A neural network algorithm for extracting pharmacological information from Russian-language Internet reviews on drugs

A G Sboev[1,2], S G Sboeva[3],A V Gryaznov[1], A V Evteeva[1], R B Rybka[1] and M S Silin[1]

[1] National Research Center "Kurchatov Institute", Moscow, Russia
[2] National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russia
[3] I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

E-mail: sag111@mail.ru

Abstract. The paper presents a neural network algorithm for analyzing online user reviews of drugs. The algorithm was validated on specially prepared and annotated corpora. The basis of the algorithm is a neural network model combining convolution and recurrent layers, context-dependent vector representations of words, conditional random fields and additional features of words obtained from different dictionaries. The proposed model showed accuracies comparable to the state-of-the-art results for this task on the corpora for other languages.

1. Introduction

Internet texts, in particular, messages from social networks, discussion groups and forums, can contain a large amount of meaningful information, including that related to consumption sphere and healthcare in general. Patients' opinions may be beneficial for evaluating efficiency of medicines in addition to clinical investigations. Messages in Internet sources often contain unique characteristics of adverse reactions which clinical trials may not have revealed due to time limitations. In addition, users may report the effect of non-standard usage of medicines on diseases other than prescribed. Such information would be very useful for a pharmacovigilance database where risks and advantages of drugs would be registered for safety monitoring and for the possibility to form hypotheses of using existing drugs for treating other diseases. The above gives relevance to developing methods for automated extraction of information on efficiency of medicines from Internet review texts.

The aim of this work is to propose the efficient neural network algorithm for the task of extracting significant pharmacological entities from Russian-language drug reviews on base of up-to-date neural network solutions and the first-created corpus of labeled examples in order to evaluate the state of the art of this task for the Russian language.

2. Related works

There are many methods in literature, devoted to analysis of English-language texts on medicines, which accuracy is based on corpora containing data of different sources: healthcare records,

Table 1. A summary of existing corpora of texts from social networks

Корпус	Кол-во сущностей	F_1^{exact}	F_1^{partial}
CADEC [4]	8118	0.661	0.905
PsyTar [5]	7414	0.49	–
TwiMed [6]	1200	–	0.648
Twitter annotated corpus [2]	2130	–	0.611

scientific articles, social net messages. In this article, we consider the last source: reviews of Internet users on medicines. The analysis of this source is assumed as more complicated because these texts in many cases wrote on speech language with mistakes, without following to formal grammar and punctuation rules.There are many methods in literature, devoted to analysis of English-language texts on medicines, which accuracy is based on corpora containing data of different sources: healthcare records, scientific articles, or social net messages. In this article, we consider the last source: reviews of Internet users on medicines. This source is assumed more complicated to analyze because texts are often written in colloquial language, contain mistakes, and do not following formal grammar and punctuation rules. The most widely known corpora of this type are CADEC, PsyTar, Twimed and Twitter.

CADEC corpus of adverse drug event annotations - is a corpus collected of 1253 medical posts with 7398 sentences. The following entities were annotated: Drug, ADR, Symptom, Disease,Findings, totally on 13 drugs.

TwiMed corpus (Twitter and PubMed comparative corpus of drugs, diseases, symptoms, and their relations)[1] contains 1000 tweets and 1000 sentences from Pubmed for 5 or 30 drugs with annotations for 3 144 entities, 2 749 relations, and 5 003 attributes.

Twitter annotated corpus [2] consists of randomly selected tweets containing drug name mentions. Two types of annotations are currently available: Binary and Span. The binary annotated part contains 10 822 tweets annotated by the presence or absence of ADRs. 1 239 (11.4%)tweets contain ADR mentions and 9583 (88.6%) do not. The span annotated part contains 2 131 tweets (which include 1 239 tweets containing ADR mention from the binary annotated part). Marked Types are: ADR, beneficial effect, indication, other (medical signs or symptoms).

The PsyTAR corpus contains 891 reviews on medicines, collected randomly from an online healthcare forum and split into 6 009 sentences. The texts have the following entities labeled: adverse drug reaction (ADR), withdrawal symptoms (WD), sign/symptoms/illness (SSI), indications for use (DL), and other.[3]

The work [5] achieved $F_1^{\text{exact}} = 0.49$ on PsyTAR used the UMLS dictionaries, morphological and syntactic features, and negation words. The work [2] demonstrates $F_1^{\text{partial}} = 0.611$ on the Twitter corpus with use of a wide range of features, including word forms, linguistic characteristics, parts of speech, semantic tags. A wide range of features is used for Twitter texts analysis: word forms, linguistic characteristics, parts of speech, semantic tags.

