Unreasonable Effectiveness of Rule-Based Heuristics in Solving Russian SuperGLUE Tasks

Iazykova Tatyana
HSE University
Moscow, Russia
`tvyazykova@edu.hse.ru`

Bystrova Olga
HSE University / Sberbank
Moscow, Russia
`ovbystrova@edu.hse.ru`

Kapelyushnik Denis
HSE University
Moscow, Russia
`dmkapelyushnik@edu.hse.ru`

Kutuzov Andrey
University of Oslo
Oslo, Norway
`andreku@ifi.uio.no`

Abstract

Leaderboards like SuperGLUE are seen as important incentives for active development of NLP, since they provide standard benchmarks for fair comparison of modern language models. They have driven the world's best engineering teams as well as their resources to collaborate and solve a set of tasks for general language understanding. Their performance scores are often claimed to be close to or even higher than the human performance. These results encouraged more thorough analysis of whether the benchmark datasets featured any statistical cues that machine learning based language models can exploit. For English datasets, it was shown that they often contain annotation artifacts. This allows solving certain tasks with very simple rules and achieving competitive rankings.

In this paper, a similar analysis was done for the Russian SuperGLUE (RSG), a recently published benchmark set and leaderboard for Russian natural language understanding. We show that its test datasets are vulnerable to shallow heuristics. Often approaches based on simple rules outperform or come close to the results of the notorious pre-trained language models like GPT-3 or BERT. It is likely (as the simplest explanation) that a significant part of the SOTA models performance in the RSG leaderboard is due to exploiting these shallow heuristics and that has nothing in common with real language understanding. We provide a set of recommendations on how to improve these datasets, making the RSG leaderboard even more representative of the real progress in Russian NLU.

Keywords: leaderboards, benchmark, heuristics, rule-based, language models, natural language understanding
DOI: 10.28995/2075-7182-2021-20-302-317

Эффективность правиловых эвристик в решении тестовых сетов Russian SuperGLUE

Языкова Татьяна
НИУ ВШЭ
Москва, Россия
tvyazykova@edu.hse.ru

Быстрова Ольга
НИУ ВШЭ / Сбербанк
Москва, Россия
ovbystrova@edu.hse.ru

Капелюшник Денис
НИУ ВШЭ
Москва, Россия
dmkapelyushnik@edu.hse.ru

Кутузов Андрей
Университет Осло
Осло, Норвегия
andreku@ifi.uio.no

Аннотация

SuperGLUE и подобные ему наборы тестовых заданий для оценки решения задач по обработке естественного языка способствуют активному развитию этой области знания, так как они предлагают методологию тестирования современных моделей и подходов. Им удается привлекать к разработке автоматических решений задач на понимание языка лучшие мировые инженерно-технические группы с их ресурсами. Количественные результаты, полученные на некоторых тестовых сетах для английского языка, сравнимы с человеческими или превышают их. Всё это привело к более пристальному изучению наборов тестовых заданий на предмет наличия в них

каких-либо неучтённых статистических особенностей, на которые могут опираться языковые модели. Так, исследования, проведенные на подобных наборах тестовых заданий для английского языка, выявили наличие артефактов разметки в представленном языковом материале. Использование таких артефактов в построении простых правил позволяет решать некоторые из тестовых заданий на уровне, сравнимом с современными языковыми моделями.

В данной статье приводится похожий анализ для Russian SuperGLUE, системы оценки и рейтинга моделей для решения задач на понимание естественного языка для русского. Уязвимость данного тестового набора задач для простых эвристик была также подтверждена, а наши решения на основе правил часто превышали результаты таких небезызвестных языковых моделей, как GPT-3 и BERT. Мы предполагаем (как наиболее простое объяснение), что языковые модели, находящиеся в верхней части рейтинга Russian SuperGLUE, также в значительной степени полагаются на простые эвристики, а значит, не представляется возможным говорить о реальном понимании языка. Основываясь на результатах нашего исследования, мы составили набор рекомендаций по улучшению Russian SuperGLUE, которые позволят обеспечить более справедливую оценку русскоязычных систем, нацеленных на моделирование 'понимания' естественного языка.

Ключевые слова: рейтинг, тестовые сеты, эвристики, правила, языковые модели, понимание естественного языка

1 Introduction

These days many researchers are coming to a dreadful realisation that we are not that much advanced in natural language understanding (NLU) as we used to think. Huge Transformer-based models are crowning the SuperGLUE leaderboard [34], yet one should not trust these shining examples so fast. It has been shown in [23] that actually these models are exploiting statistical patterns related to the lack of diversity in data or class imbalances to demonstrate amazing performance without looking deeper and truly emulating natural language understanding. The danger of having such statistical cues is not in their mere presence. The core of the problem is that they are inherent to particular datasets only and therefore are hardly applicable to the language itself. It means that the systems do not really 'understand' natural language: instead, they are utilising statistical cues that are typical to these specific datasets, so the whole process comes down to simple pattern matching. Modern language models are trained on the amount of data no native speaker will hardly ever see [20]. But are they really as superior as we believe them to be, and is it even necessary to do genuine NLU to solve the test sets at this level of performance?

The issue became even more relevant now that the SuperGLUE benchmark, that was initially created for English, was adopted for Russian in the form of Russian SuperGLUE (RSG) benchmark and the corresponding leaderboard [33]. In this paper, we study the possibility to achieve results comparable to ones in the leaderboard without using any machine learning algorithms. We manually examined the datasets in order to find statistical regularities. As a result, we came up with a list of simple rule-based heuristics (for instance, label instances as 'entailment' if they contain the word 'был' 'was'). We do not have direct proofs that machine learning based models also make use of these shallow heuristics in the case of RSG. But we do know that this was confirmed to be true for the English SuperGLUE [32], and we know that deep neural nets are extremely efficient in capturing regularities useful for their objective function. Following the Occam's Razor, we argue that finding and exploiting shallow statistical cues (not necessarily the ones we found manually) is much more plausible explanation for the observed performance of pre-trained language models than the assumption that they 'understand' Russian discourse.

