Article

Extraction of the Relations between Significant Pharmacological Entities in Russian-Language Reviews of Internet Users on Medications

Alexander Sboev [1,2,*], Anton Selivanov [1], Ivan Moloshnikov [1], Roman Rybka [1], Artem Gryaznov [1], Sanna Sboeva [1] and Gleb Rylkov [1]

[1] National Reseatch Center "Kurchatov Institute"; nrcki@nrcki.ru
[2] National Research Nuclear University "MePHI"; VVRomanychev@mephi.ru
* Correspondence: sag111@mail.ru

Abstract: Nowadays, an analysis of virtual media to predict society's reaction to any events or processes is a task of great relevance. Especially it concerns meaningful information on healthcare problems. Internet sources contain a large amount of pharmacologically meaningful information useful for pharmacovigilance purposes and repurposing drug use. An analysis of such a scale of information demands developing the methods that require the creation of a corpus with labeled relations among entities. Before, there have been no such Russian language datasets. This paper considers the first Russian language dataset where labeled entity pairs are divided into multiple contexts within a single text (by used drugs, by different users, by the cases of use, etc.), and a method based on the XLM-RoBERTa language model, previously trained on medical texts to evaluate the state-of-the-art accuracy for the task of indication of the four types of relationships among entities: ADR–Drugname, Drugname–Diseasename, Drugname–SourceInfoDrug, Diseasename–Indication. As shown based on the presented dataset from the Russian Drug Review Corpus, the developed method achieves the F1-score of 81.2% (obtained using cross-validation and averaged for the four types of relationships), which is 7.8% higher than the basic classifiers.

Keywords: pharmacological text corpus; automatic relation extraction; natural language processing; deep learning

Citation: Sboev, A.; Selivanov, A.; Rybka, R.; Gryaznov A. Extraction of the Relations between Significant Pharmacologic Entities in Russian-Language Reviews of Internet Users on Medications. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

The development of virtual communication opportunities through social networks and special Internet resources expands the possibility of discussing the use of certain medicines.

An analysis of such scale of information demands the development of extraction methods for pharmacologically meaningful information. This, in turn, requires a corpus containing relations between various pharmacologically-related entities. Such English-language corpora are widely presented in the literature, in particular, DDI (Drug-Drug Interaction), ADE (Adverse Drug Event), etc. These corpora contain selected pharmaceutically relevant entities of different types as well as the relationships between them. A more detailed analysis of the corpora is presented in Section 2. However, at the moment there is only one large domain-oriented dataset in Russian: Russian Drug Review Corpus (RDRS) of Internet user reviews with complex NER labels, that was presented by our group [1,2]. Now we present an extension of this corpus that includes highlighted relationships between the individual named entities most relevant to further research of drug efficiency (see Section 3.1).

Automation of the process of extracting meaningful information from a review written in natural language requires a sequential solution of the following separate tasks: text segmentation, Named Entity Recognition (NER), Relation Extraction (RE), structuring of the extracted information, and comprehensive evaluation of the results. In this paper, we focus on the task of automated extraction of relationships between named entities.

Such a formulation, as opposed to a comprehensive solution of the problem of Named Entity Recognition (NER) with simultaneous extraction of links between them (combined

approach, in the literature: “joint” or “end-to-end”), facilitates assessment of the complexity of the problem as a whole and the accuracy of solving its sub-tasks. The results of our review (See Section 2) show, that deep neural networks are the most promising technology for textual data analysis in terms of relation extraction. In this paper, we use a model based on the XLM-RoBERTa language model, pre-trained on a huge unlabelled corpus of drug reviews. Section 3 contains details of the model configuration and its setting.

Based on this model, a set of computational experiments was performed on the different parts of the RDRS corpus. Section 5.1 presents the accuracy of indication of only the single relation ADR–Drugname on a part of the data including only these entities (628 documents with 845 positive and 239 negative relations). The result on the macro-averaged F1-metric is equal to 95%. Next, Section 5.2 presents evaluations on a subpart of the corpus containing reviews with multiple contexts. This experiment aims to obtain the state-of-the art results for the task of relation extraction for the following four relation types: ADR–Drugname, Drugname–Diseasename, Drugname–SourceInfoDrug, Diseasename–Indication. The results of the model presented in this work are compared with the results of the set of baseline methods: Multinomial Naive Bayes Classifier, Linear Support Vector Machine,and Dummy Classifier (see Section 5).

2. Related Works

The development of textual data analysis tools depends on annotated data necessary for tuning algorithms and assessing their performance. There are a set of corpora of textual data in the English language with a markup of pharmacologically relevant entities and relationships. These corpora differ by the types of texts (online reviews, tweets, clinical extracts, etc.) and by the level of detail of the annotated named entities and relationships. Some studies provided the achievable accuracy of extracting relationships between pharmacologically relevant entities using the developed methods based on these corpora.

