

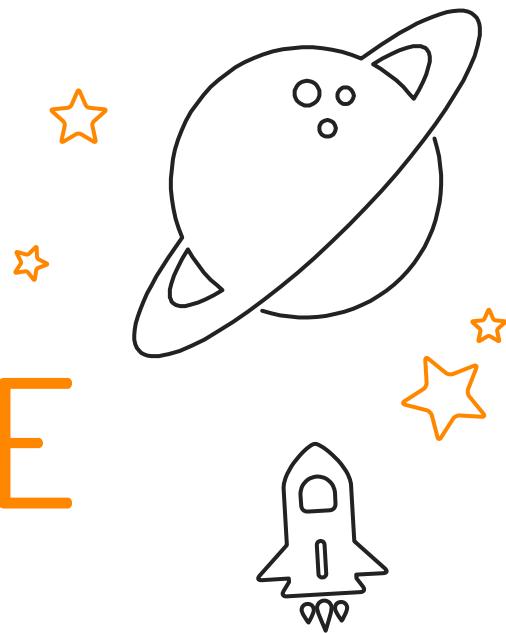
Ru_Open-Source Universe

Новые открытые ресурсы
для задач анализа русского языка



Russian SuperGLUE

A Russian Language Understanding
Evaluation Benchmark



What is general language understanding evaluation?










The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems.

GLUE consists of

- A benchmark of 9 sentence- or sentence-pair language understanding tasks
- A diagnostic dataset designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in natural language,
- A public leaderboard for tracking performance on the benchmark



Leaderboard Version: 2.0

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	Multirc	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+ 2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
4	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+ 5	Huawei Noah's Ark Lab	NEZHA-Large		83.8	85.8	93.3/95.6	91.2	78.7/42.4	87.1/86.4	88.5	73.1	90.4	58.0	87.1/74.4
+ 6	Infosys : DAWN : AI Research	RoBERTa-iCETS		77.4	84.7	88.2/91.6	85.8	78.4/37.5	82.9/82.4	83.8	69.1	65.1	35.2	93.8/68.8
7	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
8	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1	90.4/55.3
9	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
		BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7
		Most Frequent Class		47.1	62.3	21.7/48.4	50.0	61.1/0.3	33.4/32.5	50.3	50.0	65.1	0.0	100.0/50.0
		CBoW		44.5	62.2	49.0/71.2	51.6	0.0/0.5	14.0/13.6	49.7	53.1	65.1	-0.4	100.0/50.0

Russian SuperGLUE Benchmark

1. Motivation
2. Methodology
3. Collecting the data
 - a. Diagnostics
 - b. Textual Entailment, NLI
 - c. Common Sense
 - d. World Knowledge
 - e. Machine Reading
 - f. Reasoning
4. First Results