Moreover, we evaluated a set of trivial baselines, such as random choice, majority class and random balanced choice. The goal was to compare state-of-the-art (SOTA) results against those and to see whether cutting-edge deep learning architectures (GPT-3, BERT, etc) significantly outperform them. As we found out, this is not always the case.

1.1 Contributions

The contributions of this work can be formulated as follows:

1. We introduced a set of simple rule-based heuristics applicable to various datasets of Russian Super-GLUE benchmark[1], and evaluated their performance on the test data.
2. We evaluated the performance of even more trivial baselines (random choice, majority class, etc) on the Russian SuperGLUE tasks, to establish a lower boundary for language models' performance.

---

[1] https://github.com/tatiana-iazykova/2020_HACK_RUSSIANSUPERGLUE

3. A number of suggestions, spotted annotation errors and generally problematic or controversial cases are given for the authors of the Russian SuperGlUE benchmark, for further improvement.

2   Previous work

Leaderboards provide the NLP community with tools to evaluate language models. This competition ensures a fair ground for comparison as the models are required to solve the same tasks on a single independently curated set of data. For example, the GLUE leaderboard [11] was initially designed for English and consists of several diverse natural language understanding tasks and a diagnostic dataset with openly available labels to evaluate models.

By March 2021, the situation with the GLUE dataset is the following: 14 different models hold a higher ranking than the human performance which is equal to 87.1 [25]). The knowledge about language is considered as key to solve the GLUE or any NLU tasks, yet when the SOTA approach [7] (as of now) exceeded human performance by 3.8, the creators of this model hypothesised that it was not necessary for those specific datasets. With 14 other models outperforming the human level as well, it has soon become clear that the benchmark itself is no longer able to provide a challenging evaluation system. As a result, the authors of GLUE designed SuperGLUE [34] for a more representative analysis of the current progress in NLU. To track this progress for other languages, other researchers created language-specific benchmarks similar to GLUE and SuperGLUE, e.g. Russian SuperGLUE [33] explored in this paper or CLUE [4] for the Chinese language.

Although the SuperGLUE benchmark is more recent, its current SOTA score of 90.3 [6] also managed to surpass the human performance [34] by 0.5. As these competitions attract the world's best engineering teams with almost unlimited resources, models like T5 [9], GPT-3 [17], BERT [3] and its optimized versions like RoBERTa [29] usually hold top rankings and yet their performance scores differ by a mere fraction. These models prove their reputation by achieving scores that are very close to or even higher than human benchmark, and this is where some room for criticism appears.

Such complex models require considerable resources, raising questions about their general utility [8]. Indeed, for the majority of us the size and efficiency of a model is as important as the performance scores, and some trade-off has to be allowed. Through such discussions, e. g. [31], the NLP community attempts to increase the transparency of benchmarks. Fortunately, the leaderboards are open for changes and new functionality. For example, the MOROCCO project has been recently launched to evaluate Russian SuperGLUE models in two additional dimensions: inference speed and GPU RAM usage[2].

Although these issues are important, another question — probably a deeper one — is raised by how exactly large-scale language models are 'solving' certain NLU tasks. For example, BERT has skyrocketed the performance in many NLP tasks for English, yet if we take a closer look into its 'language skills', we might be disappointed [32]. It appears that BERT never misses an opportunity to use shallow heuristics while solving tasks on natural language inference [23, 13, 14], reading comprehension [12, 36, 2, 18], argument reasoning comprehension [26] and text classification [14].

The above-mentioned analysis is mostly English-centred, and we are truly grateful to the creators of the Russian SuperGLUE [33], since it is now possible to have a fair ground for comparing Russian NLU models. It is the first standardized set of diverse NLU benchmarks for Russian.

Some of the instances for its datasets were translated from the corresponding tasks in the SuperGLUE, while the others were collected by the RSG authors from scratch [10].

In this paper, we explore all the datasets thoroughly to test their vulnerability to shallow heuristics. The results are compared to other approaches represented in the Russian SuperGLUE leaderboard. It should be noted that the RSG has been created very recently, and the human performance of 0.811 is still at the top of the leaderboard. As of early May 2021, the highest score of 0.679 was achieved by an ensemble of Transformer models.

---

[2]https://russiansuperglue.com/performance/

3 Methodology

The Russian SuperGLUE benchmark consists of 9 datasets or tasks, that follow the GLUE and Super-GLUE methodology. Each task is designed to evaluate if a model or an approach can solve problems with the help of logic, common sense and reasoning. Data is split into training, validation and test samples. The true labels of the test set are not openly available and to evaluate a system on the test set, it is necessary to submit the predictions to the leaderboard. Currently there are two versions of Russian SuperGLUE present, namely 1.0 and 1.1; our research was based on the latest 1.1 version.

Our general approach was to identify shallow heuristics and design rule-based functions that would surpass the results achieved by the trivial baselines (majority class, random choice and random balanced choice) and potentially approach SOTA scores. Being native Russian speakers, we invested our efforts into manual exploration of each dataset. Additionally, ELI5[3], a tool to debug machine learning classifiers and explain their predictions, was applied to some of the tasks. It was used to check if any tokens are more specific to one of the classes in the dataset. Moreover, whenever the lemmatisation was needed, `pymorphy2` morphological analyzer [15] was used.

As the datasets differ significantly, there was no intention to identify a single heuristic to solve them all: we analyzed them separately. Heuristics found in the training sets were applied to the validation sets to get an idea of their performance. All of the heuristics that were proved to work on training and validation sets were combined into functions with a set of if-else statements. To determine the order of these statements, we tested different sequences empirically and chose the ones with the higher performance scores.

Finally, these rule-based functions were applied to the relevant test sets. To handle examples that did not trigger any of the heuristics, three aforementioned baseline methods were used to predict the label. All the predictions were grouped by their baseline function and submitted to the leaderboard to receive scores for each dataset individually as well as the total score per submission. The results are shown in the Table 8 in the section 4. Below we first describe task-specific heuristics in more detail.

3.1 Linguistic Diagnostic for Russian (LiDiRus)

Inspired by [35], the authors of the original SuperGLUE benchmark included a small curated test dataset called AX-b for the analysis of the models' overall performance. It was 'provided not as a benchmark, but as a tool for error analysis, qualitative model comparison, and development of adversarial examples' [11]. LiDiRus is a Russian version of this dataset, where each sentence was translated from English into Russian with the help of 'professional translators and linguists to ensure that the desired linguistic phenomena remain' [33].

