russian superglue

Sberbank
Artificial Intelligence

Russian SuperGLUE

Creating a Language Understanding Evaluation Benchmark

T.Shavrina

Sberbank, AGI NLP team

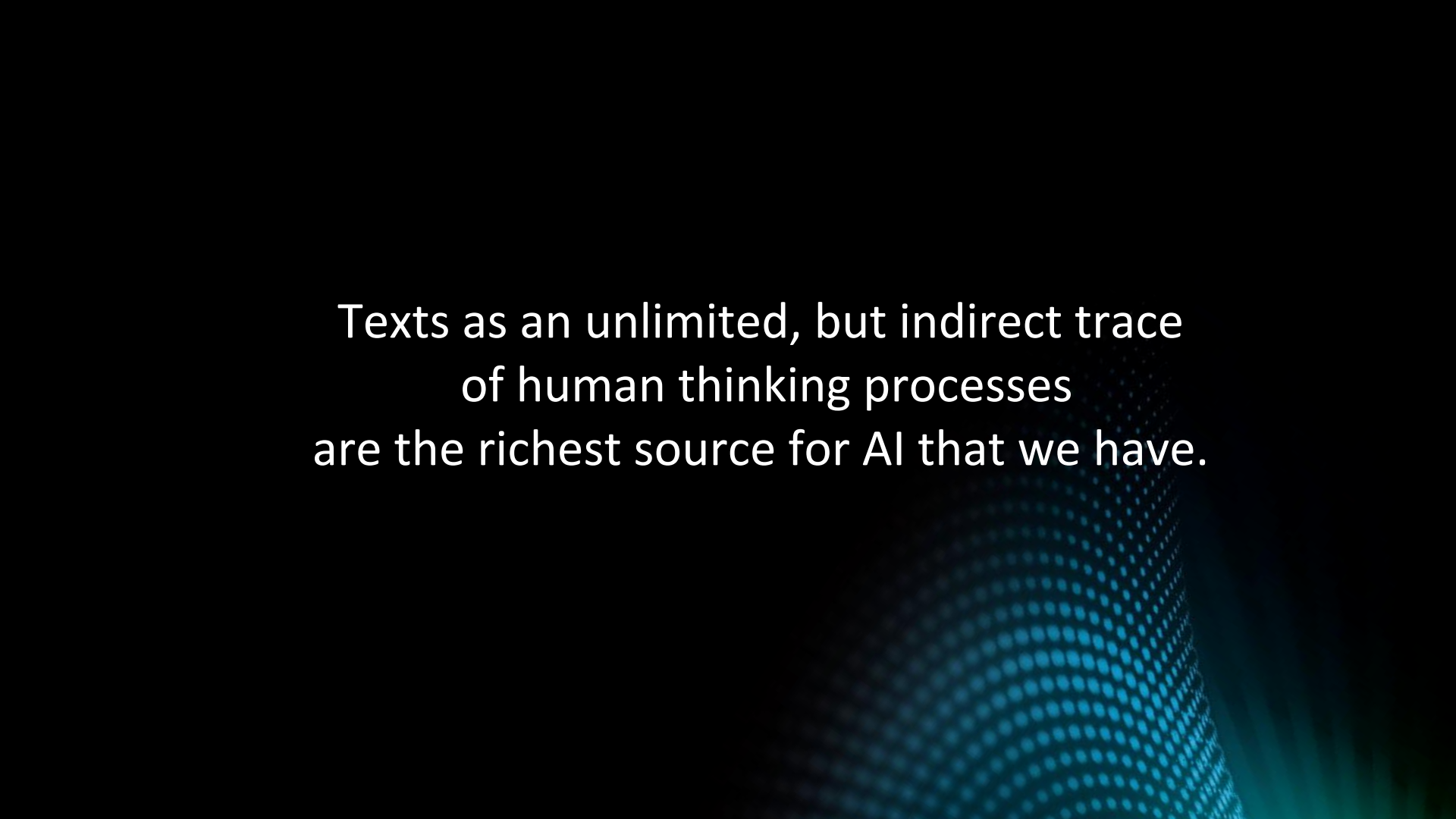
What is general language understanding evaluation?