In [3] on the DDI (Drug-Drug Interaction) dataset, which contains excerpts about drug interactions from the DrugName and MedLine databases, the model based on BERT SciBert [4] was used to solve the task of classifying the sentences for relationships between the selected drugs. The model showed a result of 84.08% on the f1-micro metric. In [5], the performance of the SpERT model based on the BERT language model on the ADE v2 dataset [6] is presented. The ADE v2 dataset contains sentences from the abstracts of PubMed scientific articles with relations between medical drugs and their adverse reactions. The model sequentially solves the problem of extracting named entities and the relationships between them. To solve the problem of identifying named entities, all possible consecutive sets of words in the text (limited in length) are generated and then classified by the model according to the type of entity. The results of the classification are filtered, forming pairs of entities for which the model determines the presence of a relationship and its type. Such model achieves the value of the f1-macro metric equal 79.24%.

From the datasets on biomedical topics with markup for solving the problem of identifying relationships between named entities, it is also possible to select the corpora of the i2b2 Competition Corps Workshop on Natural Language Processing Challenges for Clinical Records organized by the Department of Biomedical Informatics (DBMI) at Harvard Medical School provides datasets called n2c2, that also could be highlighted as the biomedical datasets with annotation for the relation extraction between named entities. The datasets consist of full texts of medical records in English. The data annotation is enriched within each competition as the scope of the competition expands and changes.

The task of extracting relationships between named entities is considered in the 2009 [7], 2010 [8] and 2018 [9] corpora. In [3] on the DDI (Drug-Drug Interaction) dataset, which contains excerpts about drug interactions from the DrugName and MedLine databases, the model based on BERT SciBert [4] was used to solve the task of classifying the sentences for relationships between the selected drugs. The model showed a result of 84.08% on the f1-micro metric. In [5], the performance of the SpERT model based on the BERT language

model on the ADE v2 dataset [6] is presented. The ADE v2 dataset contains sentences from the abstracts of PubMed scientific articles with relations between medical drugs and their adverse reactions. The model sequentially solves the problem of extracting named entities and the relationships between them. To solve the problem of identifying named entities, all possible consecutive sets of words in the text (limited in length) are generated and then classified by the model according to the type of entity. The results of the classification are filtered, forming pairs of entities for which the model determines the presence of a relationship and its type. Such model achieves the value of the f1-macro metric equal 79.24%.

At present, a fairly limited set of Russian-language corpora for the relation extraction tasks are publicly available. However, those corpora facilitate the apriori assessment of the accuracy in the extraction of the relationships between named entities of different types, not related to pharmaceuticals: RuSERRC [10] – 80 manually annotated texts with entities from computer science subjects (software, database, programming languages, etc.). RuREBus [11] – 300 annotated texts of strategic programs of the Ministry of Economic Development of the Russian Federation, containing various relations between the entities of the following types: Social Objects, Actions, Goals, Tasks; RURED [12] – a corpus of 536 annotated texts about economics, containing entities of type Geographic Objects, Names, Age, Currencies, etc., as well as relationships of various types between them; Factrueval [13] – 255 annotated texts with entities of type Persons, Locations and Organizations, and also relations: Ownership, Occupation, Meeting, and Deal; NEREL [14] – 933 annotated documents with the markup of a large number of entities, including: Persons, Organizations, Geopolitical entities, numbers, dates, time, money, age, etc., as well as links between them. On the RuSERC corpus (split by sentences) BERT-based architecture, R-BERT [15] was used to obtain a result of 67% for macro-f1 metric, on the RuRBus corpus (in documents) also R-BERT architecture [15] was used to get a result of 44% for micro-f1 metric on the corresponding corpus. On the RUED dataset (in sentences), the span-BERT architecture achieved 78% accuracy on the f1 metric (method for aggregating f1 across different classes wasn't specified). On the Factrueval (in documents) dataset, the method achieved 66% accuracy on the fact extraction task (relationships from multiple entities). On the NEREL dataset the RuBERT model achieved a precision of 51% (in documents) (the method of f1 aggregation across different classes wasn't specified).

As for the Russian-language corpora annotated to extract the relationships between pharmacologically significant entities, the only corpus of this type is the Russian Drug Review Corpus (RDRS 2800 reviews), which is considered in this paper. Therefore, the accuracy demonstrated in the works above with other types of texts is only an evaluation of the possible accuracy of determining relationships for pharmacological entities, which is an additional motivation to perform the present work.

Summarizing the information above, it can be concluded that the current trend in identifying relationships between named entities is the use of models with transformer architecture pre-trained on large datasets. Further in this work, we develop this approach based on the XLM-RoBERTa language model [16] using the Russian Drug Review Corpus (RDRS) [2] described in the Data section and available at the Sagteam [1] project.