We identified a set of heuristics for this dataset. They are grouped in Table 1, which also demonstrates how many samples in the validation set were covered by each heuristic and the percentage of their correct predictions. Only basic split on white-space is applied for pre-processing sentences for all heuristics but one. 'All lemmas in sentences 1 and 2 overlap' required lemmatisation first.

As the dataset does not assume any training and validation samples, the corresponding parts of the Textual Entailment Recognition for Russian (TERRa) dataset from the same RSG benchmark were used to make predictions to calculate the class distribution for the majority class and random weighted baseline functions if the utterances did not trigger the use of any heuristics. TERRa's class distribution differs from LiDiRus[4] but maintains the same dataset organisation.

The performance of the aforementioned heuristics (as well as heuristics for other RSG tasks) is consolidated into Table 8 which can be found in section 4. It provides SOTA scores for each task as of May 2021, performance scores of the baseline functions, as well as the results for heuristics-based approach supported by one of the three baseline functions. The evaluation metric used for LiDiRus is Matthews correlation coefficient [22]. The authors of the original benchmark for English suggested this metric, as

---

[3] https://github.com/eli5-org/eli5
[4] The labels are distributed in the following proportions: 58.4% not_entailment, 41.6% entailment for LiDiRus vs. 49.15% not_entailment, 50.85% entailment for TERRa

	Heuristic	Target label	Coverage	Correct
1	Number of tokens in sentence 1 differs from sentence 2 by more than 10	not_entailment	24.3%	65.2%
2	Sentences 1 and 2 differ by two commas	not_entailment	27.3%	64.1%
3	Sentences 1 and 2 differ by two words	not_entailment	16%	66.6%
4	The presence of 'и', 'не', 'что', 'никогда', 'вовсе', 'это' ('and', 'not', 'that', 'never', 'at all', 'this') in only one of the two sentences	not_entailment	29.3%	66.3%
5	Vocabularies of two sentences overlap by 100% (lemmatised data)	entailment	4%	64.4%
6	'Чтобы', 'будет', 'от', 'он' ('in order to', 'will', 'from', 'he') occur in both sentences	entailment	11.6%	57%

Table 1: LiDiRus: identified heuristics with their coverage of the validation set and the percentage of correct predictions

it can be applied to unbalanced binary classification problems and its values range from -1 to 1, with 0 being the performance of uninformed guessing [11].

As it is a diagnostic dataset, the SOTA approach is hardly applicable to it, though the fact that there is a small performance gap between heuristics and other models deserves to be mentioned. It supports the hypothesis that shallow heuristics might play a significant part in the results of the approaches which apply pre-trained language models to solve NLP tasks.

3.2 Russian Commitment Bank (RCB)

Russian Commitment Bank is a Natural Language Inference task dataset that consists of naturally occurring discourses where the task is to predict the relation of one phrase (hypothesis) to the given text (premise), where the options are entailment, contradiction and neutral.

The training data is distributed unequally in this dataset (46.3% — neutral, 35.4% — entailment, 18.3% — contradiction for train data; 52.7% — neutral, 33.6% — entailment, 13.6% — contradiction for validation data). This imbalance can potentially lead to a substantial bias towards a certain class for the large pre-trained language models. The model can simply predict the majority class and still achieve a rather good result, though it by any means would not be natural language understanding.

The number of instances in the training data is 438, in the validation — 220 and in the test — 438. Two metrics are used to evaluate the model's performance on solving this task: Accuracy and F1, as is the case with the corresponding Commitment Bank task in the original SuperGLUE [11]. According to the authors of the SuperGLUE, the imbalanced nature of the dataset (relatively fewer neutral examples in the English version and significantly more neutral instances in the Russian SuperGLUE) was the reason for them using two metrics, where they used macro-F1 for multi-class problems.

One of the heuristics (you can see its performance in Table 2) used for solving this dataset utilised the correlation between the label and the number of words in the hypothesis or the premise.

Instances with 5-7 words in the hypothesis would more likely have the 'neutral' label (median for it is 5) and instances with less than 5 words in the hypothesis would more likely have the 'contradiction' label (median for it is 4).

Instances with more than 30 words in them would likely belong to the 'entailment' category (median for the 'entailment' is 27).

As we can see from the Table 2, the heuristics do not cover all the data, leaving some answers to be predicted with the help of three baselines (majority class, random choice, random balanced choice). The results achieved with the help of the heuristics were comparable with results of large pre-trained language models in the RSG leaderboard, which are given in the section 4 of this paper.

	Heuristic	Target label	Coverage	Correct
1	The hypothesis is a sub-string of the premise	entailment	26%	40%
2	75% intersection of the hypothesis and premise's vocabularies (lemmatised data)	entailment	5%	45%
3	The presence of 'признать' ('admit') (lemmatised data)	entailment	6%	36%
4	The presence of 'подозревать, считать, говорить, думать, надеяться, понять, уверять' ('suspect, consider, say, think, hope, assure, realise') (lemmatised data)	neutral	6%	36%
5	Hypothesis > 5 words	contradiction	41%	23%
6	4 < hypothesis < 8 words	neutral	34%	70%
7	More than 30 words in the premise	entailment	35%	39%

Table 2: RCB: identified heuristics with their coverage of the validation set and the percentage of correct predictions

3.3 Choice of Plausible Alternatives for Russian language (PARus)

To evaluate progress in open-domain common sense casual reasoning, the authors of Russian Super-GLUE provided the Choice of Plausible Alternatives for Russian language (PARus) dataset. It is based on the English COPA [30]. A typical task in PARus consists of a premise and two alternatives, where the goal is to select the alternative that has a causal or effect relation with the premise.

There are 400 samples in the train dataset and 100 in the validation set. Since there is no semantics behind the labels, the difference between label distribution in the training and validation data should be considered irrelevant. Also because of this lack of label meaning, it was challenging to find linguistic heuristics to solve this task. All textual data was lemmatised to get better results. The heuristics used for tackling this task are shown in Table 3.