Motivation

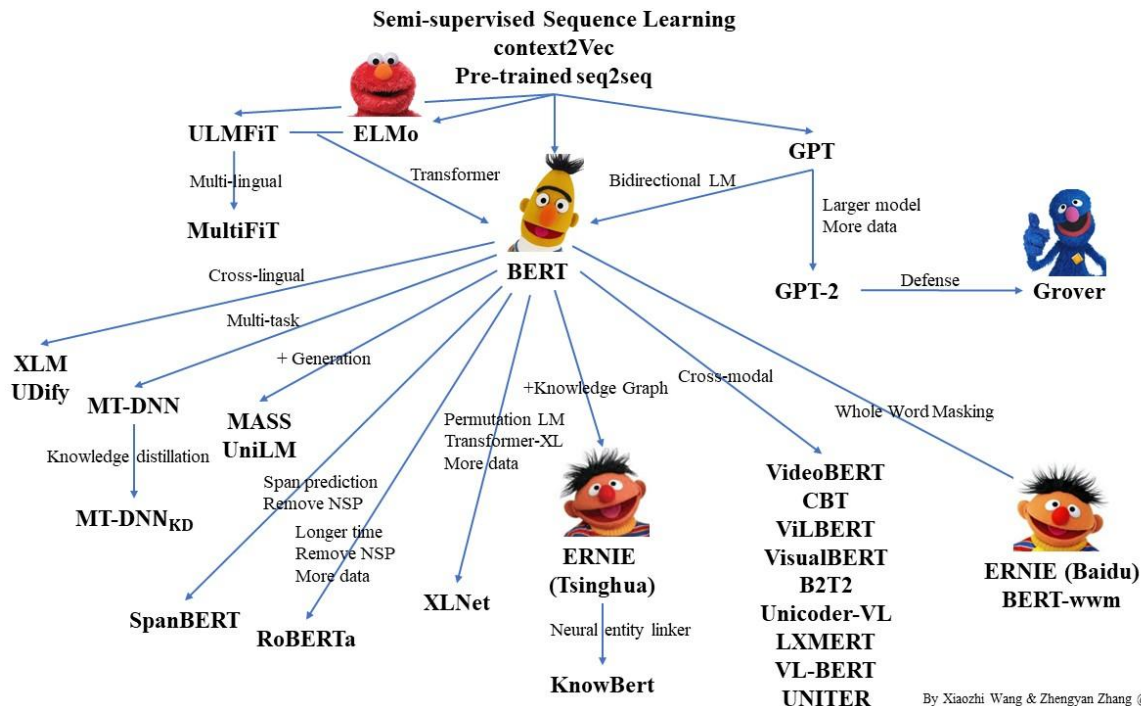
LMs went through the advanced stages of natural language modelling.

Universal transformers show ability to extract complicated relationships from texts.

Development of benchmark approach, testing general intellectual “abilities” in a text format.

Anglo-centric development of machine intelligence.

Russian suffers!



Previous Work

XNLI

The Cross-Lingual NLI Corpus (XNLI)

Alexis Conneau
Guillaume Lample
Ruty Rinott
Holger Schwenk

XGLUE

If you have questions about use of the dataset or any research outputs in your product review. For other questions, please feel free to contact us.

☐ I agree to terms and conditions. Upon accepting links to dataset will become available

Submit

XGLUE Dataset and Leaderboard

Tasks

1. NER
2. POS Tagging (POS)
3. News Classification (NC)
4. MLQA
5. XNLI
6. PAWS-X
7. Query-Ad Matching (QADSM)
8. Web Page Ranking (WPR)
9. QA Matching (QAM)
10. Question Generation (QG)
11. News Title Generation (NTG)

Relevant Links

XGLUE Submission Guideline/Github

XGLUE Paper

Unicoder Baseline

Applications

LINSPECTOR

Beta-version (Currently under test)

Language Inspector

Equal inspector to analyze word representations of your pre-trained AllenNLP models, HuggingFace embeddings for 52 languages.

Get Started

Demo Video

decaNLP

The Natural Language Decathlon

View on GitHub

Read the Paper

Blog Post

Tweet

Fork

417

GLUE



ML²



Tasks Overview

Six groups of tasks

1. Diagnostics: *LiDiRus*
2. Textual Entailment & NLI: *TERRa, RCB*
3. Common Sense: *RUSSe, PARus*
4. World Knowledge: *DaNetQA*
5. Machine Reading: *MuSeRC, RuCoS*
6. Logic: *RWSD*

[Leaderboard](#) [Tasks](#) [Diagnostic](#)

Tasks

Name	Identifier	Download	Info	Metrics
Broadcoverage Diagnostics	LiDiRus	Download	More	Matthews Corr
Russian Commitment Bank	RCB	Download	More	Avg. F1 / Accuracy
Choice of Plausible Alternatives for Russian language	PARus	Download	More	Accuracy
Russian Multi-Sentence Reading Comprehension	MuSeRC	Download	More	F1a / EM
Textual Entailment Recognition for Russian	TERRa	Download	More	Accuracy
Words in Context	RUSSE	Download	More	Accuracy
The Winograd Schema Challenge (Russian)	RWSD	Download	More	Accuracy
DaNetQA	DaNetQA	Download	More	Accuracy
Russian reading comprehension with Commonsense reasoning	RuCoS	Download	More	F1 / EM

