RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark

Tatiana Shavrina[1,2], Alena Fenogenova[1], Anton Emelyanov[1,3], Denis Shevelev[1], Ekaterina Artemova[2], Valentin Malykh[4], Vladislav Mikhailov[1,2], Maria Tikhonova[1,2], Andrey Chertok[1] and Andrey Evlampiev[1]

[1]Sberbank / Moscow, Russia
[2]National Research University Higher School of Economics / Moscow, Russia
[3]Moscow Institute of Physics and Technology / Moscow, Russia
[4]Huawei / Moscow, Russia

Abstract

In this paper, we introduce an advanced Russian general language understanding evaluation benchmark – RussianGLUE.
Recent advances in the field of universal language models and transformers require the development of a methodology for their broad diagnostics and testing for general intellectual skills - detection of natural language inference, commonsense reasoning, ability to perform simple logical operations regardless of text subject or lexicon. For the first time, a benchmark of nine tasks, collected and organized analogically to the SuperGLUE methodology (Wang et al., 2019), was developed from scratch for the Russian language. We provide baselines, human level evaluation, an open-source framework for evaluating models and an overall leaderboard of transformer models for the Russian language.
Besides, we present the first results of comparing multilingual models in the adapted diagnostic test set and offer the first steps to further expanding or assessing state-of-the-art models independently of language.

1 Introduction

With the development of technologies for text processing and then deep learning methods for obtaining better text representation, language models went through the increasingly advanced stages of natural language modelling.

Modern scientific methodology is beginning to gradually explore universal transformers as an independent object of study - furthermore, such models show the ability to extract causal relationships in texts (natural language inference), common sense and world knowledge and logic (textual entailment), to generate coherent and correct texts. An actively developing field of model interpretation develops testing procedures comparing their performance to a human level and even the ability to reproduce some mechanisms of human brain functions.

NLP is gradually absorbing all the new areas responsible for the mechanisms of thinking and the theory of artificial intelligence.

Benchmark approaches are being developed, testing general intellectual "abilities" in a text format, including complex input content, but having a simple output format. Most of these benchmarks (for more details see Section 2) make the development of machine intelligence anglo-centric, while other, less widespread languages, in particular Russian, have other characteristic linguistic categories to be tested.

In this paper, we expand the linguistic diversity of the testing methodology and present the first benchmark for evaluating universal language models and transformers for the Russian language, together with a portable methodology for collecting and filtering the data for other languages.

The contribution of RussianGLUE is two-fold. First, it provides nine novel datasets for the Russian language covering a wide scope of NLU tasks. The choice of the tasks are justified by the design of prior NLU benchmarks (Wang et al., 2018, 2019). Second, we evaluate two widely used deep models to establish baselines.

The remainder is structured as follows. We overview multiple prior works on developing NLU benchmarks, including those designed for languages other than English, in Section 2. Section 3.1 lists the tasks and novel datasets, proposed for the Russian NLU. Section 4 presents with the baselines, established for the tasks, including a human level baseline. We overview compare achieved results in Section 2 to the current state of English NLU. We discuss future work directions and emphasize the importance of NLU benchmarks for languages other than English in Section 6. Section 7 concludes.

2 Related Work

Several benchmarks have been developed to evaluate and analyze word and sentence embeddings over the past few years.

SentEval (Conneau and Kiela, 2018) is one of the first frameworks intended to evaluate the quality of sentence embeddings. A twofold set of transfer tasks is used to assess the generalization power of sentence embedding models. The transfer tasks comprise downstream tasks, in which the sentence embedding is used as a feature vector, and probing tasks, which are aimed to evaluate the capability of sentence embeddings to encode linguistic properties. The choice of the downstream tasks is limited to sentiment classification, natural language inference, paraphrase detection and image captioning tasks. The probing tasks are meant to analyse morphological, syntactical and semantical information encoded in sentence embeddings.