The heuristics check whether one of the choices has more shared lemmas with the premise than the others, and if so, then this choice should be taken as an answer. If the vocabulary overlap was the same for all choices, one of the baseline functions was applied. If one of choices had more words than the other, then this choice was taken as an answer.

	Heuristic	Coverage	Correct
1	If one of choices has more shared lemmas with the premise than the others, it is taken as an answer (lemmatised data)	22%	64%
2	If one of choices has more words than the others, then this choice should be taken as an answer (lemmatised data)	59%	52%
3	The combination of these two heuristics (lemmatised data)	66%	52%

Table 3: PARus: identified heuristics with their coverage of the validation set and the percentage of correct predictions

As we can see from Table 3, these heuristics cover less than 70% of the data, therefore many answers still depend on one of three baselines. Overall results are presented in the Table 8 in section 4. The maximal accuracy score was 0.516. To achieve SOTA performance, we probably need more complex algorithms. It proves that this task fulfills its goal and we do need to learn some open-domain common sense casual reasoning to solve such tasks.

3.4 Russian Multi-Sentence Reading Comprehension (MuSeRC)

The MuSeRC dataset is collected for the reading comprehension task. It contains more than 900 paragraphs across 5 different domains: elementary school texts, news, fiction stories, fairy tales, and summaries of TV series and books [10]. Samples were collected based on the following criteria:

1. the passage length is less than 1.5K characters;
2. the passage contains named entities;
3. if the passage contains only one named entity, then it must have one or more co-reference relations.

Furthermore, the authors of the dataset ensured correct sentence splitting and used these sentences in a crowd-sourcing effort at the Yandex.Toloka platform. In it, humans were asked to generate questions, a set of answers for each of them and to check that answering a question requires consulting with more than one sentence in the text. The answer can be either True or False, so all the answers are either correct or incorrect with no in-between. The number of correct answers varies and each question/answer pair is treated individually[5].

A set of heuristics identified for this dataset is grouped in Table 4.

	Heuristic	Target label	Coverage	Correct
1	All lemmas from the answer occur in the text (lemmatised data)	True	39.2%	58.8%
2	The answer is longer than 11 tokens	True	10.3%	72.3%
3	More than 6 overlapping lemmas between the answer and the text (lemmatised data)	True	18.9%	73.9%
4	No overlapping lemmas between the answer and the text (lemmatised data)	False	9.9%	89.1%
5	The answer is shorter than 4 tokens	False	46.4%	64.9%
6	One overlapping lemma between the answer and the text (lemmatised data)	False	18.6%	69.3%

Table 4: MuSeRC: identified heuristics with their coverage of the validation set and the percentage of correct predictions

While predicting, the if-else statements dealt with the 'True' label first, as it is less frequent in the data. The function yields the intended label as long as at least one of the heuristics gets triggered. After that, the opposite set of heuristics is applied.

The overall performance is given in Table 8 which can be found in section 4. To provide the evaluation metrics, the dataset authors roughly followed the evaluation procedure by [28, 21]. Since each answer-option can be assessed independently, F1-averaged (F1a) is applied to evaluate binary decisions over all the answer options in the dataset. It is a harmonic mean of precision and recall per question. Exact Match (EM) is the exact match per each instance, i.e. each set of predictions should be the same as of the answers [10].

We were not able to reach neither the SOTA score nor the human performance, although the obtained results are on par with some of those produced by large pre-trained language models. In fact, at the time of submission, our heuristics-based approach combined with the majority class baseline function achieved higher performance scores for this task than Multilingual Bert and RuGPT3Small[6].

3.5 Textual Entailment Recognition for Russian (TERRa)

Textual Entailment Recognition is another dataset dedicated to the Natural Language Inference task. This task requires to recognise, given two text fragments, whether the meaning of one text is entailed (can be inferred) from the other text [33]. This task is similar to the RCB, yet in TERRa there are only

---

[5]The average number of questions is approximately 20. The labels are distributed in the following proportions: 55% false and 45% true for the training set vs. 55.6% false and 44.4% true for the validation set.

[6]https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2

two categories (entailment/not_entailment) instead of three. The number of instances in the training data is 2 616, in the validation — 307 and in the test — 3 198.

Similar to the RCB, one of the heuristics (Table 5, heuristics 6 and 7) used in solving this dataset utilised the interplay between the label and the number of words. Unlike with the RCB, in this dataset it was possible to find such relations only between the label and the number of words in the premise. Instances with less than 29 words would more likely have the label 'not_entailment' (median number of words for not_entailment is 29) whereas if the number of words was more than 32, then the label is likely 'entailment' (median number of words for entailment is 32).

Another heuristic (8 in Table 5) for TERRa dealt with the presence of specific words, namely 'только' ('only') and 'мужчина' ('man') in the hypothesis. It was possible to find a rather noticeable correlation between their presence and the label.

	Heuristic	Target label	Coverage	Correct
1	The hypothesis is a sub-string of the premise	entailment	1%	50%
2	Vocabularies of the hypothesis and the premise overlap by 33% (lemmatised data)	not_entailment	11%	69%
3	Vocabularies of the hypothesis and the premise overlap by 75% (lemmatised data)	entailment	9%	52%
4	Vocabularies of the hypothesis and the premise overlap by 66% (lemmatised data)	entailment	9%	56%
5	Vocabularies of the hypothesis and the premise overlap by 100% (lemmatised data)	entailment	14%	65%
6	Less that 29 words in the premise	not_entailment	45%	58%
7	More than 32 words in the premise	entailment	45%	60%
8	The presence of 'только', 'мужчина' ('only', 'man') (lemmatised data)	not_entailment	21%	66%

Table 5: TERRa: identified heuristics with their coverage of the validation set and the percentage of correct predictions

As we can see from the Table 5, the heuristics do not cover all the data, leaving some answers to be predicted with the help of the three trivial baselines. However, the results achieved with the help of the heuristics were comparable with the results of large pre-trained language models in the RSG leaderboard and even outperformed the ones by RuGPT3Medium and RuGPT3Small.

