PROCEEDINGS OF SCIENCE

Relation Extraction from Texts Containing Pharmacologically Significant Information on base of Multilingual Language Models

Anton Selivanov,[a,*] **Artem Gryaznov,**[a] **Roman Rybka,**[a,b] **Alexander Sboev,**[a,b,c] **Sanna Sboeva**[a,d] **and Yuliya Klyueva**[d]

[a] *NRC "Kurchatov Institute",*
Akademika Kurchatova sq., 1, Moscow, Russian Federation
[b] *Russian Technological University "MIREA",*
Vernadsky av., 78, Moscow, 119454, Russian Federation
[c] *National Research Nuclear University "Moscow Engineering Physics Institute (MEPhi)",*
Kashira Hwy, 31, Moscow, Russian Federation
[d] *Sechenov First Moscow State Medical University,*
Bolshaya Pirogovskaya st., 2/4, Moscow, Russian Federation

E-mail: aaselivanov.10.03@gmail.com, sag111@mail.ru

In this paper we estimate the accuracy of the relation extraction from texts containing pharmacologically significant information on base of the expanded version of RDRS corpus, which contains texts of internet reviews on medications in Russian.
The accuracy of relation extraction is estimated and compared for two multilingual language models: XLM-RoBERTa-large and XLM-RoBERTa-large-sag. Earlier research proved XLM-RoBERTa-large-sag to be the most efficient language model for the previous version of the RDRS dataset for relation extraction using a ground-truth named entities annotation. In the current work we use two-step relation extraction approach: automated named entity recognition and extraction of relations between predicted entities. The implemented approach has given an opportunity to estimate the accuracy of the proposed solution to the relation extraction problem, as well as to estimate the accuracy at each step of the analysis.
As a result, it is shown, that multilingual XLM-RoBERTa-large-sag model achieves relation extraction macro-averaged f1-score equals to 86.4% on the ground-truth named entities, 60.1% on the predicted named entities on the new version of the RDRS corpus contained more than 3800 annotated texts. Consequently, implemented approach based on the XLM-RoBERTa-large-sag language model sets the state-of-the-art for considered type of texts in Russian.

**** The 6th International Workshop on Deep Learning in Computational Physics (DLCP2022) ****
**** 6-8 July 2022 ****
**** JINR, Dubna, Russia ****

[*] Speaker

PoS(DLCP2022)014

An analysis of biomedical texts, in particular pharmaceutical ones, is relevant and being actively investigated [1–4] due to its practical meaning for such tasks, as the pharmacovigilance [5], social research, the marketing analysis of drug market, an automatic processing of medical records, developing telemedical and decision support system. The development of machine learning methods, in particular deep neural networks, combined with access to a large datasets of biomedical texts, has significantly increased the efficiency of solving problems of automatic analysis of natural language. Modern analysis tools are based on neural network models that are pre-trained on large datasets and additionally trained for the target task using labeled data corpora. A number of research groups have proposed trained models of this type for solving problems of the considered topic, for example: PubMedBERT [6], BioBERT [7], BioRoBERTa [8], BioELECTRA [9], BioALBERT [10].

The use of language models in the analysis methods for English language allows one to achieve high accuracy in solving the basic problems of named entitiy recognition (NER) and extraction of the relations between them (RE) [6–13]. Existing research is more focused on the analysis of English texts, for which there are two main approaches to the united solution of these problems (End-to-End):

- Cascade – sequential approach that involves extracting named entities and then establishing relations with two separate neural network models;
- Joint – solution of the both tasks based on a single neural network model.