[Download all tasks](#)

Diagnostics: LiDiRus

Linguistic Diagnostic for Russian is a diagnostic dataset covering 33 linguistic phenomena

1. **Lexical Semantics:** lexical entailment, factivity, quantifiers, named entities, symmetry or collectivity, morphological negation, redundancy;
2. **Logic:** negation and double negation, intervals or numbers, upward/downward/non- monotone, temporal, conjunction and disjunction, conditionals, universal and existential;
3. **Predicate-Argument Structure:** core args, prepositional phrases, intersectivity, restrictivity, anaphora and coreference, coordination scope, active or passive voice, ellipsis or implicits, nominalization, relative clauses, datives, genitives and partitives;
4. **Knowledge:** common sense, world knowledge

Dataset size

1104 test examples

Translation of SuperGLUE Diagnostic Dataset;

Task

- Used as another “test” for TERRa;
- Linguistic features were preserved and sentences are in one-to-one correspondence.

Testing Textual Entailment + linguistic noise

- 1) Семена, мягкие части растений и насекомые охотно поедаются южноафриканскими воробьями.
- 2) Южноафриканские воробьи охотно питаются семенами, мягкими частями растений и насекомыми

Tag: predicate-argument-structure: Active/Passive

Label: Entailment - No entailment



Testing Textual Entailment + linguistic noise

- 1) Семена, мягкие части растений и насекомые охотно поедаются южноафриканскими воробьями.
- 2) Южноафриканские воробьи охотно питаются семенами, мягкими частями растений и насекомыми

Tag: predicate-argument-structure: Active/Passive

Label: Entailment - No entailment



Testing Textual Entailment + linguistic noise

- 1) Пчёлы летают не по тем же правилам аэродинамики, что и самолёты.
- 2) Пчёлы летают более энергоэффективно, чем самолёты.

Tag: knowledge: Common sense

Label: Entailment - No entailment



Testing Textual Entailment + linguistic noise

- 1) Пчёлы летают не по тем же правилам аэродинамики, что и самолёты.
- 2) Пчёлы летают более энергоэффективно, чем самолёты.

Tag: knowledge: Common sense

Label: Entailment - No entailment



Testing Textual Entailment + linguistic noise

- 1) Я не могу достичь своих целей каждый год, начиная с 1997, а сейчас 2008.
- 2) Я не достиг своих целей в 2004 году.

Tag: logic: Intervals/Numbers

Label: Entailment - No entailment



Testing Textual Entailment + linguistic noise

- 1) Я не могу достичь своих целей каждый год, начиная с 1997, а сейчас 2008.
- 2) Я не достиг своих целей в 2004 году.

Tag: logic: Intervals/Numbers

Label: Entailment - No entailment



Natural Language Inference: TERRa

The dataset is an analogue of *RTE* dataset in SuperGLUE.

Dataset size

2616 train / 307 val / 3198 test examples

Data Source

Data extracted from *Taiga web-corpus*

Task

given two text fragments (premise and hypothesis), say whether the meaning of one text is entailed from the other

Example

Premise: Автор поста написал в комментарии, что прорвалась канализация.

Hypothesis: Автор поста написал про канализацию.

Label: Entailment

Natural Language Inference: RCB

Task

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

Data source

Dataset corresponds to CommonBank dataset. We filtered out Taiga with a number of rules and manually post processed the extracted passages.

Dataset size

438/220/348

2715

Example

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

Common Sense: RUSSE

Task

Given two sentences, each containing an anchor word, define, whether the word is used in the same sense, or not.

