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

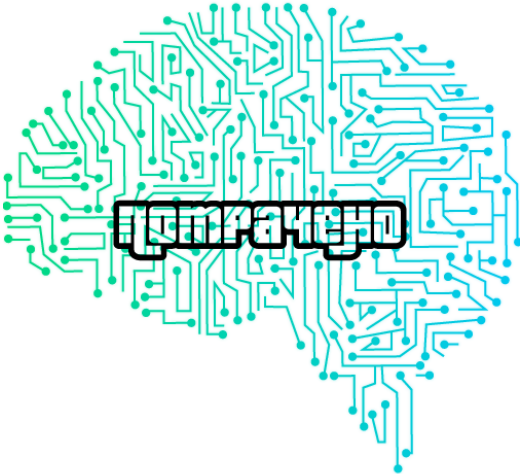
Dialogue'2021

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

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

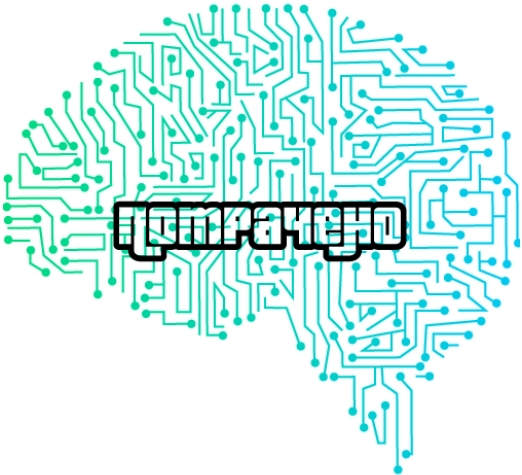
Contents

Attacking Russian SuperGLUE



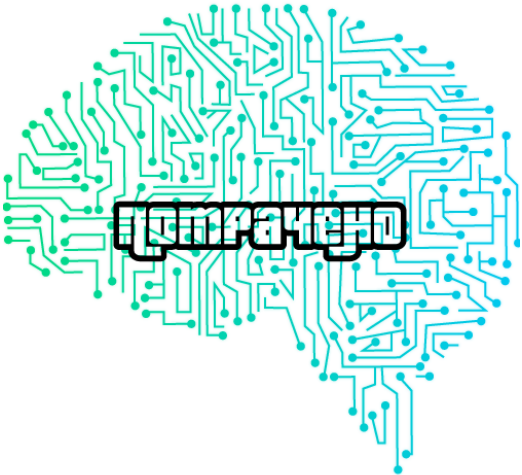
- Solve RSG with shallow heuristics

Attacking Russian SuperGLUE



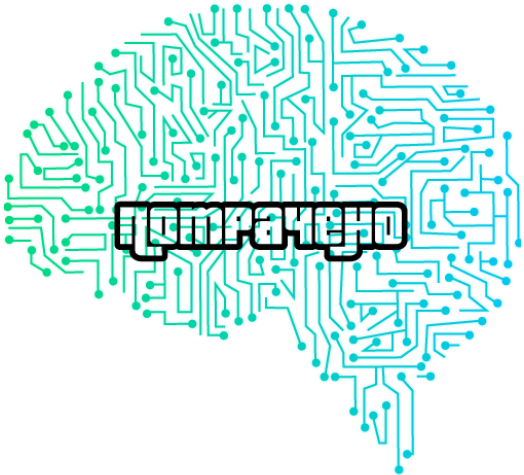
- Solve RSG with shallow heuristics
- No ML, go kozhanye meshki!

Attacking Russian SuperGLUE



- Solve RSG with shallow heuristics
- No ML, go kozhanye meshki!
- Are huge language models that good?

Attacking Russian SuperGLUE



- Solve RSG with shallow heuristics
- No ML, go kozhanye meshki!
- Are huge language models that good?
- Aim: Improve RSG datasets and contribute to Russian NLP

Previous work

Leaderboards

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

Their critique

- Statistical cues and annotation artifacts [Poliak et al., 2018]
- Why A is better than B [Ethayarajh and Jurafsky, 2020]
- Big tech only [Rogers, 2019]
- 'Right for the wrong reasons' [McCoy et al., 2019]

Russian SuperGLUE Overview

What is RSG?