The General Language Understanding Evaluation (GLUE) (Wang et al., 2018) benchmark is a collection of tools for evaluating the performance of language models across a diverse set of existing natural language understanding (NLU) tasks, adopted from different sources. These tasks are divided into two parts: single sentence classification tasks and sentence pair classifications tasks subdivided further into similarity and inference tasks. GLUE also includes a hand-crafted diagnostic test, which probes for complex linguistic phenomena, such as the ability of the model to express lexical semantics and predicate-argument structure, to pose logical apparatus and knowledge representation. GLUE is recognized as a de-facto standard benchmark to evaluate transformer-derived language models. Last but not least GLUE informs on human baselines for the tasks, so that not only submitted models are compared to the baseline, but also to the human performance. The SuperGLUE (Wang et al., 2019) follows GLUE paradigm for language model evaluation based on NLU tasks, providing with more complex tasks, of which some require reasoning capabilities and some are aimed at detecting ethical biases. A few recent projects reveal that GLUE tasks may be not sophisticated enough and do not require much tasks-specific linguistic knowledge (Kovaleva et al., 2019; Warstadt et al., 2019). Thus SuperGLUE benchmark, being more challenging, becomes much more preferable for evaluation of language models.

decaNLP (McCann et al., 2018) widens the scope for language model evaluation by introducing ten disparate natural language tasks. These tasks comprise not only text classification problems, but sequence tagging and sequence transformation problems. The latter include machine translation and text summarization, while the former include semantic parsing and semantic role labelling. Although decaNLP along with the associated research direction focuses on multi-task learning as a form of question answering, it supports zero-shot evaluation.

To evaluate models for languages other than English, several monolingual benchmarks were developed, such as FLUE (Le et al., 2019) and CLUE (Liang, 2020), being French and Chinese versions of GLUE. These benchmarks include a variety of tasks, ranging from part-of-speech tagging and syntax parsing to machine reading comprehension and natural language inference.

To the best of our knowledge, LINSPECTOR (Eichler et al., 2019) is a first multi-lingual benchmark for evaluating the performance of language models. LINSPECTOR offers 22 probing tasks to analyse for a single linguistic feature such as case marking, gender, person, or tense for 52 languages. A part of these 22 probing tasks are static, i.e. are aimed at evaluation of word embeddings, and the rest are contextual and should be used to evaluate language models. Released in early 2020 two multilingual benchmarks, (Liang et al., 2020) and XTREME (Hu et al., 2020), aim at evaluation of cross-lingual models. XGLUE includes 11 tasks, which cover both language understanding and language generation problems, for 19 languages. XGLUE provides with several multilingual and bilingual corpora that allow of cross-lingual model training. As for the Russian language, XGLUE provides with four datasets for POS tagging, a part of XNLI (Conneau et al., 2018) and two datasets, crawled from commercial news website, used for news classification and news headline generation. XTREME consists of nine tasks which cover classification, sequence labelling, question answering and retrieval problems for 40 languages. Almost a half of the datasets were translated from English to the target languages with the help of professional translators. XTREME offers for the Russian language five datasets, including NER and two question-answering datasets. Both XGLUE and XTREME offer tasks that are much simpler than SuperGLUE and are aimed at evaluation of cross-lingual models rather than at comparison of mono-lingual models in similar setups. Thus the need for novel datasets targeted at mono-lingual model evaluation for languages other than English is still not eliminated.

3 RussianGLUE Overview

We have intenooed to have the same task set in the framework as one in the SuperGLUE. There is no one-to-one mapping, but the corpora we use could be considered close to the specified tasks in the SuperGLUE framework.

We divided the tasks into six groups, covering the general diagnostics of language models and different core tasks: common sense understanding, natural language inference, reasoning, machine reading and world knowledge.

3.1 Tasks

The tasks description is provided below. The samples from the tasks are presented at figs. 1 to 7.

3.1.1 Diagnostics

LiDiRus: Linguistic Diagnostic for Russian is a diagnostic dataset that covers a large volume of linguistic phenomena, while allowing you to evaluate information systems on a simple test of textual entailment recognition. This dataset was translated from English to Russian with the help of professional translators and linguists to ensure that the desired linguistic phenomena remain. This dataset corresponds to AX-b dataset in SuperGLUE benchmark.

3.1.2 Common Sense

RUSSE: Word in context is a binary classification task, based on word sense disambiguation problem. Given two sentences and a polysemous word, which occurs in both sentences, the task is to determine, whether the word is used in the same sense in both sentences, or not. For this task we used the Russian word sense disambiguation dataset **RUSSE** (Panchenko et al., 2015) and converted it into WiC dataset format from SuperGLUE.