The current level of accuracy, on texts from Internet sources, is usually achieved using modern deep learning methods based on bi-directional LSTM layers (BiLSTM) ($F_1^{\text{partial}} = 0.648$ [6] with Part-of features -Speech (PoS) along with pretrained models of vector representation of words.Some works use a combination of BiLSTM layers with convolutional layers of character encoding [7], as well as a layer based on Conditional Random Fields (CRF), to determine the label of the token(слова) [8] ($F_1^{\text{exact}} = 0.661$ на CADEC). Therefore, further to create our model for working with Russian-language texts, we are based on the BiLSTM deep learning architecture along with the use of modern pre-trained language models of the Russian language. Table 1

shows the results obtained using different approaches and features.

3. Materials and methods

3.1. Russian drug review corpus

The corpus we created for this work contains 1 660 texts of patient reviews on medicines of different pharmacotherapeutic groups. The language of the reviews does not always follow formal grammar and punctuation rules of Russian. The corpus has mentions of the following entities manually labeled: Medication (17 779 mentions), Adverse Drug Reaction (844), Disease (9 285), Note (2 319). Labelling was performed by annotators qualified in medicine or pharmacy. A more detailed description of the corpus, as well as entity type definitions that the annotators acted in accordance with are presented at `sagteam.ru/en/med-corpus`.

Before further analysis, an input text is split into sentences and then into tokens (words and punctuation marks), and each word is assigned its part of speech with the help of the UDpipe library, which was chosen as a baseline method at the CoNLL competition [9].

3.2. Model components

Our model is built out of the following neural network modelling components:

(i) Fully-connected layer (FCL) in which each artificial neuron receives all the components of an input vector, performs a weighed sum of them, and after subtracting the neuron's bias passes the sum through its activation function to obtain the neuron's output activity value.

(ii) The embedding stage, which converts words into continuous vectors with the help of preliminarily-trained language models. As such model we use the ELMo deep neural network, trained on the Russian WMT News [10] corpus, as published in the DeepPavlov [11] library.

(iii) Convolution layer, in which each neuron receives only those components of an input vector that fall into the neuron's receptive area, and the weights by which these components are multiplied when summed are shared over the neurons.

(iv) Long short-term memory (LSTM) [12], a sequence-processing layer in which the output activity value of a cell depends, in addition to weighed sum of the cell's input components, on the memory value stored when processing the previous element of a sequence.

(v) Conditional Random Fields (CRF) [13], one of the commonly used methods that takes a sequence of tokens as input, estimates the probabilities of labels (from a predefined set),and returns the best scoring label sequence.

3.3. Entity definition neural network model

We consider the problem of mention detection and classification as a multi-label classification of tokens – words and punctuation marks – in sentences. For each of the three entities – ADR, Medication and Disease – its own neural network is trained. That way, mentions of different entities can intersect, so that one word can have several tags. The output for each token (word) is a tag in the BIO format: the "B" tag indicates the first word of a mention of the considered entity, the "I" tag is used for subsequent words within the mention, and the "O" tag means that the word is outside of an entity mention.

The topology of model is presented on Fig. 1. For encoding tokens into input vectors we use: a) character-wise one-hot encoding following by convolution layer, b) word vectors obtained from the embedding language model (see Section 3.2), c) dictionary features obtained by looking the tokens up in the Vidal [14] and Meshrus [15] dictionaries, along with part-of-speech (PoS) tags one-hot-encoded [16], e) sentiment features, including psycholinguistic markers (further denoted as PM) and features from the LIWC [17] dictionaries (denoted as LIWC). Then, the vectors thus