The advantage of the first approach is the ability to control the settings of various models as part of the general solution. The "Joint" approach uses the loss functions from both stages of the analysis simultaneously. However, currently there are no End-to-End methods for united solution of the NER and RE tasks for Russian. In a number of works [6–13] authors offers solutions for NER or RE task. At the same time, the best accuracy is obtained using language models trained on a large amount of unlabeled textual data. In [14] we proposed a method for solving the RE problem based on the expert annotation of named entities. In this paper, we propose an End-to-End method to the tasks of NER and RE based on a cascade approach, combining the best solutions based on the pre-trained language models (see Section "Materials and Methods"). To fine-tune the models that are parts of the proposed solution, the Russian Drug Review Corpus (RDRS) has been used [15] (see Section "Dataset"). The set of the named entitiy types used for the study includes: Adverse Drug Reaction (ADR), Drug name, Disease name, Indication, and Source of the medication information, that are the basic set for pharmacovigilance purposes. Computational experiments setup, model estimation technique and technical platform are described in Section "Experiments". An estimation of the accuracy depending on the size of the dataset and on the language model used is presented in Section "Results".

Materials and Methods

Named Entity Recognition The solution to this problem is based on the approach we proposed in [15], that implements multilabel classification of text tokens according to the BIO scheme. This named entity markup scheme assumes that the first token of the entity receives the tag "B-«entity class name»" (beginning of the entity), subsequent ones "I-«entity class name»" (in the

entity), and tokens that aren't included in the entity get tag "O" (out of context). The classification is based on the neural network model of the Transformer architecture.

To solve the multilabel classification problem, the activities of the last layer of the transformer are processed with the set of linear layers with softmax activation function. Each layer represents specific named entities class and contains 3 neurons for every tag from BIO.

Figure 1 shows the scheme of the NER task solution and estimation.

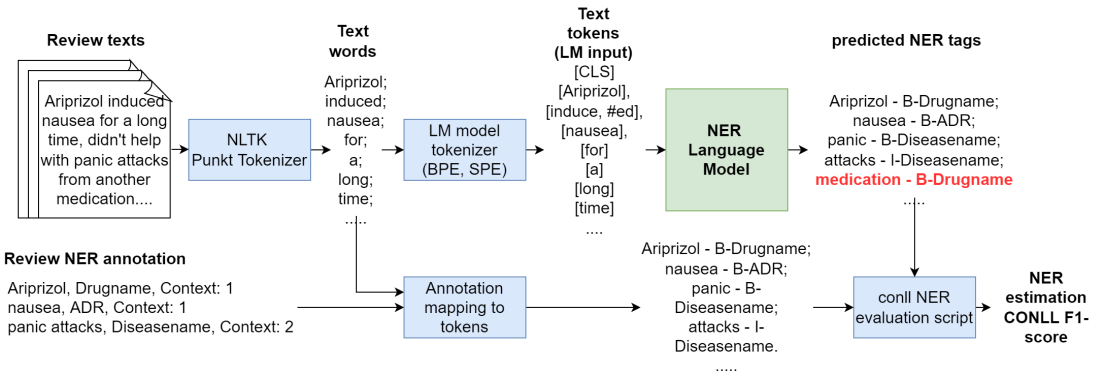

Figure 1: Named Entity Recognition pipeline scheme.

Classification is performed for each token in the text. Text tokens are obtained through application of a) NLTK tokenizer to obtain words and entity boubdaries using in subsequent matching of the ground-truth and automatically predicted entities, b) tokenization of the words into the simpler elements – tokens, using language model tokenizator (SentencePiece [16]).

After receiving a set of BIO tags from the network for each named entity class, tokens are combined into words, that used to evaluate named entity recognition accuracy.

Estimation based on the metric from the Computational Natural Language Learning 2003 competition (CoNLL [17]), further referred to as f1-conll. According to this method, entities are considered correctly predicted only if their boundaries completely aligned with the ground-truth entities when compared by tokens.

Relation Extraction Similar to the paper [14], we consider the Relation Extraction task as a classification: the task is to determine the presence of a relationship for a pair of entities extracted from the text.

Pairs are formed by exhaustive search of the combinations of entities that correspond by tags to the types of relationships under consideration. That is, for the ADR-Drugname relationship type, all possible combinations of entities of the ADR and Drugname types are considered.

If the entities of the designated types have the same context in the expert annotation, then there is a relation between them, otherwise there is no relation between the entities.