RUSSE

Context 1: *Бурые ковровые дорожки заглушали шаги.* **Context 2:** *Приятели решили выпить на дорожку в местном баре.*
Sense match: `False`

Figure 1: A sample from RUSSE dataset.

PARus: The choice of Plausible Alternatives for Russian language evaluation provides researchers with a tool for assessing progress in open-domain commonsense causal reasoning. Each question in PARus is composed of a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. The correct alternative is randomized so that the expected performance of randomly guessing is 50%. PARus is constructed as a translation of COPA dataset from SuperGLUE and edited by professional editors. The data split from COPA is retained.

PARus

Premise: *Гости вечеринки прятались за диваном.*
Question: *Почему это произошло?*
Alternative 1: *Это была вечеринка-сюрприз.*
Alternative 2: *Это был день рождения.*
Correct Alternative: `1`

Figure 2: A sample from PARus dataset.

3.1.3 Natural Language Inference

TERRa: Textual Entailment Recognition for Russian is a dataset which is devoted to capture textual entailment. The task of textual entailment has been proposed recently as a generic task that captures major semantic inference needs across many NLP applications, such as Question Answering, Information Retrieval, Information Extraction, and Text Summarization. This task requires to recognize, given two text fragments, whether the meaning of one text is entailed (can be inferred) from the other text. The corresponding dataset in SuperGLUE is RTE, which in its place is constructed from NIST RTE challenge series corpora. To collect TERRa we filtered out the large scale Russian web-corpus, Taiga (Shavrina and Shapovalova, 2017) with a number of rules to extract suitable sentence pairs and manually corrected them. The rules had the following structures: there should be a mental verb in the first sentence and the second sentence should be attached to the first one by a subordinate conjunction. To ensure the literary language of the extracted sentences, we processed only news and fiction parts of Taiga and made sure, that the sentences contain only frequently used words (i.e. number instances per million, IPM is higher than 1). The word frequencies were estimated according to Russian National Corpus[1].

RCB: The Russian Commitment Bank is a corpus of naturally occurring discourses whose final sentence contains a clause-embedding predicate

[1] http://www.ruscorpora.ru/new/en/

TERRA

Text: *Автор поста написал в комментарии, что прорвалась канализация.*
Hypothesis: *Автор поста написал про канализацию.* **Entailment:** `True`

Figure 3: A sample from TERRa dataset.

under an entailment canceling operator (question, modal, negation, antecedent of conditional). Similarly to the design of TERRa dataset, we filtered out Taiga with a number of rules and manually post processed the extracted passages. Final labelling was conducted by three of the authors. This dataset corresponds to CommonBank dataset.

RCB

Text: *Сумма ущерба составила одну тысячу рублей. Уточняется, что на место происшествия выехала следственная группа, которая установила личность злоумышленника. Им оказался местный житель, ранее судимый за подобное правонарушение.*
Hypothesis: *Ранее местный житель совершал подобное правонарушение.*
Entailment: `Yes`

Figure 4: A sample from RCB dataset.

3.1.4 Reasoning

RWSD: Winograd Schema task is devoted to coreference resolution in specifically designed experiment, where reference could be resolved only using the common sense. The Russian Winograd Schema Dataset (**RWSD**) is constructed as translation of the Winograd Schema Challenge[2].

RWSD

Text: *Кубок не помещается в коричневый чемодан, потому что он слишком большой.* **Coreference:** `True`

Figure 5: A sample from RWSD dataset.

3.1.5 Machine Reading

MuSeRC: Russian Multi-Sentence Reading Comprehension is a reading comprehension challenge in which questions can only be answered by taking into account information from multiple sentences. The dataset is the first to study multi-sentence inference at scale, with an open-ended set of question types that requires reasoning skills. The task is actually a binary classification, whether the answer to the question is correct or not. Each example consists of numerated passage, question and answers. Our dataset contains approximately 6000 questions for more than 800 paragraphs across 5 different domains, namely: 1) elementary school texts, 2) news, 3) fiction stories, 4) fairy tales, 5) brief annotations of TV series and books. First, we have collected open sources data from different domains and automatically preprocessed them, filtered only those paragraphs that corresponds to the following parameters: 1) paragraph length 2) number of named entities 3) number of coreference relations. Afterwords we have checked the correct splitting on sentences and numerate each of them. Next, in Toloka[3] we have generated the crowd sourcing task to get the following information: 1) generate questions 2) generate answers 3) check that to solve every question a human needs more than one sentence in the text. Collecting the dataset we adhere the principles of MultiRC (Khashabi et al., 2018): a) We exclude any question that can be answered based on a single sentence from a paragraph; b) Answers are not written in the full match form in the text; c) Answers to the questions are independent from each other.

