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

Tatyana Iazykova[1] Olga Bystrova[1]
Denis Kapelyushnik[1] Andrey Kutuzov[2]

[1]National Research University
Higher School of Economics (Moscow),
[2]University of Oslo (Oslo)

10 March 2022

# First things first


- Russian researchers largely do not support Putin's invasion in Ukraine
- Dozens of students were arrested during the anti-war protests in Russia
- The war must end immediately.

# First things first


- Russian researchers largely do not support Putin's invasion in Ukraine
- Dozens of students were arrested during the anti-war protests in Russia
- The war must end immediately.

But one shouldn't blame Russian language.

# Contents

# Standing on the shoulders of giants

## Leaderboards

- GLUE [Wang et al., 2018]
- SuperGLUE [Wang et al., 2019]
- CLUE [Xu et al., 2020]
- Russian SuperGLUE [Shavrina et al., 2020]

# Standing on the shoulders of giants

## Leaderboards

- GLUE [Wang et al., 2018]
- SuperGLUE [Wang et al., 2019]
- CLUE [Xu et al., 2020]
- Russian SuperGLUE [Shavrina et al., 2020]

## Critique

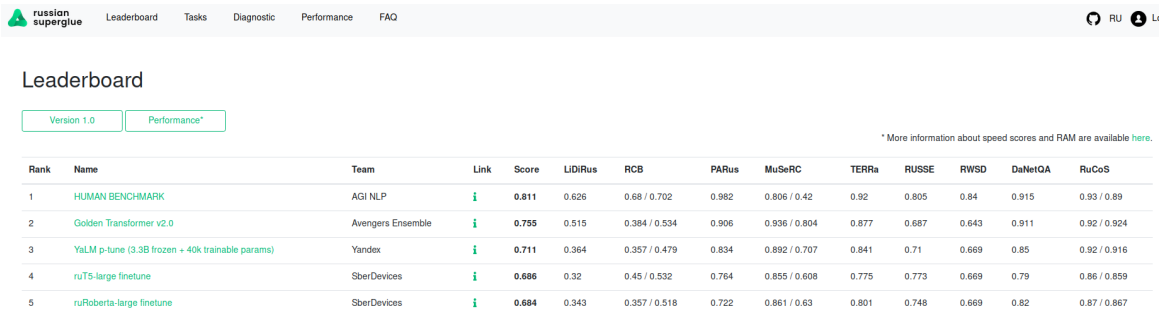
- A model or a resource? Not reproducible: big tech only [Rogers, 2019]
- Statistical cues and annotation artifacts; 'hypothesis-only' models [Poliak et al., 2018]
- 'Right for the wrong reasons' [McCoy et al., 2019]
- Poor proxies for NLP practitioners [Ethayarajh and Jurafsky, 2020]

# Russian SuperGLUE (RSG) overview

## What is RSG?

- 8 datasets + 1 diagnostic set
- Typical tasks: NLI, Common Sense, Machine Reading, World Knowledge, etc
- Data from news articles, existing datasets, English-translated datasets
- Mostly binary classification