3. Materials and Methods

3.1. Datasets

This paper uses the Russian Drug Review Corpus (RDRS) [2], which contains 2800 texts of drug reviews written by Internet users. The corpus contains markup for 18 types of named entities, which can be divided into 3 groups:

- Medication – this group includes everything related to the mentions of drugs and drugs manufacturers, including: Drug name, Drug class, Drug form, Route (how to use the drug), Dosage, SourceInfoDrug (source of the drug information) etc.;
- Disease – this group contains entities related to the diseases or reasons for using the drug (disease name, indications or symptoms), as well as the obtained effects

(NegatedADE – the drug was inefficient, Worse – mention of deterioration, BNE-pos - mention of improvement of the condition) etc.
- ADR – mentions of adverse reactions that occurred.

Among the entire corpus of 1,590 texts, entities were marked up into "lines of meaning" – "contexts", linking those entities of the review that relate to the same case of drug use described. Different contexts arise in particular when describing the use of multiple drugs in treatment, or different effects following the use of a single drug for different conditions, or when the review describes the use of a drug by different people. Thus, entities that occur in the same context are related, while entities from different contexts are considered unrelated.

An example of context annotation is shown in the Figure 1. The main (1st) context of the review is about drug "orvirem" which caused an allergy. It includes mentions with number 1 above them: "antiviral" (drugclass), "syrup" (drugform) "orvirem" (drugname), multiple mentions of "allergy" (ADR), "red spots" (ADR), "swelling on the face" (ADR), "1 day" (Duration). There are other contexts of the review:

- 2nd context: "allergy" (Diseasename), "red spots" (Indication), "zyrtek" (Drugname), "the situation did not improve" (NegatedADE), "it seems to have gotten even worse" (Worse).
- 3d context: "allergy" (Diseasename), "red spots" (Indication), "doctor" (SourceInfodrug), "On her recommendation" (SourceInfodrug), "smecta" (Drugname), "the situation did not improve" (NegatedADE), "it seems to have gotten even worse" (Worse).
- 4th context: "allergy" (Diseasename), "red spots" (Indication), "doctor" (SourceInfodrug), "On her recommendation" (SourceInfodrug), "suprastin" (Drugname), "the situation did not improve" (NegatedADE), "it seems to have gotten even worse" (Worse), "Injected" (Route), "The redness seems to pass" (BNE-Pos), ""swelling on the face still remains" (NegatedADE).
- 5th context: "allergy" (Diseasename), "red spots" (Indication), "doctor" (SourceInfodrug), "prednisone" (Drugname), "Injected" (Route), "The redness seems to pass" (BNE-Pos), ""swelling on the face still remains" (NegatedADE).

In Tables 1, 4 the quantitative characteristics of the corpus with contextual markup are presented.

Table 1. Distribution of the number of contexts

Contexts Count	1	2	3	>3
Texts Count	682	559	218	131

Table 2. Average lengths of the contexts in corpus

	Average mentions count	Average tokens count
Main context	19.9	38.9
Other contexts	3.7	6.6

In this paper, the following pairs of entities are chosen as the most interesting to analyze from the practical point of view:

- ADR-Drugname —- the relationship between the drug and its side effects;
- Drugname-SourceInfodrug —- the relationship between the medication and the source of information about it (e.g., "was advised at the pharmacy", "the doctor recommended it");
- Drugname-Diseasename —- the relationship between the drug and the disease;
- Diseasename-Indication —- the connection between the illness and its symptoms (e.g., "cough", "fever 39 degrees").

Two subsets of the original corpus were compiled for the experiments:

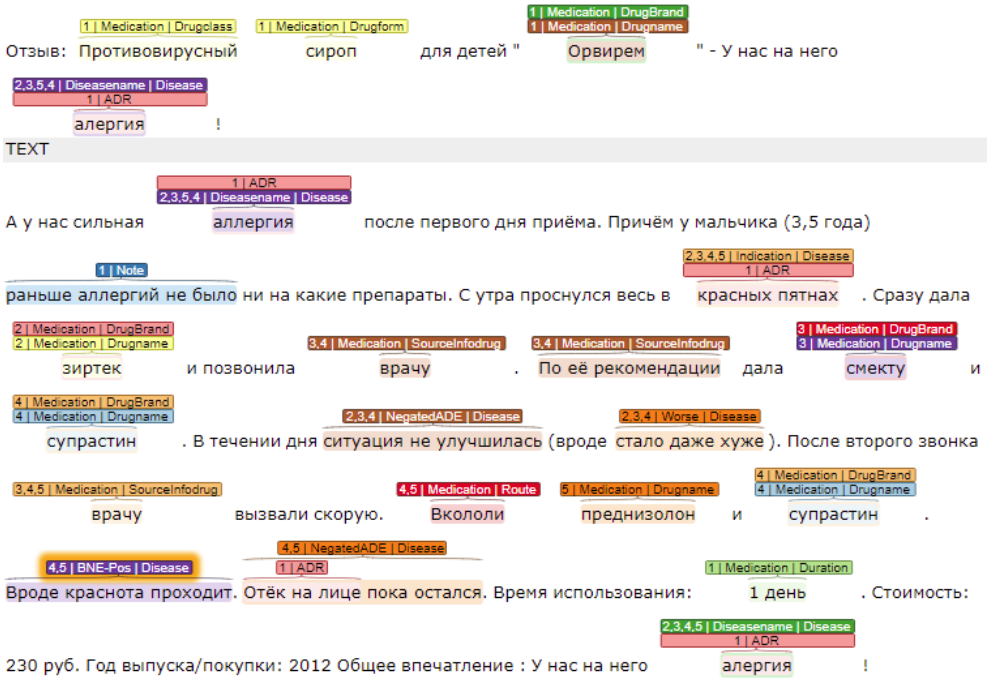

Figure 1. The text says: Review: Antiviral syrup for children "Orvirem" - We have an allergy to it! TEXT We have a severe allergy after the first day of taking it. Moreover, the boy (3.5 years old) had no allergies to any drugs before. In the morning he woke up covered in red spots. I immediately gave him zyrtek and called the doctor. On her recommendation, I gave smecta and suprastin. During the day, the situation did not improve (it seems to have gotten even worse). After the second call to the doctor, an ambulance was called. Injected with prednisone and suprastin. The redness seems to pass. The swelling on the face still remains. Usage time: 1 day. Price: 230 rubles. Year of release / purchase: 2012 Overall impression: We have an allergy to it!.

1. The first one includes 628 texts containing ADR and Drugname entity pairs. The experiments on this part were aimed at selecting the most effective combinations of input feature representations and hyper-parameters of the used methods. The texts of the RDRS corpus that contain ADR and Drugname entities were divided into training and test parts, the composition of which is presented in the Table 3.
2. The second part includes texts that contain multiple contexts, the total number of such texts is 908. Statistics on the types of relationships are presented in Table 4. This corpus is used to establish the current level of accuracy in determining the relationships between pharmacologically-significant entities in Russian-language review texts.

Table 3. Statistics on the RDRS dataset part with ADR-Drugname relations

Number of	Train	Test
Texts	502	126
Sentences	4016	1008
Words	82425	20961
"ADR" type entities	1461	356
"Drugname" type entities	1416	368
Relations	3444	845
Avg. numbers of relations per text	6.9	6.7

Table 4. Statistics on the types of relations in the RDRS corpus with 908 multi-context reviews.

Relation classes	ADR & Drugname		Drugname & Diseasename		Drugname & SourceInfoDrug		Diseasename & Indication	
	pos.	neg.	pos.	neg.	pos.	neg.	pos.	neg.
Relation number	1913	917	4277	2153	2700	1232	2588	701
Text fraction	0.273	0.204	0.634	0.514	0.598	0.457	0.416	0.148

Experiments with these subsets are described further in Section 4.

3.2. Models

3.2.1. Deep Learning Methods

Language Models

In this work the XLM-RoBERTa-sag model[2] was used. Original XLM-RoBERTa is a multilingual language model based on the transformer [17] architecture and trained on a larger multilingual corpus from the CommonCrawl project which contains 2.5TB of texts. The XLM-RoBERTa-sag is a result of additional training of the model XLM-RoBERTa [16] on a dataset of unlabeled internet texts about medicines (~1.65M texts).

This type of model is based on the Transformer topology [17], that consists of multi-head attention layers, which create vector representations of input data parts (words in case of NLP) that encode information about their context.

Text pre-processing includes text splitting into words or word parts – "tokens". In the case of XLM-RoBERTa-sag SentencePiece tokenizer [18] is used.

Language models are currently considered to be standard in modern natural language processing. During the adjustment experiments we used two versions of the model:

- XLM-RoBERTa-base-sag – 12 Transformer blocks, 768 hidden neurons, 8 Attention Heads, 125 millions of parameters, 2 epochs of additional training on Russian texts about medications;
- XLM-RoBERTa-large-sag – 24 Transformer blocks, 1024 hidden neurons, 16 Attention Heads, 355 millions of parameters, 1 epoch of additional training on Russian texts about medications;

Input text pre-processing

To solve the classification task, transformer-based language models use special token added to the input sequence: [CLS]. During the input data processing this token accumulates information about the text as a whole. At the training stage model weights are adjusted to the state in which they create a vector representation of the [CLS] token informative in terms of current task to solve, in other words, they aim at minimization of the loss function during the class prediction based on the vector of the [CLS] token.

In the approach proposed in this work, the classification is performed on the basis of the information about a pair of entities, for which the existence of a relationship is determined, and the text that mentions this pair. Figure 2 shows conceptual scheme of our approach to the task of relation extraction with the language model.