MuSeRC

Paragraph: *(1) Мужская сборная команда Норвегии по биатлону в рамках этапа Кубка мира в немецком Оберхофе выиграла эстафетную гонку. (2) Вторыми стали французы, а бронзу получила немецкая команда. (3) Российские биатлонисты не смогли побороться даже за четвертое место, отстав от норвежцев более чем на две минуты. (4) Это худший результат сборной России в текущем сезоне. (5) Четвёртыми в Оберхофе стали австрийцы. (6) В составе сборной Норвегии на четвёртый этап вышел легендарный Уле-Эйнар Бьорндален. (7) Впрочем, Норвегия с самого начала гонки была в числе лидеров, успешно проведя все четыре этапа. (8) За сборную России в Оберхофе выступали Иван Черезов, Антон Шипулин, Евгений Устюгов и Максим Чудов. (9) Гонка не задалась уже с самого начала: если на стрельбе из положения лежа Черезов был точен, то из положения стоя он допустил несколько промахов, в результате чего ему пришлось бежать один дополнительный круг. (10) После этого отставание российской команды от соперников только увеличивалось. (11) Напомним, что днем ранее российские биатлонистки выиграли свою эстафету. (12) В составе сборной России выступали Анна Богалий-Титовец, Анна Булыгина, Ольга Медведцева и Светлана Слепцова. (13) Они опередили своих основных соперниц - немок - всего на 0,3 секунды.*
Question: *На сколько секунд женская команда опередила своих соперниц?*
Candidate answers:
Всего на 0,3 секунды. (`T`),
На 0,3 секунды. (`T`),
На секунду. (`F`),
На секунды. (`F`)

Figure 6: A sample from MuSeRC dataset.

RuCoS: Russian reading comprehension with Commonsense reasoning is a large-scale dataset for machine reading comprehension requiring commonsense reasoning. The dataset construction is based on ReCoRD methodology (Zhang et al., 2018). RuCoS consists of passages and cloze-style queries automatically generated from Russian news articles, namely Lenta[4] and Deutsche Welle[5]. Each sample from the dev and test sets was validated by crowd workers. The answer to each query is a text

[2] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

[3] https://toloka.yandex.ru

[4] https://lenta.ru/

[5] https://www.dw.com/ru/

span that corresponds to one or more referents of the answer entity in the context. The answer entity may be expressed by an abbreviation, an acronym or a set of surface forms. Hence, the task requires understanding of rich inflectional morphology and lexical variability of Russian. The goal of RuCoS is to test a machine's ability to infer the answer based on the commonsense reasoning and knowledge.

RuCoS

Paragraph: Мать двух мальчиков, брошенных отцом в московском аэропорту *Шереметьево*, забрала их. Об этом сообщили *ТАСС* в пресс-службе министерства образования и науки *Хабаровского края*. Сейчас младший ребенок посещает детский сад, а старший ходит в школу. В учебных заведениях с ними по необходимости работают штатные психологи. Также министерство социальной защиты населения рассматривает вопрос о бесплатном оздоровлении детей в летнее время. Через несколько дней после того, как *Виктор Гаврилов* бросил своих детей в аэропорту, он явился с повинной к следователям в городе *Батайске* Ростовской области.
Query 26 января <placeholder> бросил сыновей в возрасте пяти и семи лет в Шереметьево.
Correct Entities: `Виктор Гаврилов`

Figure 7: A sample from RuCoS dataset.