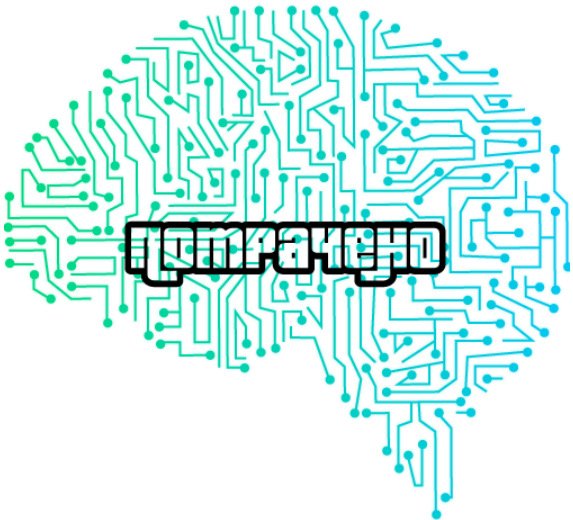
# What is this about




| Rank | Name | Team | Link | Score | LiDiRus | RCB | PARus | MuSeRC | TERRa | RUSSE | RWSD | DaNetQA | RuCoS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | HUMAN BENCHMARK | AGI NLP | i | **0.811** | 0.626 | 0.68 / 0.702 | 0.982 | 0.806 / 0.42 | 0.92 | 0.805 | 0.84 | 0.915 | 0.93 / 0.89 |
| 2 | Golden Transformer v2.0 | Avengers Ensemble | i | **0.755** | 0.515 | 0.384 / 0.534 | 0.906 | 0.936 / 0.804 | 0.877 | 0.687 | 0.643 | 0.911 | 0.92 / 0.924 |
| 3 | YaLM p-tune (3.3B frozen + 40k trainable params) | Yandex | i | **0.711** | 0.364 | 0.357 / 0.479 | 0.834 | 0.892 / 0.707 | 0.841 | 0.71 | 0.669 | 0.85 | 0.92 / 0.916 |
| 4 | ruT5-large finetune | SberDevices | i | **0.686** | 0.32 | 0.45 / 0.532 | 0.764 | 0.855 / 0.608 | 0.775 | 0.773 | 0.669 | 0.79 | 0.86 / 0.859 |
| 5 | ruRoberta-large finetune | SberDevices | i | **0.684** | 0.343 | 0.357 / 0.518 | 0.722 | 0.861 / 0.63 | 0.801 | 0.748 | 0.669 | 0.82 | 0.87 / 0.867 |