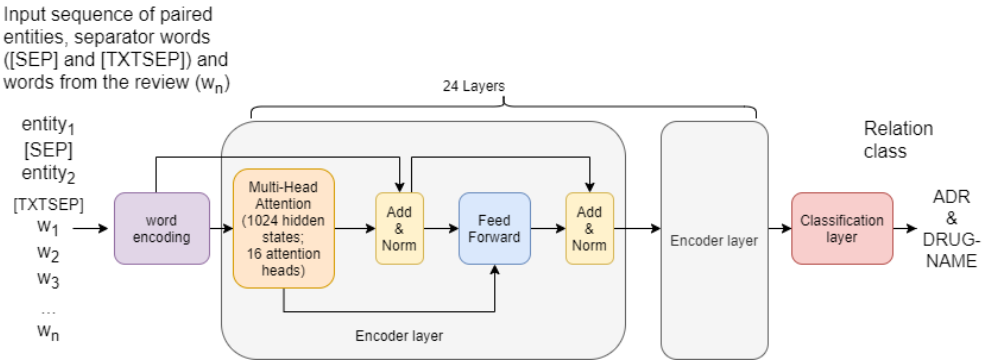

Figure 2. Conceptual scheme of our approach to the task of relation extraction based on the language model XLM-RoBERTa

The following text representation variants were considered during the experiments:

1. the whole text – tokenized input text was used as an input, target entities were highlighted using special start and end tokens, for example [T_ADR], [\T_ADR] ;
2. the text of target entities only – only text of the target entities was used as input data;
3. the text of target entities and the text between target entities;
4. the text of target entities and the whole text.

Depending on the variant of the textual representation, a single text sequence was formed. "Entities highlighting" is a way to highlight parts of texts with high importance in a way that the model "notices" them.

Example of a single text sequence for variant No. 1:

«[CLS].... [T_DRUG]"Orvirem"[\T_DRUG] - We have an allergy to it! TEXT We have a severe allergy after the first day of taking it. Moreover, the boy (3.5 years old) had no allergies to any drugs before. In the morning he woke up covered in [T_ADR]red spots[\T_ADR]. I immediately gave him zyrtek... »

Example of a single text sequence for variant No. 4:

[CLS]«text of first target entity»[SEP]«text of second target entity»[TXTSEP]«whole text»

In these cases several special tokens were used:

- [SEP] – separation token that is placed between two target entities;
- [TXTSEP] – separation token that is placed between a pair of entities and the text.
- [T_DRUG], [T_ADR], [\T_DRUG], [\T_ADR] – start and end tokens of Drugname and ADR entities;

Potentially, this way of organizing the input data makes it possible to build a more informative vector representation due to the Attention mechanism inside the Transformer layers, and facilitates solving the problem in a classification formulation. Previously the effectiveness of such text representation was demonstrated in the paper [19].

As it was mentioned before, there are many degrees of freedom in such models that require consideration in order to achieve higher accuracy, in scope of the current research the following options were analyzed:

- maximum input sequence length (in tokens);
- learning rate;
- batch size;
- maximum learning epoch number;
- learning rate decay technique [20];
- early stopping technique [21].

3.2.2. Other machine learning methods

Basic machine learning methods perform decently in many applications [22][23][24]. These methods are highly efficient in terms of computational complexity. Due to that fact it is possible to search for the optimal set in an extensive space of hyperparameters and to test hypotheses relatively quickly.

The first goal of using basic machine learning methods was to obtain a strong baseline for relation extraction of medical entities in Russian language that exceeds "Dummy" models' results.

As a text data representation for the baseline on basic machine learning methods concatenation of a frequency features (tf-idf) of the character n-grams of the target entities was used. The size of the n-gram n and the frequency filter of tf-idf were considered as the hyperparameters to tune during the experiments.

The second goal of using basic machine learning methods was to check if the information about the entities' text is sufficient to achieve competitive accuracy for the task.

The following methods were used during the experiments with basic machine learning:

- Logistic regression [25] – a basic linear model for text classification using a logistic function to estimate the probability of an example to be of a certain class;
- Support vector machine [26] – a linear model based on building a hyperplane that maximizes the margin between two classes;
- Multinomial Naive Bayes model [27] – a popular solution for baselines in such text analysis tasks as spam filtering or text classification. It performs text classification based on words'/n-grams' co-occurrence probability;
- Gradient Boosting [28] – a strong decision tree-based ensemble model, which iteratively "boosts" the result of each tree by building a next tree, that should classify examples that the previous tree did not classify correctly.

Also for comparison the RuBERT [29] language model was considered, which is a BERT [30] model with 12 layers, 768 hidden neurons each, 12 attention heads, 180M parameters. RuBERT was trained on the Russian part of Wikipedia and news data. When solving the problem, the language model is used to form a vector representation of the text, which is fed into the linear layer. The output activities of the linear layer are used to determine if there is a relationship between the pair of entities fed to the input.

3.2.3. Dummy models

"Dummy" models were considered to be the low-level baseline. Such models generate labels randomly or according to some simple principle. The following methods were checked as methods for "dummy" classification:

- most frequent class labeling – each prediction is the most frequent class in the dataset (in case of extraction of relations between adverse reaction and medication in Russian Drug Review dataset it counts each input example as an example with relation);
- uniform random labeling – labels are predicted randomly according to a uniform probability distribution, without taking into account any characteristics of the input dataset;
- stratified random labeling – labels are predicted randomly, but unlike the previous option, the probability distribution is similar to the one in the training data.