Data source

Re-use of RUSSE dataset [Panchenko et al., 2018], in which ambiguous words are annotated with sense labels.

Dataset size

19845 train, 8508 dev, 12151 test examples

Example

Context 1: Бурые ковровые дорожки заглушали шаги.

Context 2: Приятели решили выпить на дорожку в местном баре.

Sense match: False

Panchenko, A., Lopukhina, A.,
Ustalov, D., Lopukhin, K., Arefyev, N.,
Leontyev, A., Loukachevitch, N.:
[RUSSE'2018: A Shared Task on
Word Sense Induction for the
Russian Language](#). (2018)

Common Sense: PARus

PARus is an analogue of COPA in SuperGLUE.

Task

Choose the best of plausible alternatives for the question based on the text.

Data source

PARus is constructed as a translation of COPA dataset from SuperGLUE and edited by professional editors. The data split from COPA is retained.

Dataset size

500/100/400

1000 sentences

Example

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

World Knowledge: DaNetQA

Task

Given a passage, answer a yes/no question to it.

Data source

- 1) Crowdsourced questions are used as queries to Wikipedia
- 2) Wikipedia pages are retrieved via Google API
- 3) Passages are retrieved by Deep Pavlov SQuAD models
- 4) Crowd workers are used to answer the questions based on the passages

Dataset size

800 train, 200 dev, 200 test examples; 562 (~59%) unique questions

Example

- **Passage:** В период с 1969 по 1972 год по программе «Аполлон» было выполнено 6 полётов с посадкой на Луне.
- **Question:** Был ли человек на луне?
- **Answer:** Yes

Both yes and no answers are possible on the same question. The passages needs to be carefully read.

Machine Reading: MuSeRC

MuSeRC is an analogue of MultiRC in SuperGLUE.

Task

Reading comprehension challenge, questions can be answered only based on multiple sentences from the paragraph.

Dataset size

500/100/322

Data source

+800 paragraphs ~6k questions

5 different domains collected from open sources:

- 1) elementary school texts
- 2) news
- 3) fiction stories
- 4) fairy tales
- 5) brief annotations of TV series and books

Example

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. (Т), На 0,3 секунды. (Т), На секунду. (F), На секунды. (F)

Machine Reading: RuCoS

The dataset is based on ReCoRD methodology from SuperGLUE.

Dataset size

72193 train / 4370 val / 4147 test;

Data source

Lenta & Deutsche Welle

Task

Find the correct entity in the paragraph that best fits the placeholder in the query.

Example

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

Logic: RWSD

This dataset is a trancreation of Winograd Schema Challenge

Dataset size

606 train/204 val/154 test pairs of sentences

Data source

Editing and translation of the Winograd Schema Challenge

Task

Each sentence has two objects for coreference and segment markup

Example

Text: Кубок не помещается в коричневый чемодан, потому что он слишком большой.

Coreference: True

Experiments: baselines

Naive Baseline

TF-IDF model on 20 thousand sample from Wikipedia + Logistic Regression

Advanced Baselines

- 1) *Multilingual BERT (MultiBERT)* - a single language model pre-trained in 104 languages
- 2) *Russian BERT (RuBERT)* trained on large-scale corpus of news and Wikipedia in Russian

Human evaluation

All tasks were solved by Yandex.Toloka annotators' majority vote.

