and Ivanov, 2014). In the field of aspect-level sentiment analysis, neural networks are popularly utilized (Zhang et al., 2018). Ma et al. proposed Interactive Attention Network which interactively learns attentions in the contexts and targets, and generates the representations for targets and contexts separately (Ma et al., 2017). The model compared with different modifications of Long Short Term Memory (LSTM) models and performed greatest results with 78.6% and 72.1% of accuracy on restaurant and laptop corpora respectively. Song et al. introduced Attentional Encoder Network (AEN) (Song et al., 2019). AEN eschews recurrence and employs attention based encoders for the modeling between context and target. The model obtained 72.1% and 69% f accuracy on restaurant and laptop corpora respectively. Wang et al. utilized Attention-based LSTM, which takes into account aspect information during attention (Wang et al., 2016). This neural network achieved 77.2% and 68.7% of accuracy restaurant and laptop corpora respectively. The Attention-over-Attention neural network proposed by Huang et al. models aspects and sentences in a joint way and explicitly captures the interaction between aspects and context sentences (Huang et al., 2018). This approach achieved the best results among the described articles wit 81.2% and 74.5% of accuracy on restaurant and laptop corpora.

To sum up this section, we note that there has been little work on utilizing neural networks for entity-level ADR classification task. Most of the works used classical machine learning models, which are limited to linear models and manual feature engineering (Liu and Chen, 2013; Sarker et al., 2015; Niu et al., 2005; Bian et al., 2012; Alimova and Tutubalina, 2017; Aramaki et al., 2010; Miftahutdinov et al., 2017; Rastegar-Mojarad et al., 2016). Most methods for extracting ADR so far dealt with extracting information from the mention itself and a small window of words on the left and on the right as a context, ignoring the broader context of the text document where it occurred (Korkontzelos et al., 2016; Dai et al., 2016; Alimova and Tutubalina, 2017; Bian et al., 2012; Aramaki et al., 2010). Finally, in most of the works experiments were conducted on a single corpus.

3 Corpora

We conducted our experiments on four corpora: CADEC, PsyTAR, Twitter, TwiMed. Further, we
briefly describe each dataset.

**CADEC** CSIRO Adverse Drug Event Corpus (CADEC) consists of annotated user reviews written about Diclofenac or Lipitor on askapatient.com (Karimi et al., 2015). There are five types of annotations: ‘Drug’, ‘Adverse effect’, ‘Disease’, ‘Symptom’, and ‘Finding’. We grouped diseases, symptoms, and findings as a single class called ‘non-ADR’.

**PsyTAR** Psychiatric Treatment Adverse Reactions (PsyTAR) corpus (Zolnoori et al., 2019) is the first open-source corpus of user-generated posts about psychiatric drugs taken from Askapatient.com. This dataset includes reviews about four psychiatric medications: Zoloft, Lexapro, Effexor, and Cymbalta. Each review annotated with 4 types of entities: adverse drug reactions, withdrawal symptoms, drug indications, sign/symptoms/illness.

**TwiMed** TwiMed corpus consists of sentences extracted from PubMed and tweets. This corpus contains annotations of diseases, symptoms, and drugs, and their relations. If the relationship between disease and drug was labeled as ‘Outcome-negative’, we marked disease as ADR, otherwise, we annotate it as ‘non-ADR’ (Alvaro et al., 2017).

**Twitter** Twitter corpus include tweets about drugs. There are three annotations: ‘ADR’, ‘Indication’ and ‘Other’. We consider ‘Indication’ and ‘Other’ as ‘non-ADR’ (Nikfarjam et al., 2015).

Summary statistics of corpora are presented in Table 1. As shown in this table, the CADEC and PsyTAR corpora contain a much larger number of annotations than the TwiMed and Twitter corpora.

4 Models

4.1 Interactive Attention Network

The Interactive Attention Network (IAN) network consists of two parts, each of which creates a representation of the context and the entity using the vector representation of the words and the LSTM layer (Ma et al., 2017). The obtained vectors are averaged and used to calculate the attention vector. IAN uses attention mechanisms to detect the important words of the target entity and its full context. In the first layer of attention, the vector of context and the averaged vector of the entity and in the second, the vector of the entity and the averaged vector of context are applied. The resulting vectors are concatenated and transferred to the layer with the softmax activation function for classification.

4.2 Attention-over-Attention

Attention-over-Attention (AOA) model was introduced by Huang et al. (Huang et al., 2018). This model consists of two parts which handle left and right contexts, respectively. Using word embeddings as input, BiLSTM layers are employed to obtain hidden states of words for a target and its context, respectively. Given the hidden semantic representations of the context and target the attention weights for the text is calculated with AOA module. At the first step, the AOA module calculates a pair-wise interaction matrix. On the second step, with a column-wise softmax and row-wise softmax, the module obtains target-to-sentence attention and sentence-to-target attention. The final sentence-level attention is calculated by a weighted sum of each individual target-to-sentence attention using column-wise averaging of sentence-to-target attention. The final sentence representation is a weighted sum of sentence hidden semantic states using the sentence attention from AOA module.

4.3 Attentional Encoder Network

The Attentional Encoder Network (AEN) eschews complex recurrent neural networks and employs attention based encoders for the modeling between context and target (Song et al., 2019). The model architecture consists of four main parts: embedding layer, attentional encoder layer, target-specific attention layer, and output layer. The embedding layer encodes context and target with pre-trained word embedding models. The attentional encoder layer applies the Multi-Head Attention and the Point-wise Convolution Transformation to the context and target embedding representation. The target-specific attention layer employs another Multi-Head Attention to the introspective context representation and context-perceptive target representation obtained on the previous step. The output layer concatenates the average pooling outputs of previous layers and uses a fully connected layer to project the concatenated vector into the space of the targeted classes.