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

GLUE consists of

- Modelling human intellectual operations on textual tasks
- A benchmark of 9 sentence- or sentence-pair NLU 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

Texts as an unlimited, but indirect trace
of human thinking processes
are the richest source for AI that we have.



Things we can find in the text:

- common objects and abstractions
- actions
- entailment, contradiction, temporal relations, etc
- empathy
- commonsense
- pure knowledge

NLP

Theory of Mind

Things we cannot find in the text:

- Linguistic typology:
  - every language has something that cannot be said
  - temporal relations and predicates are not at all universal
  - unwanted associations

- Cognition and processes are universal!

Motivation

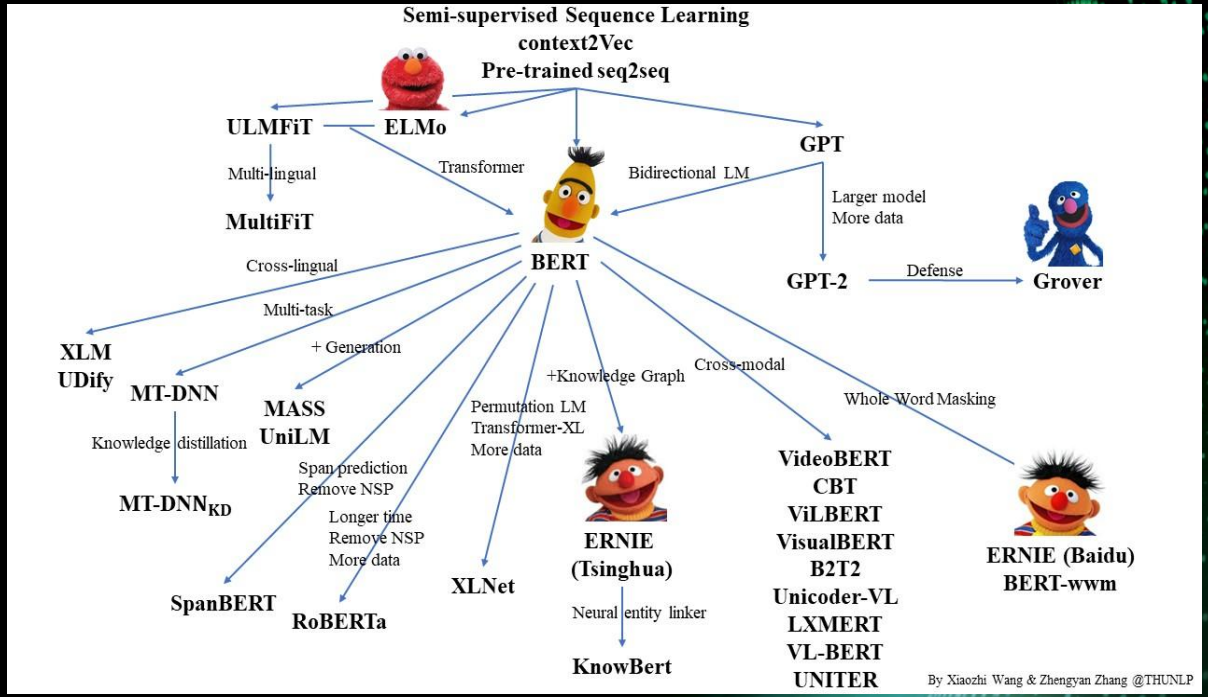
Anglo-centric development of machine intelligence.

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.

Other languages suffer!