3.6 Russian Words in Context (RUSSE)

Depending on its context, a word can have multiple, potentially unrelated, senses. For example, the Russian word 'лук' ('onion'/'bow') can mean either vegetable or weapon depending on its surrounding words. The 'word in context' task can be described as a binary classification problem, and the goal is to predict whether a given word has the same meaning in both given sentences. The Russian SuperGLUE task borrows original data from the Russe Word Sense Induction and Disambiguation shared task [27].

To find out whether the decision can be made based on simple rules, we checked whether the target word appears in the same form in both sentences. In addition, we calculated the proportion of shared tokens to all tokens in both sentences and the difference in their lengths.

The heuristics cover about 50% of the data and make correct predictions in about 65% cases. According to Table 8 in Section 4, we managed to achieve 0.595 accuracy score.

3.7 The Winograd Schema Challenge for Russian (RWSD)

The original purpose of the Winograd Schema Challenge (WSC) was to serve as an alternative Turing test to evaluate an automatic system's capacity for common sense inference [19]. The challenge evaluates the models' ability to identify the antecedent of the pronoun, which might be critical, for example, for translation purposes [5]. The performance scores on the WSC for English quickly progressed from a

	Heuristic	Target label	Coverage	Correct
1	Target word in the same form	True	14%	58%
2	Tokens overlap by more than 10%	True	4%	50%
3	Number of tokens in sentence 1 differs from sentence 2 by more than 6	False	49 %	65 %

Table 6: RUSSE: identified heuristics with their coverage of the validation set and the percentage of correct predictions

simple guess to near-human level [1] after neural language models trained on massive corpora were applied to solve this challenge.

The RWSD dataset is a Russian translation of the pronoun disambiguation task used in the SuperGLUE benchmark [24]. RWSD maintains the same structure providing a pair or a batch of sentences that differ by one or two words:

Example 1: 'Кубок не помещается в коричневый чемодан, потому что он слишком большой.' ('The trophy doesn't fit into the brown suitcase because it is too large.')

Example 2: 'Кубок не помещается в коричневый чемодан, потому что он слишком маленький.' ('The trophy doesn't fit into the brown suitcase because it is too small.')

There is an ambiguity in these sentences, namely 'он' ('it') might refer to either 'кубок' ('the trophy') or 'чемодан' ('suitcase'). Each sentence is followed by an antecedent and a pronoun for disambiguation, which can be successfully resolved if a model assigns a 'false' label to the first example for the pair of 'чемодан'('suitcase') and 'он' ('it'), and if 'true' is assigned to the second example for the same pair.

The model cannot rely on the word order or the structure of a sentence, as the task is organised so that they cannot be used for the disambiguation process [24]. Each sentence might be either 'true' or 'false' depending on a suggested pair of antecedents and pronouns. For example, one has to pay attention to a special word, i.e. 'большой' ('large') or 'маленький' ('small') in the aforementioned sentences.

There is a clearly unequal distribution of classes in the RWSD dataset. The labels for the training and validation sets are distributed as follows: 51% 'false' and 49% 'true' labels for the former and 55.4% 'false' and 44.6% 'true' labels for the latter. However, the 'false' values appear 67% of the times in the *test set*, which is very different from the datasets provided for training and validation.

We were not able to identify any heuristic that would surpass the performance score of predictions made by the majority class baseline (see Table 8 for reference), but this misfortune carries one of our most important findings.

Apparently, the very same approach to choose the most common value was used by many sophisticated models listed in the Russian SuperGLUE leaderboard by the time of our submission. The SOTA score, which is the score achieved by Multilingual T5, several BERT variations (trained on multilingual data and on Russian corpora only), RuGPT3Medium and RuGPT3Small, is 0.669: the same as achieved by our majority class baseline function. While solving the task, these models allegedly opted to predict using the majority class rather than try to actually solve the Winograd Schema Challenge. Such models as Golden Transformer, RuGPT3XL few-shot and RuGPT3Large apparently made an attempt to really predict something, but their results are sub-optimal:0.545, 0.649 and 0.636 respectively, which is in fact below the 0.662 tf-idf baseline provided by the RSG creators. The problem is similar to that with Winograd Schema Challenge (WSC) in the original SuperGLUE [11].

3.8 Yes/no Question Answering Dataset for Russian (DaNetQA)

DaNetQA is a question answering dataset for yes/no questions. Each example is a triplet of (passage, question, answer), with the title of the page as optional context [33]. The answers are encoded in a True/False formal similar to the corresponding SuperGLUE 'BoolQ' dataset. As with the Russian Commitment Bank task, here we can also notice the unequal distribution of labels (Train: True — 60.7%,

	Heuristic	Target label	Coverage	Correct
1	The question starts with 'был' ('was/were')	True	45%	58%
2	The question starts with 'есть' ('is/are')	True	13%	81%
3	The question starts with 'входит ли' ('does it belong to')	False	37%	100%
4	The question starts with 'едят ли' ('do they eat')	False	2%	53%
5	The question starts with 'правда ли' ('is it true')	False	18%	89%
6	More than 5 words in the question	False	46%	71%
7	More than 90 words in the passage	False	48%	53%

Table 7: DaNetQA: identified heuristics with their coverage of the validation set and the percentage of correct predictions

False — 39.3%) and the mismatch of this relation among training and validation data (Vaidation: True — 50.2%, False — 49.8%). The number of instances in the training data is 1 749, in the validation — 821 and 805 in the test set.

Like with the RCB and TERRa, one of the heuristics used in solving this dataset utilised the relations between the label and the number of words in questions (Table 7, heuristic 6) or passage (heuristic 7).

Instances with more than 5 words would more likely have the label 'False' (median number of words for False is 6 in the training data).

Heuristic 3 exploits correlation between the beginning of the question and the label: if the question starts with 'входит ли' ('does it belong to'), the label in the validation dataset is False 100% of the time.

As we can see from the Table 7, the heuristics do not cover all the data, leaving some answers to be predicted with the help of the three trivial baselines. However, the results achieved with the help of the heuristics were comparable with the results of large pre-trained language models in the RSG leaderboard.