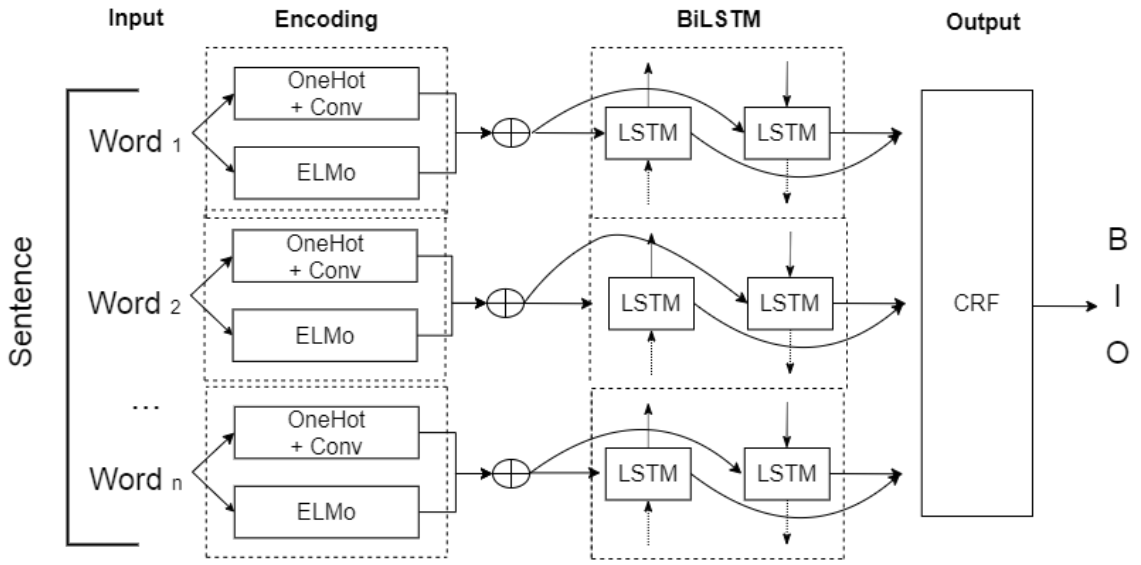

Figure 1. Summary of the neural network architecture. An input word sequence is processed by a bidirectional LSTM, where hidden states of forward LSTM and backward LSTM are concatenated. Finally, the resulting vector is processed by a CRF-based layer (in the model denoted Basemodel_FC, the latter is replaced with a fully-connected layer with the SoftMax activation function). The outputs of the model are the possibilities for a word to be outside of a mention (and thus to have the O-tag), to be at the beginning (B-tag) or inside (I-tag) of a mention of the entity the model is trained for.

encoded are presented to the neural network model, consisting of BiLSTM layers and an output layer, either fully-connected with the SoftMax activation function or CRF-based.

The performance of the model is characterised by two metrics commonly used for this type of tasks (for instance, in the works surveyed in Section 2):

(i) F_1^{exact} – F1-score calculated over mentions whose starting and ending points match the ground truth labelling exactly;

(ii) F_1^{partial}, which is calculated on base of the number of words for which belonging to mentions was correctly identified. This metric is thus able to characterize partial match of the labelling to the ground truth.

3.4. Experiments

We compared the following variations of the model depicted in Fig. 1: different combinations of input features, different number of BiLSTM layers, and two types of output layers.

The variation denoted in the tables as Basemodel_CRF repeats the model configuration from Fig. 1 without the use of additional features. Basemodel_FC differs from the Basemodel_CRF only in that the output layer is a fully-connected one with Softmax activation function. Further configuration changes are noted the tables by the abbreviations defined in Section 3.3.

The performance of the models was assessed by 5-fold cross-validation. The training set of each fold was further divided into the training set proper and the validation set, in the 9:1 proportion. Training was stopped upon meeting the early stopping criterion on the loss function on the validation set, but not later than after 70 epochs. The loss function was cross-entropy, and training was performed by the Adam optimizer with cyclical learning rate [18].

The performance of different model variations is presented in Table 2. The best scores by the

Table 2. Entity Definition Accuracy(%).