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

Russian SuperGLUE Overview

What is RSG?

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

Now, what we did to solve it?

Contents

Trivial Baselines

We started with the usual dumb candidates:

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

Heuristics overview

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

Categories of heuristics detected in the benchmark:

1. interplay between the label and the number of words (e. g. ‘More than 30 words in the premise’)
2. presence of specific words (e. g. ‘Presence of ‘чтобы’, ‘будет’, ‘от’, ‘он’)
3. word forms or lemmas overlap (e. g. ’Sentences 1 and 2 use the same set of lemmas’)
4. Other task-specific heuristics.

Heuristics overview

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

Categories of heuristics detected in the benchmark:

1. interplay between the label and the number of words (e. g. ‘More than 30 words in the premise’)
2. presence of specific words (e. g. ‘Presence of ‘чтобы’, ‘будет’, ‘от’, ‘он’)
3. word forms or lemmas overlap (e. g. ’Sentences 1 and 2 use the same set of lemmas’)
4. Other task-specific heuristics.

Let’s look at some examples of heuristics.

Contents

Task Overview (result on test — 0.549)

- Natural Language Inference task
- Labels: entailment/not_entailment

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

Textual Entailment Recognition for Russian

The presence of ‘только’, ‘мужчина’ (‘only’, ‘man’) in the premise leads to 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 27 year old **man**. 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.*
(27 year old man was kidnapped.)
Label: not_entailment

The Winograd Schema Challenge (Russian)

Logic and Reasoning, World knowledge, Binary Classification: true / 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.)

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)

Logic and Reasoning, World knowledge, Binary Classification: true / 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.)

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, MT5...
- majority class predictions accuracy: ...

The Winograd Schema Challenge (Russian)

Logic and Reasoning, World knowledge, Binary Classification: true / 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.)

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, MT5...
- majority class predictions accuracy: ...
- 0.669!

The Winograd Schema Challenge (Russian)

Logic and Reasoning, World knowledge, Binary Classification: true / 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.)

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, MT5...
- majority class predictions accuracy: ...
- 0.669!
- Huge LMs are in fact just predicting the majority class.

The Winograd Schema Challenge (Russian)

Logic and Reasoning, World knowledge, Binary Classification: true / 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.)

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, MT5...
- majority class predictions accuracy: ...
- 0.669!
- Huge LMs are in fact just predicting the majority class.
- The same was observed for English [Wang et al., 2019]

Contents

Overall results

	Metrics	Human performance	SOTA	Majority class	Random	Random balanced	Heuristics + majority class
LiDiRus	M. Corr	0.626	0.231	0.000	0.024	0.000	0.147
RCB	Avg. F1 Acc.	0.680 0.702	0.452 0.546	0.217 **0.484**	0.332 0.347	0.319 0.374	0.400 0.438
PARus	Acc.	0.982	0.908	0.498	0.474	0.480	0.478
MuSeRC	F1a EM	0.806 0.420	0.941 0.819	0.000 0.000	0.477 0.078	0.450 0.071	**0.671** 0.237
TERRa	Acc.	0.920	0.871	0.513	0.503	0.483	**0.549**
RUSSE	Acc.	0.805	0.729	0.587	0.501	0.528	**0.595**
RWSD	Acc.	0.840	0.669	**0.669**	0.487	0.597	**0.669**
DaNetQA	Acc.	0.915	0.917	0.503	0.494	0.520	**0.642**
RuCoS	F1 EM	0.930 0.890	0.920 0.924	0.250 0.247	0.250 0.247	0.250 0.247	**0.260** **0.257**
Average		0.811	0.679	0.374	0.372	0.385	**0.468**

Results overview

- We did hack the RSG: heuristics are indeed effective

Results overview

- 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.

Results overview

- 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.

Results overview

- 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 strikingly higher than surface heuristics
- We need to protect our benchmarks from hacking:
  - adversarial examples
  - data from different sources

Results overview

- 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 strikingly higher than surface heuristics
- We need to protect our benchmarks from hacking:
  - adversarial examples
  - data from different sources
- Let's discuss how to address the problem!

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.

References III

- 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.
- 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.

References IV

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 V

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.