3.1.6 World Knowledge

DaNetQA: This question-answering corpus follows BoolQ (Clark et al., 2019) design: it comprises natural yes/no questions. Each question is paired with a paragraph from Wikipedia and an answer, derived from the paragraph. The task is to take both the question as input and a paragraph and come up with a yes/no answer, i.e. to produce a binary output. DaNetQA was collected in a few steps: 1) we used crowd workers to compose candidate yes/no questions; 2) we used Google API to retrieve relevant Wikipedia pages by treating each question as a search query; 3) we queried a pretrained BERT-based model for SQuAD (Kuratov and Arkhipov, 2019) to extract relevant paragraphs from Wikipedia pages, using candidate questions; 4) finally, we used crowd workers to evaluate each question and paragraph pair and provide the desired yes/no answers. We ensure high quality of the dataset by using a high overlap for annotation at the last step and a number of control gold-standard control questions, labelled by two of the authors. The core difference of DaNetQA to BoolQ is that some question may occur multiple times in the dataset, as at the step 3) we may retrieve more than one relevant paragraph. To make the dataset more challenging, we admit contradictory answers to a question if these answers are implied from the passages.

3.1.7 Statistics for the Tasks

Table 1 below presents the characteristics of the collected datasets - examples partitioning by train/val/test, as well as the total volume in tokens and sentences. As one can see, the size of the RuCoS task significantly exceeds the rest of the tasks due to the articles included in the task.

Task	Samples	Sents	Tokens
LiDiRus	0/0/1104	2210	$3.6 \cdot 10^4$
Common Sense			
RUSSE	19845/8508/12151	90862	$1.1 \cdot 10^6$
PARus	500/100/400	1000	$5.4 \cdot 10^3$
NLI			
TERRa	2616/307/3198	13706	$2.53 \cdot 10^5$
RCB	438/220/348	2715	$3.7 \cdot 10^4$
Reasoning			
RWSD	606/204/154	1541	$2.3 \cdot 10^3$
Machine Reading			
MuSeRC	500/100/322	12805	$2.53 \cdot 10^5$
RuCoS	72193/4370/4147	583930	$1.2 \cdot 10^7$
World Knowledge			
DaNetQA	392/295/295	6231	$1.31 \cdot 10^5$

Table 1: Cumulative task statistics. The size train/validation/test splits is provided in "Samples" column.

3.2 Scoring

Following (Wang et al., 2019), we calculate scores for each of the tasks based on their individual metrics. All metrics are scaled by 100x (i.e., as percentages). These scores are then averaged to get the final score. For the tasks with multiple metrics, the metrics are averaged.

4 Experiments

4.1 Baselines

In this section, we provide a two-step baseline design. At first we have developed a naïve baseline based on the TF-IDF model (section 4.1.1), and then evaluate state-of-the-art models for Russian language (section 4.1.2).

4.1.1 Naïve Baseline

We used Scikit-learn package (Pedregosa et al., 2011) to train a TF-IDF model. We used a 20 thousand sample from Wikipedia, from Russian and English sites equally. We restricted a vocabulary to 10 thousand most common words. Then for each task set a logistic regression was trained to predict an answer.

Dataset	**Metrics**	`RuBERT`	`MultiBERT`	TF-IDF	Human
LiDiRus	MCC	0.186	0.157	0.059	**0.626**
RCB	F_1/Acc.	0.432/0.468	0.383/0.429	0.45	**0.68/0.702**
PARus	Acc	0.61	0.588	0.48	**0.982**
MuSeRC	F_1/EM	0.656/0.256	0.626/0.253	0.589/0.244	**0.806/0.42**
TERRa	Acc	0.639	0.62	0.47	**0.92**
RUSSE	Acc	**0.894**	0.84	0.66	0.747
RWSD	Acc	0.675	0.675	0.66	**0.84**
DaNetQA	Acc	0.749	0.79	0.68	**0.879**
RuCoS	F_1/EM	0.255/0.251	0.371/0.367	0.256/0.251	**0.93/0.924**
Average		0.546	0.542	0.461	0.802

Table 2: Results of the human benchmark and the baseline models. MCC stands for Matthews Correlation Coefficient; Acc - Accuracy; EM - Exact Match.