Thus, the task is to classify each generated pair of entities for the presence of a relation.

Language models are used for the classification, the input data is the text of the review and pair of the entities, pre-processed by the following procedure:

1. The text of the considered entity pair is concatenated via the service token [ESEP];

2. The resulting construct is concatenated with the text containing the target entities via the service token [TXTSEP];

3. A service token [CLS] is concatenated to the generated text, which aggregates information about the rest of the text tokens in the process of network training and is used to classify a pair of entities.

Example of the result of the processing for the pair of entities "aspirin" and "stomach hurt" from the text: "Took an aspirin, it didn't help much for my head, but my stomach hurt." The text after conversion will be presented as follows: "[CLS] aspirin [ESEP] stomach hurt [TXTSEP] Took aspirin, it didn't help much for my head, but my stomach hurt."

The text converted in this way is split into "tokens" using the tokenization tool that was used when setting up the language model. The tokenized text is the input of the neural network. Transformer language model is used to represent tokens as the vectors, relation class is determined using linear layer with the Softmax activation function (the dimension of the layer corresponds to the number of classes of the problem being solved). The input of the classification layer is the vector of the [CLS] token.

The training of the neural network is performed on the basis of the ground-truth annotation of the named entities, and the accuracy is determined as the macro-averaged f1: the predicted classes are compared to the ground-truth annotation.

In the case of automatically predicted named entities, relationships are considered correctly predicted if both the entities and the relationship type are correctly defined.

Figure 2 shows the scheme of the relation extraction task solution and estimation in either cases of expert named entities annotation and automatically predicted named entities.

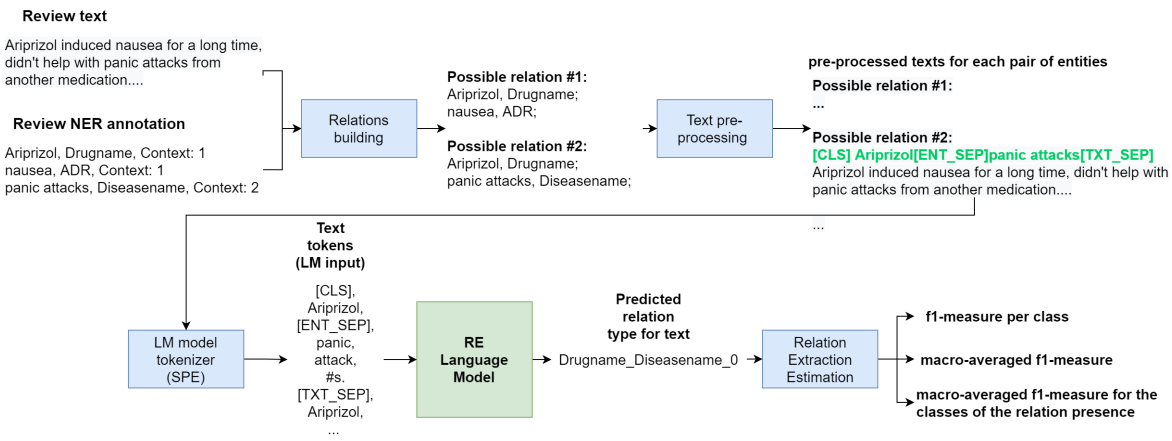

Figure 2: Relation Extraction pipeline scheme

Language Models According to the results of our previous research, the best results for the considered task are achieved with a XLM-RoBERTa language model type. In this paper, we explore the use of two models of this type:

1. XLM-RoBERTa-large is a multilingual model trained on the masked language modeling task. Training process used 2.5 TB of the data from the CommonCrawl project, containing texts

written in 100 languages. At the same time, Russian is the second most representative language in the entire corpus after English. The model contains 24 layers, 1024 hidden neurons, 16 attention heads and 550 million parameters, the dictionary contains 150K SentencePiece tokens [18].