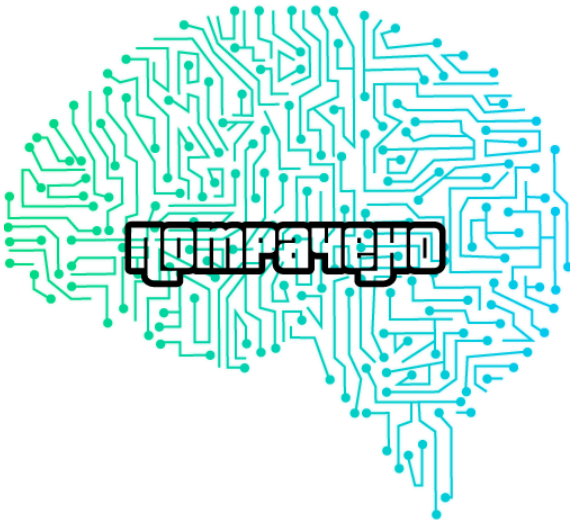
https://russiansuperglue.com/

# Attacking Russian SuperGLUE (RSG)



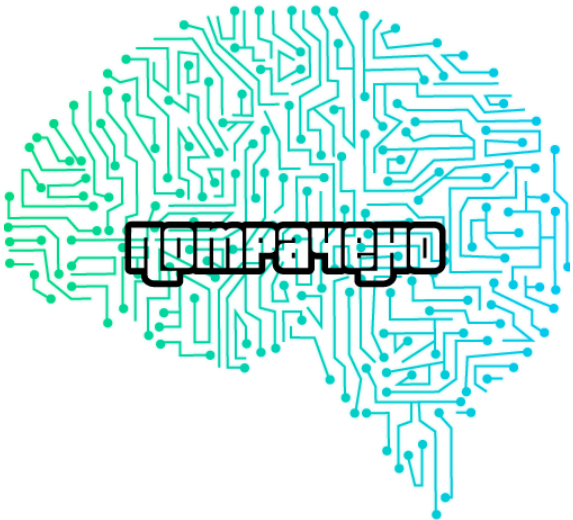
- Our 1st aim: Improve RSG datasets

# Attacking Russian SuperGLUE (RSG)



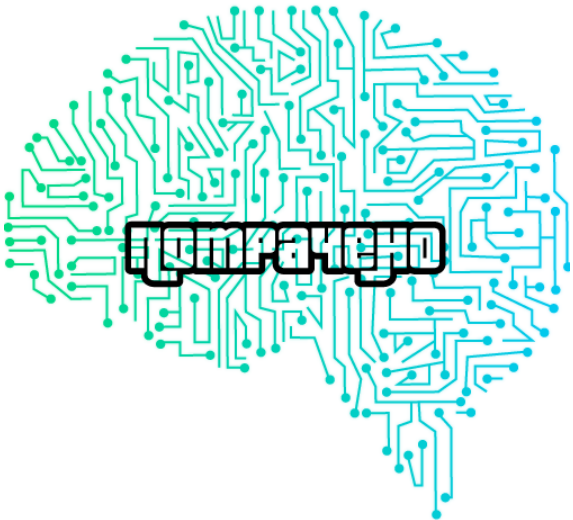
- Our 1st aim: Improve RSG datasets
- Our 2nd aim: are ‘foundational models’ that good in RSG?

# Attacking Russian SuperGLUE (RSG)




- Our 1st aim: Improve RSG datasets
- Our 2nd aim: are ‘foundational models’ that good in RSG?
- Now, what did we do?
- Let’s try to solve RSG with shallow heuristics!

# Attacking Russian SuperGLUE (RSG)



- Our 1st aim: Improve RSG datasets
- Our 2nd aim: are 'foundational models' that good in RSG?
- Now, what did we do?
- Let's try to solve RSG with shallow heuristics!
- No ML, make humans great again, etc
- Semi-manual search for useful rules.

Paper: `http://www.dialog-21.ru/media/5514/iazykovatplusetal037.pdf`

# Contents

# Trivial Baselines

The usual naive approaches:

- Random Choice:
  - for every item, predict the class randomly
- Majority Class:
  - always predict the most common class from the training data
- Random Balanced Choice:
  - for every item, predict the class randomly taking into account the label distribution in the train data

# Heuristics

‘Heuristic — proceeding to a solution by trial and error or by rules that are only loosely defined’ (Oxford English Dictionary)

## Categories of heuristics we are using in the RSG benchmarks:

1. Interplay between the class label and text length (e. g. ‘more than 30 words in the premise’)
2. Presence of specific words (e. g. ‘чтобы’, ‘будет’, ‘от’, ‘он’)
3. Overlapping lemmas or inflected word forms (e. g. ‘sets of lemmas in sentences 1 and sentence 2 overlap 80% ’)
4. Other task-specific heuristics.

# Heuristics

'Heuristic — proceeding to a solution by trial and error or by rules that are only loosely defined' (Oxford English Dictionary)

**Categories of heuristics we are using in the RSG benchmarks:**

1. Interplay between the class label and text length (e. g. 'more than 30 words in the premise')
2. Presence of specific words (e. g. 'чтобы', 'будет', 'от', 'он')
3. Overlapping lemmas or inflected word forms (e. g. 'sets of lemmas in sentences 1 and sentence 2 overlap 80% ')
4. Other task-specific heuristics.

Let's look at some examples of heuristics.

# Contents

# Textual Entailment Recognition for Russian (TERRa)

## TERRa Overview

- Natural Language Inference task
- Given: hypothesis and premise
- Labels: entailment/not_entailment

| Heuristic | Target label | Coverage | Correct |
|---|---|---|---|
| Lemmas in the hypothesis and the premise overlap <33% | not_entailment | 11% | 69% |
| Lemmas of the hypothesis and the premise overlap 100% | entailment | 14% | 65% |
| More than 32 words in the premise | entailment | 45% | 60% |
| The presence of ‘только’, ‘мужчина’ (‘only’, ‘man’) in the premise | not_entailment | 21% | 66% |

# Textual Entailment Recognition for Russian (TERRa)

## TERRa Overview

- Natural Language Inference task
- Given: hypothesis and premise
- Labels: entailment/not_entailment

| Heuristic | Target label | Coverage | Correct |
|---|---|---|---|
| Lemmas in the hypothesis and the premise overlap <33% | not_entailment | 11% | 69% |
| Lemmas of the hypothesis and the premise overlap 100% | entailment | 14% | 65% |
| More than 32 words in the premise | entailment | 45% | 60% |
| The presence of 'только', 'мужчина' ('only', 'man') in the premise | not_entailment | 21% | 66% |

Accuracy on the test set: 0.549 (higher than the `RuGPT3Medium` model by SberDevices).