Модель	ADR		Medication		Disease	
	F_1^{partial}	F_1^{exact}	F_1^{partial}	F_1^{exact}	F_1^{partial}	F_1^{exact}
Basemodel_FC	45.0 ± 2.4	28.7 ± 1.3	82.0 ± 0.6	75.6 ± 0.4	62.0 ± 0.8	50.9 ± 2.2
Basemodel_FC + Meshrus	**45.6 ± 3.4**	27.3 ± 2.9	82.2 ± 1.0	76.0 ± 1.1	62.7 ± 0.5	51.1 ± 1.5
Basemodel_FC + PoS, Meshrus	42.4 ± 3.7	27.7 ± 3.2	**82.2 ± 0.4**	75.5 ± 0.5	62.3 ± 0.9	51.3 ± 0.9
Basemodel_CRF + Meshrus	45.2 ± 1.7	30.2 ± 1.1	**82.2 ± 0.4**	75.7 ± 0.6	62.7 ± 0.9	52.1 ± 0.3
Basemodel_CRF + 3-layer BiLSTM + LIWC + PM + PoS + Meshrus + Vidal	45.0 ± 4.9	**33.2 ± 2.9**	82.0 ± 1.1	**76.6 ± 0.8**	**63.7 ± 0.9**	**56.0 ± 0.6**

F_1^{exact} metric are 76.6, 33.2, and 56.0 for Medication, ADR и Disease entity classes respectively. The results obtained are comparable with the level of accuracy obtained for similar cases in other languages, and can be considered as state of the art for the task.

4. Conclusion

The proposed algorithm for extracting pharmacological information from Russian-language texts demonstrates accuracy comparable to those obtained for other languages. The relatively low accuracy results for detecting mentions of adverse drug effects (ADR) is explained by the low number of such examples in the current version of the collected dataset. Further, we plan to replenish the dataset with more ADR entities to increase the accuracy of their detection.

Acknowledgments

This work has been supported by the Russian Science Foundation grant 20-11-20246 and carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru/.

References

[1] Website, Alvaro N, Miyao Y and Collier N 2019 Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations

[2] Wang W 2016 Mining adverse drug reaction mentions in twitter with word embeddings *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*

[3] Zolnoori M, Fung K W, Patrick T B, Fontelo P, Kharrazi H, Faiola A, Shah N D, Wu Y S S, Eldredge C E, Luo J *et al.* 2019 *Data in brief* **24** 103838

[4] Dai X, Karimi S and Paris C 2017 Medication and adverse event extraction from noisy text *Proceedings of the Australasian Language Technology Association Workshop 2017* pp 79–87

[5] Zolnoori M, Fung K W, Patrick T B, Fontelo P, Kharrazi H, Faiola A, Wu Y S S, Eldredge C E, Luo J, Conway M *et al.* 2019 *Journal of biomedical informatics* **90** 103091

[6] Gupta S, Gupta M, Varma V, Pawar S, Ramrakhiyani N and Palshikar G K 2018 Multi-task learning for extraction of adverse drug reaction mentions from tweets *European Conference on Information Retrieval* (Springer) pp 59–71

[7] Li F, Zhang M, Fu G and Ji D 2017 *BMC bioinformatics* **18** 198

[8] Tang B, Hu J, Wang X and Chen Q 2018 *Wireless Communications and Mobile Computing* **2018**

[9] Straka M 2018 *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* 197–207

[10] Website 2019 Statmt - internet resource about research in the field of statistical machine translation accessed: 2019-05-24 URL `www.statmt.org`

[11] Website 2019 Deeppavlov - an open source library for deep learning end-to-end dialog systems and chatbots accessed: 2019-05-24 URL `https://deeppavlov.ai`

[12] Kalchbrenner N, Danihelka I and Graves A 2015 *arXiv preprint arXiv:1507.01526*

[13] Lafferty J D, McCallum A and Pereira F C N 2001 Conditional random fields: Probabilistic models for segmenting and labeling sequence data *Proceedings of the Eighteenth International Conference on Machine Learning* ICML '01 (San Francisco, CA, USA) pp 282–289 ISBN 1558607781 URL `http://repository.upenn.edu/cis_papers/159/`

[14] Vidal S 2011 *M.: AstraFarmService*

[15] Website 2019 Umls metathesaurus - mshrus (mesh russian) - synopsis accessed: 2019-05-24 URL `http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/`

[16] Harris D and Harris S 2010 *Digital design and computer architecture* (Morgan Kaufmann)

[17] Litvinova O, Seredin P, Litvinova T and Lyell J 2017 Deception detection in Russian texts *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics* pp 43–52

[18] Smith L N 2017 Cyclical learning rates for training neural networks *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE) pp 464–472