2. XLM-RoBERTa-sag [15] – version of the XLM-RoBERTa-large model trained on a large set of untagged drug reviews in Russian, including: RuDReC texts [19] (about 1.4 million texts), and collected reviews on medicines from the site irecommend.ru (about 250 thousand texts)).

Dataset Our research based on the corpus RDRS [15]. The corpus contains the texts of Internet-reviews on the medications from the site Otzovik.ru. The corpus is provided with a rich annotation including 3 groups of named entities:

- Medication – a group describing the drug, it's class, form of release, methods of administration, dosage, etc.;
- Disease – a group describing the disease, including the name of the disease, symptoms, and observed effects from using the Medication;
- Adverse Drug Reaction (ADR) – Adverse reactions observed when using Medication.

When writing a review, users can describe different use cases with different effects, or mention several drugs at once. To separate several referenced drugs or different uses of the same drug, references are combined into groups that we call "contexts". Each reference may refer to one or more contexts. Each context includes references to attributes of the drug, reasons for taking the drug, if any, (disease or symptoms), and effects resulting from taking the drug (change of state or adverse reactions that occurred). An example of the context and named entities marked up in the review is shown on the Figure 3.

The current main version of the corpus includes 2800 randomly selected reviews, annotated by expert pharmacists. The corpus has been expanded with the 1,000 additional texts, tagged annotated by the same principle, but selected to increase the variety of drug and disease nomenclature.

In continuation of the work [14], we considered a set of the most interesting pairs of named entities from a practical point of view:

- ADR–Drugname – adverse effect of the particular medication;
- Drugname–SourceInfodrug – source of the information about medication (e.g. "my brother gave me advice", "apothecary mentioned");
- Drugname–Diseasename – a link between the disease and medication that user administered against it;
- Diseasename–Indication – symptoms of the particular disease (e.g., "red rash", "high temperature").

Antiviral drug **Ingavirin** **capsules**[1]
Caution, may cause a severe **allergic reaction**! [1,2]
Recently, almost all of my famili was ill with severe **VRI**. [1]
I bought the famous advertised **Ingavirin**. [1]
At night, I **started to choke**. **It's like sand was poured into respiratory tract**,
The throat became all red, **the nose was completely blocked**. [1,2]
I woke up my mother, she **injected** me and **ampoule** of **suprastin**,
it seemed to let go. [2]

	Medication	Disease	ADR
1	Ingavirin Antiviral capsules	VRI	Allergic reaction Started to choke It's like sand was poured into respiratory tract The throat became all red the nose was completely blocked
2	Suprastin injected ampoule	Allergic reaction Started to choke It's like sand was poured into respiratory tract The throat became all red the nose was completely blocked It seemed to let go	

Figure 3: Example of the context annotation. Numbers after the sentences show context assigned to the entity mentions in sentence. Phrases highlighted in green and listed in the first column are mentions of medication attributes, red color indicates disease, symptoms and dynamics, blue color is used for ADR mentions, phrases highlighted with purple color are annotated both as Disease and ADR but included in different contexts. 1st context is about taking Ingavirin against VRI and ADR it caused. 2nd context is about Suprastin that was taken against ADR caused by Ingavirin.

Further computational experiments were performed on a subsample of texts, including only a) the types of entities that are contained in the relationships under consideration, and b) more than one context.

The table 1 provides brief information about the selected subsamples of data from the main and additional text sets of the RDRS corpus.

In the process of training the neural network, examples of pairs of entities without relation were also used. As examples with no connection in the corpus, those pairs were considered, the entities of which occur only in different contexts, therefore, should not be defined as related entities.