Experiments: results

Russian GLUE

[Leaderboard](#)
[Tasks](#)
[Diagnostic](#)
[FAQ](#)
[Our team](#)
[Lo](#)

Leaderboard

Rank	Name	Team	Info	Score	Diagnostic	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	AGI NLP	i	0.802	0.626	0.68/0.702	0.982	0.806/0.42	0.92	0.747	0.84	0.879	0.93/0.924
2	RuBERT conversational	AGI NLP	i	0.546	0.186	0.432/0.468	0.61	0.656/0.256	0.639	0.894	0.675	0.749	0.255/0.251
3	Multilingual BERT	AGI NLP	i	0.542	0.157	0.365/0.425	0.588	0.626/0.253	0.62	0.84	0.675	0.79	0.371/0.367
4	Baseline TF-IDF	AGI NLP	i	0.372	-0.004	0.288/0.395	0.522	0.477/0.03	0.496	0.632	0.338	0.763	0.0/0.002

Experiments: results

Dataset	Metrics	ConvBERT	MultiBERT	TF-IDF	Human
<i>LiDiRus</i>	<i>MCC</i>	0.186	0.157	0.059	0.626
<i>RCB</i>	<i>F1/Acc.</i>	0.432/0.468	0.383/0.429	0.45	0.68/0.702
<i>PARus</i>	<i>Acc</i>	0.61	0.588	0.48	0.982
<i>MuSeRC</i>	<i>F1/EM</i>	0.656/0.256	0.626/0.253	0.589/0.244	0.806/0.42
<i>TERRa</i>	<i>Acc</i>	0.639	0.62	0.47	0.92
<i>RUSSE</i>	<i>Acc</i>	0.894	0.84	0.66	0.747
<i>RWSD</i>	<i>Acc</i>	0.675	0.675	0.66	0.84
<i>DaNetQA</i>	<i>Acc</i>	0.749	0.79	0.68	0.879
<i>RuCoS</i>	<i>F1/EM</i>	0.255/0.251	0.371/0.367	0.256/0.251	0.93/0.924
<i>Average</i>		0.546	0.542	0.461	0.802

Comparison to SuperGLUE

Diagnostic gives a possibility to compare models in English and Russian.

Sequential MultiBERT pre-training on RTE and TERRa and testing on two diagnostics.

Observations

- 1) the English model performs slightly better,
- 2) some categories are much better solved in one language and fail in the other.

For comprehensive analysis other linguistic features should be used!

	English	Russian
Overall MCC	0.2	0.15
Named entities	0.17	0.28
Redundancy	0	-0.58
Factivity	0.37	0.68
Morphological negation	0.033	0.056
Lexical entailment	0.02	0
Quantifiers	0.12	-0.12
Coordination scope	0	0.28
Anaphora/Coreference	-0.047	0
Ellipsis/Implicits	0.28	0
Intersectivity	0.21	0
Genitives/Partitives	0.45	0
Nominalization	0	0
Prepositional phrases	0.47	0
Relative clauses	0.21	0.21
Active/Passive	0.38	0.48
Datives	0.64	0.28
Core args	0.28	0.2
Restrictivity	0	0.33
Temporal	-0.24	0.13
Upward monotone	0.16	0
Conditionals	0.39	0.084
Negation	0.081	0
Existential	0.15	1
Conjunction	0.38	0.14
Double negation	0.1	0.072
Intervals/Numbers	-0.19	0
Non-monotone	-0.21	0
Disjunction	0.26	0
Universal	0.24	0.12
Downward monotone	0.012	-0.25
Knowledge	0.17	0.11
Common sense	0.071	0.024

Conclusion

- First benchmark on *General Language Understanding* evaluation for Russian.
- 8 (9) novel datasets for the Russian language covering a wide scope of NLU tasks
- Baselines and human evaluation to compare your model with.
- Platform for testing models in Russian



Our team



Denis Shevelev



Alena Fenogenova



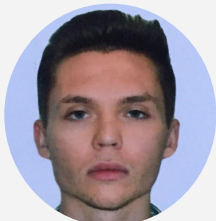
Tatiana Shavrina



Anton Emelyanov



Maria Tikhonova



Vladislav Mikhailov



Ekaterina Artemova



Valentin Malykh



Andrey Evlampiev



Taisia Glushkova

See you on the leaderboard!

Join us at russiansuperglue.com



NATIONAL RESEARCH
UNIVERSITY



SBERBANK



HUAWEI

Thank you!