# Textual Entailment Recognition for Russian

Yes, if the words ‘только’, ‘мужчина’ (‘only’, ‘man’) occur in the premise, it is very probable that the answer is not_entailment!

## Example

- Premise: ‘Была установлена личность подозреваемого - 27-летнего мужчины. По словам задержанного, он был давно влюблен в жену убитого и различными способами добивался ее внимания.’
  *(The suspect was identified as man of 27. According to the apprehended, he had long been in love with the killed man’s wife and tried hard to win her over.)*
- Hypothesis: 27-летний мужчина похищен.
  *(A man of 27 was kidnapped.)*
- Label: not_entailment

# Textual Entailment Recognition for Russian

Yes, if the words ‘только’, ‘мужчина’ (‘only’, ‘man’) occur in the premise, it is very probable that the answer is not_entailment!

## Example

- Premise: ‘Была установлена личность подозреваемого - 27-летнего мужчины. По словам задержанного, он был давно влюблен в жену убитого и различными способами добивался ее внимания.’
  *(The suspect was identified as man of 27. According to the apprehended, he had long been in love with the killed man’s wife and tried hard to win her over.)*
- Hypothesis: 27-летний мужчина похищен.
  *(A man of 27 was kidnapped.)*
- Label: not_entailment

Sad but true.

# The Winograd Schema Challenge Russian (RWSD)

Logic and Reasoning, World knowledge, Binary Classification: true / false

- False: 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.)
- True: 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.)

# The Winograd Schema Challenge Russian (RWSD)

Logic and Reasoning, World knowledge, Binary Classification: true / false

- False: 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.)
- True: 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.)

**Everyone really struggles with RWSD!**

- SOTA Accuracy: 0.669
- BERT, GPT3, RoBerta, MT5...
- majority class predictions accuracy: ...

# The Winograd Schema Challenge Russian (RWSD)

Logic and Reasoning, World knowledge, Binary Classification: true / false

- False: 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.)
- True: 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.)

**Everyone really struggles with RWSD!**

- SOTA Accuracy: 0.669
- BERT, GPT3, RoBerta, MT5...
- majority class predictions accuracy: ...
- 0.669!

# The Winograd Schema Challenge Russian (RWSD)

Logic and Reasoning, World knowledge, Binary Classification: true / false

- False: 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.)
- True: 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.)

## Everyone really struggles with RWSD!

- SOTA Accuracy: 0.669
- BERT, GPT3, RoBerta, MT5...
- majority class predictions accuracy: ...
- 0.669!
- Language models are in fact predicting the majority class.

# The Winograd Schema Challenge Russian (RWSD)

Logic and Reasoning, World knowledge, Binary Classification: true / false

- False: 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.)
- True: 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.)

## Everyone really struggles with RWSD!

- SOTA Accuracy: 0.669
- BERT, GPT3, RoBerta, MT5...
- majority class predictions accuracy: ...
- 0.669!
- Language models are in fact predicting the majority class.
- The same was observed for English [Wang et al., 2019]

# Contents

# What we found

## Overall RSG scores:

| | | | | |
|---|---|---|---|---|
| 12 | RuBERT plain | DeepPavlov | i | 0.521 |
| 13 | SBERT_Large_mt_ru_finetuning | SberDevices | i | 0.514 |
| 14 | SBERT_Large | SberDevices | i | 0.51 |
| 15 | RuGPT3Large | SberDevices | i | 0.505 |
| 16 | RuBERT conversational | DeepPavlov | i | 0.5 |
| 17 | Multilingual Bert | DeepPavlov | i | 0.495 |
| 18 | heuristic majority | hse_ling | i | 0.468 |
| 19 | RuGPT3Medium | SberDevices | i | 0.468 |
| 20 | RuGPT3Small | SberDevices | i | 0.438 |
| 21 | Baseline TF-IDF1.1 | AGI NLP | i | 0.434 |
| 22 | Random weighted | hse_ling | i | 0.385 |
| 23 | majority_class | hse_ling | i | 0.374 |