3.9 Russian Reading Comprehension with Commonsense Reasoning (RuCOS)

Russian reading comprehension with Commonsense reasoning (RuCoS) is a large-scale reading comprehension dataset which requires common sense reasoning. Unlike MuSeRC, the main data domain for RuCoS is news articles and there is more data for this task. Also in this task, a system is given a list of named entities from which it should choose the right answer (while in the MuSeRC, the answers do not have to be named entities at all). RuCoS consists of queries automatically generated from news articles; the answer to each query is a text span from a summarizing passage of the corresponding article.

This task is based on the English ReCoRD benchmark [28]. All text examples were collected from Russian media. The texts were then filtered by the IPM frequencies of the contained words and, finally, manually reviewed.

The heuristics applied to this task dealt with the presence of name entities in the question. The algorithm simply predicted all the entities present in the question. A modification of this approach was to sort the remaining entities based on the frequency of their appearance in the text. We tried several threshold values for this rule and finalized our choice on the following rule: all entities whose stems appeared less than three times were removed from our predictions.

Both heuristics were applied every time we made predictions. Thus, their coverage is 100%. We managed to outperform the tf-idf baseline with both F1 score end EM metric around 0.26, but the SOTA results are on par with human performance score, which is 0.93/ for F1 and 0.89 for EM.

	Metrics	Hum.	SOTA	maj.	rand.	r.(b)	H maj.	H rand.	H r.(b)
LiDiRus	M. Corr	0.626	**0.231**	0.000	0.024	0.000	0.147	0.149	**0.182**
RCB	Avg. F1	0.680	**0.452**	0.217	0.332	0.319	0.400	**0.401**	**0.401**
	Acc.	0.702	**0.546**	**0.484**	0.347	0.374	0.438	0.436	0.438
PARus	Acc.	0.982	**0.908**	0.498	0.474	0.480	0.478	**0.508**	0.470
MuSeRC	F1a	0.806	**0.941**	0.000	0.477	0.450	**0.671**	0.669	0.669
	EM	0.420	**0.819**	0.000	0.078	0.071	0.237	0.195	**0.202**
TERRa	Acc.	0.920	**0.871**	0.513	0.503	0.483	**0.549**	0.547	0.548
RUSSE	Acc.	0.805	**0.729**	0.587	0.501	0.528	**0.595**	0.497	0.543
RWSD	Acc.	0.840	**0.669**	**0.669**	0.487	0.597	**0.669**	0.565	0.604
DaNetQA	Acc.	0.915	**0.917**	0.503	0.494	0.520	**0.642**	0.629	0.629
RuCoS	F1	0.930	**0.920**	0.250	0.250	0.250	**0.260**	0.260	0.260
	EM	0.890	**0.924**	0.247	0.247	0.247	**0.257**	0.257	0.257
Total		0.811	**0.679**	0.374	0.372	0.385	**0.468**	0.445	0.454

Table 8: Performance scores at the time of submission. 'Maj.' is the majority class baseline function, 'rand.' — random choice, 'r.(b)' — random balanced choice, H — the heuristics-based approach.

4 Discussion

Table 8 shows the best results after applying the heuristics described above to the Russian SuperGLUE test sets. The heuristic based approach (H) was combined with one of the trivial baseline functions. The majority value and weights for baseline functions were obtained from combined training and validation sets. For every task, we chose heuristic (or combination of several heuristics) that led towards the best score. Even if for some tasks we are still far away from beating SOTA performance, simple baselines and heuristics can achieve relatively good results. For RWSD task one can achieve SOTA performance just by using the majority class baseline.

Our heuristics approach works well for the RCB task with the difference between H maj. model and SOTA being about 5%. For TERRa, RUSSE and DaNetQA we are far from SOTA results but, still, our results are on the same level with RuBERT [16] and GPT models from the leaderboard.

Yet for several tasks, the heuristics approach did not work as well. RuCoS, MuSeRC and PARus proved that it is not enough to use dataset-specific statistical cues to solve them, so for these three tasks it seems that the large pre-trained language models really pick up some peculiarities of Russian language.

Since our approaches can be divided into two groups (trivial baselines and rule-based heuristics), we will look closer at them separately.

4.1 Trivial baselines

As it was mentioned before, first we chose three baseline methods to solve all tasks in Russian Super-GLUE: majority class, random choice and random weighted choice. When comparing these baselines to the other methods, we should keep in mind that they do not rely on any linguistic knowledge at all.

All three baselines show very interesting results. For the majority class baseline, the best result is the one for the RWSD task. It should be emphasized again that not only one can achieve the SOTA performance with the majority baseline here, but, at the moment of submission, half of the leaderboard models probably use this approach as their solver method, since they all have the same performance score. Simple random choice worked good on the RCB and RWSD as well. Random balanced choice outperforms the majority class approach on the DaNetQA, RWSD, TERRa, PARus, and RUSSE.

Across all RSG tasks, the balanced random choice baseline achieves the average score of 0.385. Language models obviously outperform this score, but the difference is marginal: only half of the systems in the leaderboard achieve a score higher than 0.5, and the best ensemble of transformers reaches 0.679

(the human performance is 0.811). For some benchmarks (for example, RuCOS), the random balanced baseline *outperforms* BERT and GPT-3 models. In one specific case of the RWSD benchmark, no model managed to outperform the *simple majority class baseline*.

From this, we conclude that the RSG leaderboard scores should be taken with a grain of salt and compared to the trivial baselines. For example, the 0.669 accuracy of the SOTA models on the RWSD dataset is not a sign of their 'human-like comprehension abilities': it is just that these models (or their authors) could not do any better than simply predict the same label for all the instances in the test set. For other tasks, the picture is only slightly better: in most cases, the leaderboard participants managed to improve the random balanced baseline only by a small margin. Another important finding is that the class balances in the RSG test sets are similar to those in the validation and training sets: this allows one to achieve boosted accuracies by simply replicating these distributions in the test answers. This is true for all the tasks evaluated by accuracy (six of the RSG tasks). If the class labels in these six datasets were perfectly balanced, the expected average accuracy of the random baseline would be 0.472. In the real RSG, this value is 0.538. This is certainly an undesired property for a test set in general; in this case it additionally makes it difficult to assess to what extent the large-scale language models' NLU performance for Russian is actually an artifact of this data leakage.