4.1.2 Advanced Baselines

We leverage two BERT-derived models as baseline. Multilingual BERT (`MultiBERT`), released by (Devlin et al., 2019), is a single language model pre-trained from monolingual corpora in 104 languages, Russian texts being a part of training data. `MultiBERT` uses a shared vocabulary for all languages. The capabilities of `MultiBERT` for zero-shot cross-lingual tasks have been recently studied by (Pires et al., 2019). Russian BERT (`RuBERT`) was trained on large-scale corpus of news and Wikipedia in Russian. To alleviate the training all weights except sub-word embeddings were borrowed from `MultiBERT`. The sub-word vocabulary was obtained from the same training corpus and the new mono-lingual embeddings were transformed from the multi-lingual ones. This allowed to incorporate longer Russian sub-word units into the vocabulary. This model is part of DeepPavlov framework (Kuratov and Arkhipov, 2019).

4.2 Human Evaluation

We include human performance estimates for all provided benchmark tasks, including the diagnostic set. We estimate human performance by hiring crowd workers via Toloka platform to re-annotate a sample from each task test set. We suggest a two step procedure: 1) a crowd worker is provided with an instruction and completes a short training phase before proceeding to the annotation phase, 2) a crowd worker that passed through the training phase solves the original test set.

For the annotation phase we ask crowd workers to annotate the full test sets except for the RUSSE and the RuCoS datasets, where we randomly sampled only 5000 and 1000 examples from the tasks' test sets, respectively. For each sample, we collect annotations from three to five crowd workers and take a majority vote to estimate human performance. In annotation phase we add control questions to prevent the crowd workers from cheating. As a result, we reject the annotations from the crowd workers that fail the training phase and do not include the results of those who achieved low performance on the control tasks. The results of human evaluation are presented in Table 2. The example of a Toloka task is provided in Appendix.

5 Results

The analysis of Table 2 can give an exact representation of the baseline model performance, which still remains significantly different from the human level. Nevertheless, the task of resolving the ambiguity of the word meaning in context (RUSSE) was solved by both monolingual and multilingual BERT at a level significantly exceeding the human one (0.89 vs 0.74). Besides, the monolingual model is showing a slightly higher quality than that of the multilingual one, especially prevailing textual entailment tasks (RCB, TERRA, PARus), disambiguating word meaning (RUSSE) and reading comprehension (MuSeRC). The multilingual model shows the most excellent result on the smallest dataset on commonsense QA task (DaNetQA) and also on commonsense-related task on machine reading (RuCoS).

We hope that our benchmark will help to excel the performance of models for the Russian language in the future, and will favour achieving comparably high results.

Can the results of a multilingual BERT on Russian and English data be considered analogous? Based on the results of the assessment, `MultiBERT` in English gets an overall score of 60.8 [6], while on RussianGLUE task set an overall score of 54.2 is achieved– 6% lower, but noting that the English benchmark includes additionally Winograd Gender Parity (Levesque et al., 2012) dataset, giving SOTA models from 90 to 93% of accuracy added to the overall assessment. In the next section, a detailed comparison of the multilingual model performance is provided.

5.1 Comparison to SuperGLUE

As mentioned in Section 3.1, the diagnostic dataset has been obtained by professional translation with preservation of the original linguistic features mentioned. Thus being said, this diagnostic data is the first of its kind that allows drawing a multilingual analogy of comparable models.

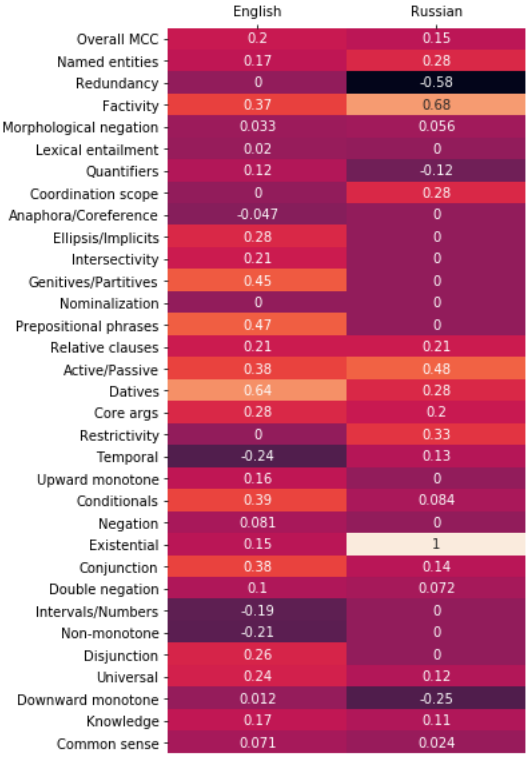

Figure 8: Russian and English Diagnostic Evaluation on Multilingual transformer, scored using Matthews' correlation (MCC).