The accuracy of the "dummy" methods based on the random label generation was averaged over 100 launches in order to operate with more stable results and prevent the occurrence probability of random outliers.

4. Experiments

4.1. Accuracy metric

The performance of relation extraction is estimated by the f1-macro metric, in which the f1 score is calculated separately for each class:

$$\text{f1score} = \frac{2 \cdot P \cdot R}{P + R}, \tag{1}$$

$$P = \frac{TP}{TP + FP}, \tag{2}$$

$$R = \frac{TP}{TP + FN}. \tag{3}$$

Here P is precision, the proportion of correctly predicted objects of the class A under consideration as compared to the number of objects that the model assigned to the class under consideration; R is recall, the proportion of correctly predicted items of the class under consideration to the real number of items of the class under consideration; TP is the number of *true positive* instances, the number of relations of class A correctly identified by the model; FP is the number of *false positive* examples, the number of relations assigned to class A while actually having a different class; FN is *false negative*, the number of relations that actually have class A while being incorrectly assigned to a different class by the model.

The overall performance of the model is estimated by averaging the f1-score over the two classes. This method of averaging allows to assess on a common basis classes containing different numbers of relations.

4.2. Selection of the model features and hyperparameters

In these experiments we used a subset of RDRS that contains texts with ADR and Drugname entities only. The following experimental setup was used:

- Fixed stratified split into training (80%) and testing (20%) sets; In order to avoid overfitting, entity pairs from each review all go either to the training set or to the testing set, but no review is split between the sets;
- Hyperparameters of the language model fine-tuning were chosen so that to maximize the accuracy (by the f1-macro metric) on the validation part of the training set;
- Language model used early stopping and learning rate decay (Experiments show the positive effect of such techniques on model accuracy);

Experiments on language models were carried out using computing cluster node with the following configuration: CPU Intel® Xeon™ E5-2650v2 (2.6 GHz) x 8, 128 Gb RAM, NVIDIA Tesla V100 (16 Gb).

The hyperparameters of the language models were searched manually in consequential experiments with the analysis of train and validation loss during the training phase. Thus, the language model's hyperparameters were determined based on the validation accuracy, without taking into account the accuracy on the test set.

4.3. Estimation of effectiveness of selected methods

In that case, a part of the RDRS containing the texts of reviews with multiple contexts was used. The calculations were performed using cross-validation with the data divided into 5 parts. Thus, at each iteration of the cross-validation 80% of the texts were used for fine-tuning the model and 20% – for testing. Accuracy was also estimated using f1-score metric for all positive and negative classes. Moreover, f1-macro was calculated to obtain the general accuracy estimation.

For the most complete analysis of the model's performance, we compared the accuracy of different machine learning models in terms of complexity and type, as well as a classifier based on the probability distribution of positive and negative examples of the pairs of entities in question (Stratified random labeling).

"Dummy" models and basic machine learning method experiments were carried out on a local machine with the following configuration: CPU Intel® Core™ i5-7400 @ 3.00GHz × 4, 16 Gb RAM. The experiments with language models were performed on the same equipment as the experiments in the previous section.

The programming language python 3.8 and software libraries numpy [31], sklearn [32], pytorch [33] and simpletransformers [34] were used for software implementation of the described method. As part of a series of experiments, the parameters of python random number generator, as well as the random number generators of numpy, sklearn, and pytorch libraries were fixed to ensure repeatability of the experiments.

5. Results

5.1. Comparison of the model features and hyperparameters

This section compares the results of experiments on identification of entity relations using XLM-RoBERTa-large-sag and XLM-RoBERTa-sag depending on the input representation. Table 5 demonstrates the results of the experiments with different text representation methods.

Table 5. A comparison of language model accuracies with different methods of text representation. "LM-base" stands for XLM-RoBERTa-base-sag, "LM-large" for XLM-RoBERTa-large-sag.

Text representation	LM-base, f1-macro	LM-large, f1-macro
Text of target entities only	0.75	0.76
Whole text with highlighting target entities	0.78	0.82
Text of target entities and text between them	0.81	0.80
Text of target entities and the whole text	**0.91**	**0.95**

The Table 5 shows that both the information about the target entities separated from the text and the entire text are important to achieve high accuracy and to overcome basic machine learning methods. In case of RDRS dataset part with ADR-Drugname relations the input representation as an entity-only text concatenated with the whole text makes it possible to achieve the macro-averaged f1-score value of 95%.