4.2 Rule-based heuristics

Rule-based heuristics tend to improve trivial baselines in cases of TERRa, RUSSE, RCB (considering Avg. F1 score). Here we categorize the described rules. Note that most of them are language-agnostic and can be tested on benchmarks for other languages as well.

1. Using text length (e. g. 'More than 30 words in the premise'): these rules are useful for RCB, PARus, MuSeRC, TERRa, RUSSE, DaNetQA.
2. Using binary lexical features (e. g. 'Presence of 'чтобы', 'будет', 'от', 'он'): these rules are useful for LiDiRus, RCB, TERRa, DaNetQA.
3. Using word forms or lemmas overlap (e. g. 'Sentences 1 and 2 use the same set of lemmas'): these rules are useful for LiDiRus, RCB, PARus, MuSeRC, TERRa, RUSSE.
4. Other task-specific heuristics.

The existence of such statistical cues is not a problem in itself: after all, this is what machine learning is after. What we see as problematic is the fact that the large over-parameterized models seem to mostly rely on them (judging by their performance scores which are not radically higher, and sometimes even lower than the scores of our rule-based approach). This means they do not employ valid inference strategies, and do not demonstrate anything close to much-praised 'natural language comprehension'. We again emphasize that our heuristics are extremely simplistic and often boil down to counting the number of words in the sentence or to finding the lexical overlap between the question and the answer (sometimes after lemmatisation). There is no doubt that billion-parameter language models can find much more statistical cues in the training data than the authors of this paper were able to come up with. But these regularities will only work on the test instances drawn from the same general population (annotated or generated according to the same guidelines). This is *pattern matching*, not *language understanding*.

5 Conclusion

The recently introduced Russian SuperGLUE (RSG) set of natural language understanding benchmarks [33] has already attracted well-deserved attention from the Russian NLP practitioners. The RSG leaderboard is filled with the impressive performance scores produced by sophisticated language models trained with bleeding-edge deep learning architectures (BERT, GPT-3, etc) on titanic corpora of Russian. But are these scores really so impressive? In this paper, we studied what performance can be achieved for the RSG benchmarks *without training any language models*.

First, we established the performance boundaries of the trivial baselines: random choice, majority class choice and balanced random class choice (probabilities weighted by the distribution of class labels in the training data). We found that in some cases, these baselines outperform large-scale language models. Second, we moved on to find out whether the RSG datasets contain other statistical regularities.

In has been shown in prior work for English and other languages that deep learning models are very prone to collecting low-hanging fruits and tracing shallow semantic and structural phenomena which help minimizing the loss on a particular dataset, instead of actually learning real linguistic generalizations.

To this end, we manually compiled a set of very simple custom rule-based heuristics for each RSG dataset (for example, '**set the label '*contradiction*' if the word** не **'not' is present in the hypothesis**', etc).

It turned out that up to 50% and more of instances (depending on a particular dataset) might be covered by this or that heuristic. Moreover, applying these rules to actually solve the RSG (with fallback to the majority class baseline if no rule is applicable) constitutes a system which achieves a very competitive RSG average score of 0.468. This outperforms RuGPT3-Small, on par with RuGPT3-Medium, and is very close to the BERT performance.

We conclude that most RSG datasets abound in statistical regularities which can easily be found at training time and employed at test time, without expensive and complicated language model pre-training. The reasons are arguably the same as with the English test sets[7]: compilation of benchmarks by crowd-sourcing and the natural desire of crowd-workers to fulfill the job in the easiest way possible.

To sum up, we recommend the RSG maintainers to 1) modify the test sets to minimize the data leakage from label distributions; 2) diversify the datasets so as to eliminate at least the most striking statistical cues (it shouldn't be possible to find the correct answer by simply counting words); 3) provide official majority class and random weighted baselines. We believe this will make the Russian SuperGLUE leaderboard even more informative of the real state of the art in Russian natural language processing.

In the future, it will be useful to develop a Russian equivalent of the HANS benchmark [23]: a test set containing adversarial examples, or even simply examples drawn from sources substantially different from those in the RSG. It will allow to evaluate the generalization abilities of large pre-trained language models for Russian. It would also be interesting to study the correlations between the predictions of our heuristics and the predictions of the language models in the RSG leader-board, in order to find out whether they actually exploit similar rules.

Finally, in the course of working on this paper, we collected a large trove of annotation errors and generally problematic or controversial cases in the RSG datasets. We have shared these findings with the RSG maintainers, in the hope of its future improvement.

References

[1] An Analysis of Dataset Overlap on Winograd-Style Tasks / Ali Emami, Kaheer Suleman, Adam Trischler, Jackie Chi Kit Cheung // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 5855–5865. — Access mode: `https://www.aclweb.org/anthology/2020.coling-main.515`.

[2] Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets / Saku Sugawara, Pontus Stenetorp, Kentaro Inui, Akiko Aizawa // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Apr. — Vol. 34, no. 05. — P. 8918–8927. — Access mode: `https://ojs.aaai.org/index.php/AAAI/article/view/6422`.

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // 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). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: `https://www.aclweb.org/anthology/N19-1423`.

[4] CLUE: A Chinese Language Understanding Evaluation Benchmark / Liang Xu, Hai Hu, Xuanwei Zhang et al. // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona,

---

[7]In fact, many RSG test sets are translated from English.

Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 4762–4772. — Access mode: `https://www.aclweb.org/anthology/2020.coling-main.419`.

[5] Davis Ernest. Winograd Schemas and Machine Translation // CoRR. — 2016. — Vol. abs/1608.01884. — 1608.01884.

[6] He Pengcheng, Liu Xiaodong, Gao Jianfeng, Chen Weizhu. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. — 2021. — 2006.03654.

[7] ERNIE: Enhanced Language Representation with Informative Entities / Zhengyan Zhang, Xu Han, Zhiyuan Liu et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 1441–1451. — Access mode: `https://www.aclweb.org/anthology/P19-1139`.

[8] Ethayarajh Kawin, Jurafsky Dan. Utility is in the Eye of the User: A Critique of NLP Leaderboards // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online : Association for Computational Linguistics, 2020. — Nov. — P. 4846–4853. — Access mode: `https://www.aclweb.org/anthology/2020.emnlp-main.393`.