Procedure: using the original `MultiBERT` (Devlin et al., 2019), we conducted sequential model pretraining in English and Russian using the RTE dataset, and then tested the models on the diagnostic set, as long as the task requires exactly the same format. Predictions were further scored using Matthews' correlation (MCC), and correlation for different linguistic features was computed. The results are presented in Figure 8.

First of all, it could be noticed that the English variant of the model performs slightly better and shows a higher overall correlation of 0.2 compared to 0.15 for the Russian variant. This could be due to an asymmetry of the quality of the multilingual model and its better understanding of the English language in general.

As for the models' performance in the context of different linguistic features, the results generally coincide. For those categories for which correlation is low in English, the result in Russian is in most cases poor as well (for instance, *Redundancy, Nominalization, Intervals/Numbers*). However, there exist several categories which are much better solved in English than in Russian such as *PA Structure, Ellipsis/Implicits, Genetives/Partitives, Prepositional phrases, Datives* – mostly low-level and/or syntactically driven categories, that may indicate that optimal hyperparameters of BERT architecture are much more suitable for English syntax and may not be linguistically universal. Similarly, we could find categories which show an extremely high correlation in Russian and low correlation in English (*Factivity, Coordination scope, Restrictivity* and *Existential*) – high-level logical and semantic categories.

These numbers compared to the ones for English could be explained by the fact that the language features now included in the diagnostics are not exactly linguistically universal in different languages and are mostly focused on the English language (at least those syntactic ones). Thus, for the comprehensive cross-linguistic typological analysis of possible linguistic features should be reviewed.

6 Discussion

We hope that our project will give a start to new research in the application of universal language models and transformers, including multilingual ones. Our example of an analysis of translated diagnostics shows that even in languages of the same European family (which Russian and English belong to), significant differences in the influence of linguistic categories on model performance are possible. One of the directions of the next studies, we consider detailed experiments on the influence of model parameters and language categories in data on the quality of the model in different languages.

[6] Jiant, full SuperGLUE task set

An independent problem for the English original leaderboard is that a gradual improvement in the quality of models allows us to exceed the human performance level in individual tasks, as happened with the T5 (Raffel et al., 2019) model. We expect that a similar situation will soon happen on Russian data, which means that when releasing straight off with complex SuperGLUE tasks, we will still be focused on adding tasks of a higher level of complexity in the future. Such tasks can become those that are obviously inaccessible to models for the "understanding" of long texts and documents, seq2seq tasks, tasks that require knowledge graphs.

In the further development of our leaderboard, we also see the possibility of adding an industrial assessment of models: for fair ranking and ease of use, all models could receive an estimate of the required memory resources, an estimate of performance, and so on.

7 Conclusion

In this paper we present the first benchmark on general language understanding evaluation for the Russian language. The benchmark including nine task sets is aimed to test BERT-like models for their ability to perform entailment recognition, commonsense reasoning and machine reading while denoising various linguistic features added on the level of semantics, logical and syntactic structure.

We invite developers, researchers, and AI experts to join our project. Further development of the benchmark includes areas such as evaluation of industrial performance of models on the leaderboard and multilingual diagnostics.

Acknowledgements

Ekaterina Artemova works within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project "5-100".

References

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Max Eichler, Gözde Gül Şahin, and Iryna Gurevych. 2019. LINSPECTOR WEB: A multilingual probing suite for word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 127–132, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4356–4365.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Xu Liang. 2020. Chinese glue. https://github.com/chineseGLUE/chineseGLUE.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, pages arXiv–2004.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Alexander Panchenko, Natalia Loukachevitch, Dmitry Ustalov, Denis Paperno, Christian Meyer, and Natalia Konstantinova. 2015. Russe: the first workshop on russian semantic similarity. *Dialogue*, 14:89–105.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:“taiga”. syntax tree corpus and parser. *Corpus linguitics–2017*, page 78.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating bert's knowledge of language: Five analysis methods with npis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2870–2880.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.