Experiments We used a 5-fold cross-validation (fixed random seed equals to 42) with the division of the RE-RDRS-main subsample into training, validation, and test parts. The training and validation parts have been extended with additional examples of the RE-RDRS-ext corpus. Thus, the test part in both subsamples was the same, which makes it possible to evaluate the effect of increasing the

Table 1: Brief statistics on the considered datasets. RE-RDRS-main and RE-RDRS-ext – subsamples of the main and main+additional parts of the RDRS 2022 dataset

	RE-RDRS-main	RE-RDRS-ext
text number	230	426
avg text length(in words)	173	186
min text length	49	49
max text length	241	321
Entity number	2431	5149
by type:		
ADR	180	728
Drugname	902	1784
SourceInfoDrug	293	592
DiseaseName	522	826
Indication	534	1219
Relation number	5228	11784
by type:		
ADR_Drugname	775	3378
Drugname_Diseasename	1997	3437
Drugname_SourceInfoDrug	1190	2555
Diseasename_Indication	1266	2414

corpus size.

All the experiments on named entity recognition and relation extraction were carried out using the hyperparameters in the Table 2 (chosen manually):

Table 2: Hyperparameters of the computational experiments on relation extraction and named entity recognition.

Hyperparameter	NER	RE
Loss function	Categorical crossentropy	Categorical crossentropy
Optimization function	Adam	NAdam
Imit learning rate	0.00005	0.00001
Batch size	8	8
Max sequence length	512	512
Epoch number	10	10
data type	float32	float16

Other parameter values are set as default parameters of simple transformers software library. Tech platform parameters: CPU Intel Xeon E5-2650v2 8 cores; GPU NVIDIA V100 16 GB, RAM 128 GB.

The final accuracy score of the NER model is calculated as macro-averaged f1-conll for all entity types.

The macro-averaged f1-score is used as a metric for assessing the accuracy of the model for the relation extraction task.

In the case of using named entities predicted by the model in the problem of identifying relationships between entities, the relationship is considered correctly predicted only if the predicted entities exactly match the entities from the ground-truth annotation both in terms of boundaries and classes. The proportion of correctly identified related entities is calculated based on the method from [13].

Results The table 3 presents the accuracy of experiments on the extraction of named entities.

Table 3: Estimation of the language models accuracy on the NER task. XLMR – XLM-RoBERTa-large, XLMR-sag – XLM-RoBERTa-sag, Drug – DrugName, Source – SourceInfoDrug entity tag; macro-f1 – macro-averaged metric over f1-score for each entity

Dataset	LM	ADR	Disease	Indication	Drug	Source	macro-f1
RE-RDRS-main	XLMR	47.9	85.6	64.5	93.6	69.2	72.2
	XLMR-sag	47.4	87.3	66.1	94.1	67.6	72.5
RE-RDRS-ext	XLMR	55.7	86.2	67.6	96.6	73.4	75.9
	XLMR-sag	60.1	87.2	67.5	96.6	72.5	76.7

The table shows that using the extension version of dataset (RE-RDRS-ext) leads to higher accuracy in prediction of the ADR and Drugname types of named entities.

The table 4 shows the accuracy of identifying relationships between entities both with the reference expert annotation (Ground-truth) and with the annotation obtained by the neural network for NER (Predicted) with macro-averaging over the considered set of relations.

Table 4: Estimation of the language models accuracy on the relation extraction task using two versions of datasets.

RE entities origin	Language model	RE-RDRS-main	RE-RDRS-ext
Ground-truth Named Entities	XLM-RoBERTa large	86.7	87.2
	XLM-RoBERTa large sag	87.3	86.4
Predicted Named Entities	XLM-RoBERTa large	55.9	60.0
	XLM-RoBERTa large sag	57.1	60.1

Analysis of the table 4 shows that the use of a specialized language model configured on a large set of unlabeled texts of the subject area gives an advantage with a small number of annotated examples in solving the problem of relation extraction. If there is a sufficiently large set of annotated data available for training, the influence of the choice of the language model as part of the overall solution on the accuracy of the result is leveled out.

Conclusion The results of the study establish the current level of accuracy for the relation extraction from the Russian-language text as part of a general solution for analyzing texts of reviews on pharmaceutical products (60%).

The comparison of language models of the same type, in the settings of which texts of general vocabulary and texts of reviews about pharmaceuticals were used, showed that with a sufficient number of annotated examples for the target task, the difference in their results isn't significant. However, in the case of a limited set of labeled data, it is more promising to use a language model adjusted on the texts of reviews as part of the overall solution. Further research will be aimed at developing the relation extraction approach based on a single neural network model with an assessment of its efficiency on multilingual corpora of annotated examples.