[9] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer / Colin Raffel, Noam Shazeer, Adam Roberts et al. // Journal of Machine Learning Research. — 2020. — Vol. 21, no. 140. — P. 1–67. — Access mode: `http://jmlr.org/papers/v21/20-074.html`.

[10] Fenogenova Alena, Mikhailov Vladislav, Shevelev Denis. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 6481–6497. — Access mode: `https://www.aclweb.org/anthology/2020.coling-main.570`.

[11] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding / Alex Wang, Amanpreet Singh, Julian Michael et al. // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. — Brussels, Belgium : Association for Computational Linguistics, 2018. — Nov. — P. 353–355. — Access mode: `https://www.aclweb.org/anthology/W18-5446`.

[12] Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks / Anna Rogers, O. Kovaleva, Matthew Downey, Anna Rumshisky // AAAI. — 2020.

[13] HellaSwag: Can a Machine Really Finish Your Sentence? / Rowan Zellers, Ari Holtzman, Yonatan Bisk et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4791–4800. — Access mode: `https://www.aclweb.org/anthology/P19-1472`.

[14] Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment / Di Jin, Zhijing Jin, Joey Tianyi Zhou, Peter Szolovits // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Apr. — Vol. 34, no. 05. — P. 8018–8025. — Access mode: `https://ojs.aaai.org/index.php/AAAI/article/view/6311`.

[15] Korobov Mikhail. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts / Ed. by Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko et al. — Springer International Publishing, 2015. — Vol. 542 of Communications in Computer and Information Science. — P. 320–332. — online; accessed: `http://dx.doi.org/10.1007/978-3-319-26123-2_31`.

[16] Kuratov Yuri, Arkhipov Mikhail. Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". — 2019. — Access mode: `http://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf`.

[17] Language Models are Few-Shot Learners / Tom Brown, Benjamin Mann, Nick Ryder et al. // Advances in Neural Information Processing Systems / Ed. by H. Larochelle,

M. Ranzato, R. Hadsell et al. — Vol. 33. — Curran Associates, Inc., 2020. — P. 1877–1901. — Access mode: `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[18] Learning and Evaluating General Linguistic Intelligence / Dani Yogatama, Cyprien de Masson d'Autume, J. Connor et al. // ArXiv. — 2019. — Vol. abs/1901.11373.

[19] Levesque Hector J., Davis Ernest, Morgenstern Leora. The Winograd Schema Challenge // Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. — KR'12. — Rome, Italy : AAAI Press, 2012. — P. 552–561.

[20] Linzen Tal. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 5210–5217. — Access mode: `https://www.aclweb.org/anthology/2020.acl-main.465`.

[21] Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences / Daniel Khashabi, S. Chaturvedi, Michael Roth et al. // NAACL-HLT. — 2018.

[22] Matthews B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme // Biochimica et Biophysica Acta (BBA) - Protein Structure. — 1975. — Vol. 405, no. 2. — P. 442–451. — Access mode: `https://www.sciencedirect.com/science/article/pii/0005279575901099`.

[23] McCoy Tom, Pavlick Ellie, Linzen Tal. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 3428–3448. — Access mode: `https://www.aclweb.org/anthology/P19-1334`.

[24] Morgenstern Leora, Davis Ernest, Ortiz Charles L. Planning, Executing, and Evaluating the Winograd Schema Challenge // AI Magazine. — 2016. — Apr. — Vol. 37, no. 1. — P. 50–54. — Access mode: `https://ojs.aaai.org/index.php/aimagazine/article/view/2639`.

[25] Nangia Nikita, Bowman Samuel R. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4566–4575. — Access mode: `https://www.aclweb.org/anthology/P19-1449`.

[26] Niven Timothy, Kao Hung-Yu. Probing Neural Network Comprehension of Natural Language Arguments // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 4658–4664. — Access mode: `https://www.aclweb.org/anthology/P19-1459`.

[27] RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language / Alexander Panchenko, Anastasia Lopukhina, Dmitry Ustalov et al. // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". — Moscow, Russia : RSUH, 2018. — P. 547–564. — Access mode: `http://www.dialog-21.ru/media/4539/panchenkoaplusetal.pdf`.

[28] ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension / Sheng Zhang, Xiaodong Liu, Jingjing Liu et al. // CoRR. — 2018. — Vol. abs/1810.12885. — 1810.12885.

[29] RoBERTa: A Robustly Optimized BERT Pretraining Approach / Yinhan Liu, Myle Ott, Naman Goyal et al. // arXiv preprint arXiv:1907.11692. — 2019.

[30] Roemmele Melissa, Bejan Cosmin Adrian, Gordon Andrew S. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. // AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning. — 2011. — P. 90–95.

[31] Rogers Anna. How the Transformers broke NLP leaderboards. — 2019. — Jun. — Access mode: `https://hackingsemantics.xyz/2019/leaderboards/`.

[32] Rogers Anna, Kovaleva Olga, Rumshisky Anna. A Primer in BERTology: What We Know About How BERT Works // Transactions of the Association for Computational Linguistics. — 2020. — Vol. 8. — P. 842–866. — Access mode: `https://www.aclweb.org/anthology/2020.tacl-1.54`.

[33] RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online : Association for Computational Linguistics, 2020. — Nov. — P. 4717–4726. — Access mode: `https://www.aclweb.org/anthology/2020.emnlp-main.381`.

[34] SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems / Alex Wang, Yada Pruksachatkun, Nikita Nangia et al. // Advances in Neural Information Processing Systems / Ed. by H. Wallach, H. Larochelle, A. Beygelzimer et al. — Vol. 32. — Curran Associates, Inc., 2019. — Access mode: `https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf`.

[35] Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task / Allyson Ettinger, Sudha Rao, Hal Daumé III, Emily M. Bender // Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems. — Copenhagen, Denmark : Association for Computational Linguistics, 2017. — Sep. — P. 1–10. — Access mode: `https://www.aclweb.org/anthology/W17-5401`.

[36] What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? / Chenglei Si, Shuohang Wang, Min-Yen Kan, Jing Jiang // ArXiv. — 2019. — Vol. abs/1910.12391.