4.4 Attention-based LSTM with Aspect Embedding

The main idea of Attention-based LSTM with Aspect Embedding (ATAE-LSTM) is based on ap-
Table 1: Summary statistics of corpora.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>ADR</th>
<th>non-ADR</th>
<th>Max sentence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADEC (Karimi et al., 2015)</td>
<td>1231</td>
<td>5770</td>
<td>550</td>
<td>236</td>
</tr>
<tr>
<td>PsyTAR (Zolnoori et al., 2019)</td>
<td>891</td>
<td>4525</td>
<td>2987</td>
<td>264</td>
</tr>
<tr>
<td>TwiMed-Pubmed (Alvaro et al., 2017)</td>
<td>1000</td>
<td>264</td>
<td>983</td>
<td>150</td>
</tr>
<tr>
<td>TwiMed-Twitter (Alvaro et al., 2017)</td>
<td>637</td>
<td>329</td>
<td>308</td>
<td>42</td>
</tr>
<tr>
<td>Twitter (Nikfarjam et al., 2015)</td>
<td>645</td>
<td>569</td>
<td>76</td>
<td>37</td>
</tr>
</tbody>
</table>

5 Experiments

In this section, we compare the performance of the discussed neural networks with Interactive Attention Neural Network.

5.1 Settings

We utilized vector representation trained on social media posts from (Miftahutdinov et al., 2017). Word embedding vectors were obtained using word2vec trained on a Health corpus consists of 2.5 million reviews written in English. We used an embedding size of 200, local context length of 10, the negative sampling of 5, vocabulary cutoff of 10, Continuous Bag of Words model. Coverage statistics of word embedding model vocabulary: CADEC – 93.5%, Twitter – 80.4%, PsyTAR – 54%, TwiMed-Twitter – 81.2%, TwiMed-Pubmed – 76.4%. For the out of vocabulary words, the representations were uniformly sampled from the range of embedding weights. We used a maximum of 15 epochs to train IAN and ATAE-LSTM and 30 epochs to train AEN and AOA on each dataset. We set the batch size to 32 for each corpus. The number of hidden units for LSTM layer is 300, the learning rate is 0.01, 12 regularization is 0.001. We applied the implementation of the model from this repository.

5.2 Experiments and Results

All models were evaluated by 5-fold cross-validation. We utilized the F-measure to evaluate the quality of the classification.

The results are presented in Table 2. The results show that IAN outperformed other models on all corpora. IAN obtained the most significant increase in results compared to other models on Cadec and Twitter-Pubmed corpora with 81.5% and 87.4% of the macro F-measures, respectively. We assume that the superiority of the IAN results in comparison with other models is due to the small number of parameters being trained and the small size of the corpora.

The AOA model achieved the second-place result on all corpora except Twitter. The AOA results for PsyTAR (81.5%) and Twimed-Twitter (79.5%) corpora state on par with IAN model, while for the rest corpora, the results are significantly lower. This leads to the conclusion that the model is unstable for highly imbalanced corpora.

The ATAE-LSTM model with 78.6% of macro F-measure outperformed AEN and AOA models results on Twitter corpora and achieved comparable with AOA results on Twimed-Pubmed corpora (80.1%). This result shows that ATAE-LSTM applicable to a small size imbalanced corpora.

The AEN model achieved comparable with other models results on PsyTAR (80.2%) corpora and significantly lower results on Twitter (66.7%), Cadec (49%) and Twimed-Pubmed (74.3%) corpora. 72.4% of F-measure on Twimed-Twitter corpus states on par with the ATAE-LSTM model (73.5%). This leads to the conclusion that the presence of multiple attention layers did not give the improvement in results.

6 Conclusion and Feature Research Directions

We have performed a fine-grained evaluation of state-of-the-art attention-based neural network models for entity-level ADR classification task. We have conducted extensive experiments on four benchmarks. Analyzing the results, we have found that that increasing the number of attention layers...
Table 2: Macro F-measure classification results of the compared methods for each datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Twitter</th>
<th>Cadec</th>
<th>PsyTAR</th>
<th>Twimed-Twitter</th>
<th>Twimed-PubMed</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAN</td>
<td>.794</td>
<td>.815</td>
<td>.817</td>
<td>.819</td>
<td>.874</td>
</tr>
<tr>
<td>AEN</td>
<td>.667</td>
<td>.490</td>
<td>.802</td>
<td>.742</td>
<td>.743</td>
</tr>
<tr>
<td>AOA</td>
<td>.752</td>
<td>.752</td>
<td>.815</td>
<td>.795</td>
<td>.803</td>
</tr>
<tr>
<td>ATAE-LSTM</td>
<td>.786</td>
<td>.702</td>
<td>.807</td>
<td>.735</td>
<td>.801</td>
</tr>
</tbody>
</table>

did not give an improvement in results. Addition an aspect vector to the input layer also did not give significant benefits. IAN model showed the best results for entity-level ADR classification task in all of our experiments.

There are three future research directions that require, from our point of view, more attention. First, we plan to add knowledge-based features as input for IAN model and evaluate their efficiency. Second, apply these models to the entity-level ADR classification task for texts in other languages. Finally, we plan to explore the potential of new state-of-the-art text classification methods based on BERT language model.

Acknowledgments

This research was supported by the Russian Science Foundation grant no. 18-11-00284.

References


Nestor Alvaro, Yusuke Miyao, and Nigel Collier. 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. JMIR public health and surveillance, 3(2).