# Overall results

| | Metrics | Human performance | SOTA | Majority class | Random | Random balanced | Heuristics + majority class |
|---|---|---|---|---|---|---|---|
| **LiDiRus** | M. Corr | 0.626 | 0.515 | 0.000 | 0.024 | 0.000 | 0.147 |
| **RCB** | Avg. F1<br>Acc. | 0.680<br>0.702 | 0.452<br>0.484 | 0.217<br>**0.484** | 0.332<br>0.347 | 0.319<br>0.374 | **0.400**<br>0.438 |
| **PARus** | Acc. | 0.982 | 0.908 | 0.498 | 0.474 | 0.480 | 0.478 |
| **MuSeRC** | F1a<br>EM | 0.806<br>0.420 | 0.941<br>0.819 | 0.000<br>0.000 | 0.477<br>0.078 | 0.450<br>0.071 | **0.671**<br>**0.237** |
| **TERRa** | Acc. | 0.920 | 0.877 | 0.513 | 0.503 | 0.483 | **0.549** |
| **RUSSE** | Acc. | 0.805 | 0.773 | 0.587 | 0.501 | 0.528 | **0.595** |
| **RWSD** | Acc. | 0.840 | 0.675 | **0.669** | 0.487 | 0.597 | **0.669** |
| **DaNetQA** | Acc. | 0.915 | 0.917 | 0.503 | 0.494 | 0.520 | **0.642** |
| **RuCoS** | F1<br>EM | 0.930<br>0.890 | 0.920<br>0.924 | 0.250<br>0.247 | 0.250<br>0.247 | 0.250<br>0.247 | **0.260**<br>**0.257** |
| **Average** | | 0.811 | 0.755 | 0.374 | 0.372 | 0.385 | **0.468** |

# Contents

# Summing up

- We did hack the RSG: heuristics are indeed effective.

# Summing up

- We did hack the RSG: heuristics are indeed effective.
- Unreasonably? Or reasonably?
- 50% and more of instances (depending on a particular dataset) are covered by heuristics.

# Summing up

- We did hack the RSG: heuristics are indeed effective.
- Unreasonably? Or reasonably?
- 50% and more of instances (depending on a particular dataset) are covered by heuristics.
- Competitive performance can be achieved for the RSG benchmarks without training any language models.

# Summing up

- We did hack the RSG: heuristics are indeed effective.
- Unreasonably? Or reasonably?
- 50% and more of instances (depending on a particular dataset) are covered by heuristics.
- Competitive performance can be achieved for the RSG benchmarks without training any language models.
- The performance of large Russian LMs is not much higher than naive heuristics
- 'Human-like comprehension abilities'? No.
- Pattern matching, not language understanding.

# Summing up

## What to do?

Protect our benchmarks from hacking:

- adversarial examples (HANS benchmark [McCoy et al., 2019])
- test sets drawn from different sources
  - ...so that surface cues could not be exploited
- uniform distribution of labels
- more transparency: publish models' predictions.

# Summing up

## What to do?

Protect our benchmarks from hacking:

- adversarial examples (HANS benchmark [McCoy et al., 2019])
- test sets drawn from different sources
  - ...so that surface cues could not be exploited
- uniform distribution of labels
- more transparency: publish models' predictions.

A tool to analyze benchmark datasets (work in progress):
`https://check.rusvectores.org/`

# References I


- Ethayarajh, K. and Jurafsky, D. (2020).
  Utility is in the eye of the user: A critique of NLP leaderboards.
  In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.
- McCoy, T., Pavlick, E., and Linzen, T. (2019).
  Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.
  In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

# References II


- Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Rogers, A. (2019). How the transformers broke NLP leaderboards.
- Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

# References III


- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019).
  SuperGLUE: A stickier benchmark for general-purpose language understanding systems.
  In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018).
  GLUE: A multi-task benchmark and analysis platform for natural language understanding.
  In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

# References IV


Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., and Lan, Z. (2020).
CLUE: A Chinese language understanding evaluation benchmark.
In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.