The results in the table are the best obtained during the set of experiments with different hyperparameters' values. Resulting values for XLM-RoBERTa-base-sag are:

- maximum input length – 512;
- early stopping active;
- learning rate – 0.00005;
- batch size – 32;
- maximum epochs – 10;
- learning rate decay active;

Resulting hyperparameters' values for XLM-RoBERTa-large-sag are:

- maximum input length – 512;
- early stopping active;
- learning rate – 0.00001;
- batch size – 8 (there wasn't enough memory for bigger batch size with XLM-RoBERTa-large);
- maximum epochs – 10;
- learning rate decay active;

5.2. Estimation of relation extraction efficiency

As a result of the conducted experiments on the RDRS part with multi-context the accuracy of solving the problem of determining the relationships between pharmacologi-

cally significant named entities using the developed method based on the XLM-RoBERTa language model was estimated. The accuracy on the f1-score metric averaged over five folds of the cross-validation for each class of relations and its comparison with the basic accuracy based on the simplest classifiers is given in Table 6. In case of using basic machine learning methods, the input information consisted of the target entity pairs encoded with tf-idf. The best results presented in the table were obtained using the tf-idf method to encode n-grams of 3-8 characters.

Table 6. Accuracy of the models for the relation extraction task on the 908 multicontext reviews of the RDRS dataset

Methods	ADR & Drugname		Drugname & Diseasename		Drugname & Source Info Drug		Diseasename & Indication	
	pos	neg	pos	neg	pos	neg	pos	neg
Ourmodel	**92.7**	**91.1**	**89.9**	**76.2**	**92.9**	**82.7**	**87.1**	**44.3**
	91.9		**83.05**		**87.8**		**65.7**	
RuBERT	88.8	76.2	86.1	66.2	89.4	72.6	85.7	21.6
	82.5		76.15		81		53.65	
LinearSVM	72.8	45.0	75.6	44.9	77.9	45.2	83.2	24.4
	58.9		60.25		61.55		53.8	
MultinomialNaive Bayes	66.3	33.8	68.8	26.1	73.4	14.3	80.2	5.4
	50.05		47.45		43.85		42.8	
Stratified RandomGeneration	66.5	31.8	66.5	33.3	69.8	32.9	77.8	22.0
	49.15		49.9		51.35		49.9	

As follows from this table, the developed model determines a set of 4 relationships under consideration with the following accuracy (according to f1-score metrics): between adverse drug reactions and drugs (ADR–Drugname) 92.7%, between drugs and diseases (Drugname–Diseasename) 89.%, between a drug and its source of information (Drugname–SourceInfoDrug) 92.9%, between diseases and symptoms (Diseasename–Indication) 87.1%. This is 43.5%,40%,41.5%,38.2% higher than Dummy Classifier accuracy and higher than RuBERT accuracy by 3.9%,3.8%,3.5%,2.1% respectively. At the same time, for the non-coupling identification class, the accuracies are more volatile and reach lower values of 91.1%,76.2%,82.7%,44.3%.However, they exceed the Dummy Classifier accuracy by 59.3%,42.9%,49.8%,22.3% and RuBERT by 14.9%,10.0%,20.1%,22.7% respectively. On average, the accuracy of the developed model with cross-validation estimation exceeds the RuBERT by 7.8% and is equal to 82.1%.

6. Discussion

The results of this work show that the accuracy of entity relation identification depend on the input text representation. Using the neural network model XLM-RoBERTa-large-sag, selected on the preliminary stage, with the text representation in the form of target entities text fragment followed by the whole text, we received accuracy high enough for the task in view. But it is worth mentioning the volatility of accuracy depending on the type of relation. It is explained by the varying numbers of relations of different types, and may be corrected in the future by enlarging the corresponding parts of the context-labeled dataset. Overall, the received results evaluate the state-of-the-art accuracy level for the task of pharmacological entity relation identification in Russian-language reviews text and may be viewed as a basis for a future task of automated analysis of the meaning of reviews.

Author Contributions: Conceptualization, Sboev and Rybka; methodology, Sboev and Moloshnikov; software, Selivanov, Rylkov, Moloshnikov, Gryaznov; validation, Rylkov and Selivanov; investigation, Sboev, Selivanov, Rylkov, Moloshnikov; resources, Sboev, Rybka; data curation, Sboev, Sboeva and Gryaznov; writing—original draft preparation, Rybka, Selivanov, Sboev; writing—review and editing, Sboev, Rybka, Selivanov; visualization, Gryaznov and Rylkov; supervision, Sboev; project administration, Rybka; funding acquisition, Sboev

Funding: This work has been supported by the Russian Science Foundation grant No. 20-11-20246.

Data Availability Statement: Data can be obtained by sending a request from the website of our project: https://sagteam.ru/en/med-corpus/; Models will be presented on the page of our team on the huggingface repository: https://huggingface.co/sagteam; Code will be prepared and uploaded to the github repository https://github.com/sag111