Acknowledgement The study was supported by a grant from the Russian Science Foundation (project no. 20-11-20246);

This work has been 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] M. Nadif and F. Role, *Unsupervised and self-supervised deep learning approaches for biomedical text mining*, *Briefings in Bioinformatics* **22** (2021) 1592.

[2] A. Ben Abacha, Y. Mrabet, Y. Zhang, C. Shivade, C. Langlotz and D. Demner-Fushman, *Overview of the MEDIQA 2021 shared task on summarization in the medical domain*, in *Proceedings of the 20th Workshop on Biomedical Language Processing*, (Online), pp. 74–85, Association for Computational Linguistics, June, 2021, DOI.

[3] K. Roberts, T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff et al., *Searching for scientific evidence in a pandemic: An overview of trec-covid*, *Journal of Biomedical Informatics* **121** (2021) 103865.

[4] G. Mourits, *Identification of adverse drug reactions in dutch electronic health records*, Master's thesis, 2022.

[5] I. Segura-Bedmar and P. Martínez, *Pharmacovigilance through the development of text mining and natural language processing techniques*, *Journal of Biomedical Informatics* **58** (2015) 288.

[6] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu et al., *Domain-specific language model pretraining for biomedical natural language processing*, .

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*, *Bioinformatics* **36** (2019) 1234.

[8] P. Lewis, M. Ott, J. Du and V. Stoyanov, *Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art*, in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, (Online), pp. 146–157, Association for Computational Linguistics, Nov., 2020, https://www.aclweb.org/anthology/2020.clinicalnlp-1.17.

[9] K. raj Kanakarajan, B. Kundumani and M. Sankarasubbu, *Bioelectra: Pretrained biomedical text encoder using discriminators*, in *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 143–154, 2021.

[10] U. Naseem, A.G. Dunn, M. Khushi and J. Kim, *Benchmarking for biomedical natural language processing tasks with a domain specific albert*, *arXiv preprint arXiv:2107.04374* (2021) .

[11] E.V. Tutubalina, Z.S. Miftahutdinov, R.I. Nugmanov, T.I. Madzhidov, S.I. Nikolenko, I.S. Alimova et al., *Using semantic analysis of texts for the identification of drugs with similar therapeutic effects*, *Russian Chemical Bulletin* **66** (2017) 2180.

[12] S. Wu and Y. He, *Enriching pre-trained language model with entity information for relation classification*, in *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 2361–2364, 2019.

[13] M. Eberts and A. Ulges, *Span-based joint entity and relation extraction with transformer pre-training*, in *ECAI 2020*, pp. 2006–2013, IOS Press (2020).

[14] A. Sboev, A. Selivanov, I. Moloshnikov, R. Rybka, A. Gryaznov, S. Sboeva et al., *Extraction of the relations among significant pharmacological entities in russian-language reviews of internet users on medications*, *Big Data and Cognitive Computing* **6** (2022) 10:1.

[15] A. Sboev, S. Sboeva, I. Moloshnikov, A. Gryaznov, R. Rybka, A. Naumov et al., *Analysis of the full-size russian corpus of internet drug reviews with complex ner labeling using deep learning neural networks and language models*, *Applied Sciences* **12** (2022) 491:1.

[16] T. Kudo and J. Richardson, *Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*, *arXiv preprint arXiv:1808.06226* (2018) .

[17] E.F. Tjong Kim Sang and F. De Meulder, *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition*, in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 142–147, Association for Computational Linguistics, 2003.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán et al., *Unsupervised cross-lingual representation learning at scale*, *arXiv preprint arXiv:1911.02116* (2019) .

[19] E. Tutubalina, I. Alimova, Z. Miftahutdinov, A. Sakhovskiy, V. Malykh and S. Nikolenko, *The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews*, *Bioinformatics* (2020) .