Acknowledgments: 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/.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sboev, A.; Sboeva, S.; Gryaznov, A.; Evteeva, A.; Rybka, R.; Silin, M. A neural network algorithm for extracting pharmacological information from russian-language internet reviews on drugs. Journal of Physics: Conference Series. IOP Publishing, 2020, Vol. 1686, p. 012037.
2. Sboev, A.; Sboeva, S.; Moloshnikov, I.; Gryaznov, A.; Rybka, R.; Naumov, A.; Selivanov, A.; Rylkov, G.; Ilyin, V. An analysis of full-size Russian complexly NER labelled corpus of Internet user reviews on the drugs based on deep learning and language neural nets, 2021, [arXiv:cs.CL/2105.00059].
3. Asada, M.; Miwa, M.; Sasaki, Y. Using Drug Descriptions and Molecular Structures for Drug-Drug Interaction Extraction from Literature. *Bioinformatics* **2021**.
4. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: Pretrained Language Model for Scientific Text. EMNLP, 2019, [arXiv:1903.10676].
5. Eberts, M.; Ulges, A. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. In *ECAI 2020*; IOS Press, 2020; pp. 2006–2013.
6. Gurulingappa, H.; Rajput, A.M.; Roberts, A.; Fluck, J.; Hofmann-Apitius, M.; Toldo, L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics* **2012**, *45*, 885 – 892. Text Mining and Natural Language Processing in Pharmacogenomics, doi:https://doi.org/10.1016/j.jbi.2012.04.008.
7. Patrick, J.; Li, M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association* **2010**, *17*, 524–527.
8. Anick, P.; Hong, P.; Xue, N.; Anick, D. I2B2 2010 challenge: machine learning for information extraction from patient records. Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2, 2010.
9. Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; Uzuner, O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association* **2019**, *27*, 3–12. doi:10.1093/jamia/ocz166.
10. Bruches, E.; Pauls, A.; Batura, T.; Isachenko, V. Entity Recognition and Relation Extraction from Scientific and Technical Texts in Russian. 2020 Science and Artificial Intelligence conference (SAI ence). IEEE, 2020, pp. 41–45.
11. Ivanin, V.; Artemova, E.; Batura, T.; Ivanov, V.; Sarkisyan, V.; Tutubalina, E.; Smurov, I. Rurebus-2020 shared task: Russian relation extraction for business. Computational Linguistics and Intellectual Technologies, 2020, pp. 416–431.
12. Bondarenko, I.; Berezin, S.; Pauls, A.; Batura, T.; Rubtsova, Y.; Tuchinov, B. Using Few-Shot Learning Techniques for Named Entity Recognition and Relation Extraction. 2020 Science and Artificial Intelligence conference (SAI ence). IEEE, 2020, pp. 58–65.
13. Gordeev, D.; Davletov, A.; Rey, A.; Akzhigitova, G.; Geymbukh, G. Relation extraction dataset for the russian language. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog"[Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]. Moscow, Russia, 2020.
14. Loukachevitch, N.; Artemova, E.; Batura, T.; Braslavski, P.; Denisov, I.; Ivanov, V.; Manandhar, S.; Pugachev, A.; Tutubalina, E. NEREL: A Russian Dataset with Nested Named Entities and Relations. *arXiv preprint arXiv:2108.13112* **2021**.
15. Wu, S.; He, Y. Enriching pre-trained language model with entity information for relation classification. Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 2361–2364.
16. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* **2019**.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017, Vol. 30, pp. 5998–6008.
18. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226* **2018**.

19. Sboev, A.; Selivanov, A.; Rybka, R.; Moloshnikov, I.; Rylkov, G. Evaluation of Machine Learning Methods for Relation Extraction Between Drug Adverse Effects and Medications in Russian Texts of Internet User Reviews.
20. Smith, L.N. Cyclical learning rates for training neural networks. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, pp. 464–472.
21. Caruana, R.; Lawrence, S.; Giles, L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems* **2001**, pp. 402–408.
22. Sahoo, K.S.; Tripathy, B.K.; Naik, K.; Ramasubbareddy, S.; Balusamy, B.; Khari, M.; Burgos, D. An evolutionary SVM model for DDOS attack detection in software defined networks. *IEEE Access* **2020**, *8*, 132502–132513.
23. Chun, P.j.; Izumi, S.; Yamane, T. Automatic detection method of cracks from concrete surface imagery using two-step light gradient boosting machine. *Computer-Aided Civil and Infrastructure Engineering* **2021**, *36*, 61–72.
24. Xu, F.; Pan, Z.; Xia, R. E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management* **2020**, *57*, 102221.
25. Hosmer Jr, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied logistic regression*; Vol. 398, John Wiley & Sons, 2013.
26. Suykens, J.A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters* **1999**, *9*, 293–300.
27. Rish, I.; others. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, Vol. 3, pp. 41–46.
28. Mason, L.; Baxter, J.; Bartlett, P.; Frean, M. Boosting algorithms as gradient descent in function space. Proc. NIPS, 1999, Vol. 12, pp. 512–518.
29. Kuratov, Y.; Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2019, pp. 333–339.
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
31. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M.H.; Brett, M.; Haldane, A.; del Río, J.F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T.E. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. doi:10.1038/s41586-020-2649-2.
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
33. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. Pytorch: An imperative style, high-performance deep learning library. 33rd Conference on Neural Information Processing Systems, 2019, Vol. 32, *Advances in neural information processing systems*, pp. 8026–8037.
34. Rajapakse, T.C. Simple Transformers. https://github.com/ThilinaRajapakse/simpletransformers, 2019.