SuperGLUE

SuperGLUE GLUE | Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

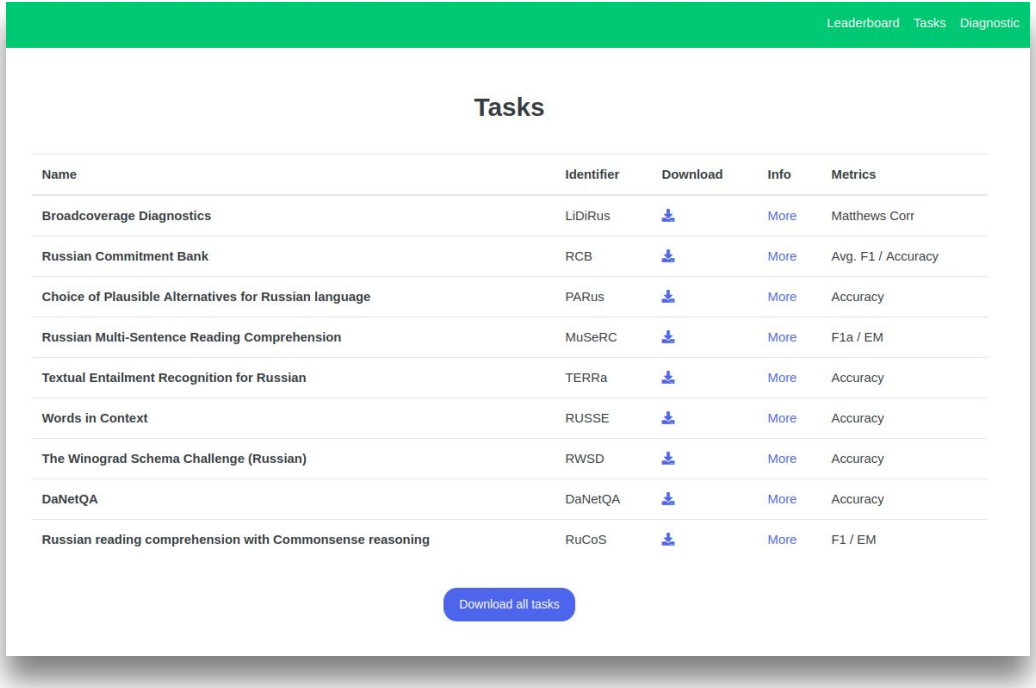
Leaderboard Version: **2.0**

	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

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*



8 new tasks for Russian from scratch

Translation + adaptation

- Diagnostics
- COPA (Choice of Plausible Alternatives)
- WSD (Winograd Schema)

Corpus Filtration +Manual Correction&Annotation

- RUSSE (WiC)
- RCB (Rus Commitment Bank)
- TERRa (Textual Entailment Recognition)

Web Crawling + Automatic Filtration + Manual Annotation

- DaNetQA (BoolQ)
- MuSeRC (Multi-Sentence Reading Comprehension)
- RuCoS (Russian RC with Commonsense)

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

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

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- mono-tone, 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

Text: The pupils of the House of Youthful Technical Creativity designed the fairytale sled. They are based on a tank sled that the pupils made earlier, they just cut off the turret, built a new corpus and painted it with Khokhloma - since we are facing the year of the Rooster.

Label: Entailment - No entailment

Testing Textual Entailment + linguistic noise

Text: The pupils of the House of Youthful Technical Creativity designed the fairytale sled. They are based on a tank sled that the pupils made earlier, they just cut off the turret, built a new corpus and painted it with Khokhloma - since we are facing the year of the Rooster.

Label: Entailment - No entailment

(Khokhloma national ornament casually contains peacocks,

flowers and other natural motifs)

Natural Language Inference: TERRa

Example

- **Premise:** The author of the post wrote in a comment that the sewage system had broken through.
- **Hypothesis:** *The author of the post wrote about the sewage system.*
- **Label:** Entailment

World Knowledge: DaNetQA

Example

- **Passage:** In the period from 1969 to 1972, according to the Apollo program, 6 flights were performed with landing on the moon.
- **Question:** *Was there a man on the moon?*
- **Answer:** Yes

Baselines

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

Advanced Baselines

- Multilingual BERT (MultiBERT) - a single language model pre-trained in 104 languages
- 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.

Current leaderboard

russian superglue

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		**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		**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		**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	Plain RuBERT	DeepPavlov		**0.524**	-0.026	0.338/0.393	0.532	0.712/0.309	0.636	0.877	0.662	0.78	0.38/0.379
5	Baseline TF-IDF	AGI NLP		**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

Adding Some More Linguistics to the GLUE

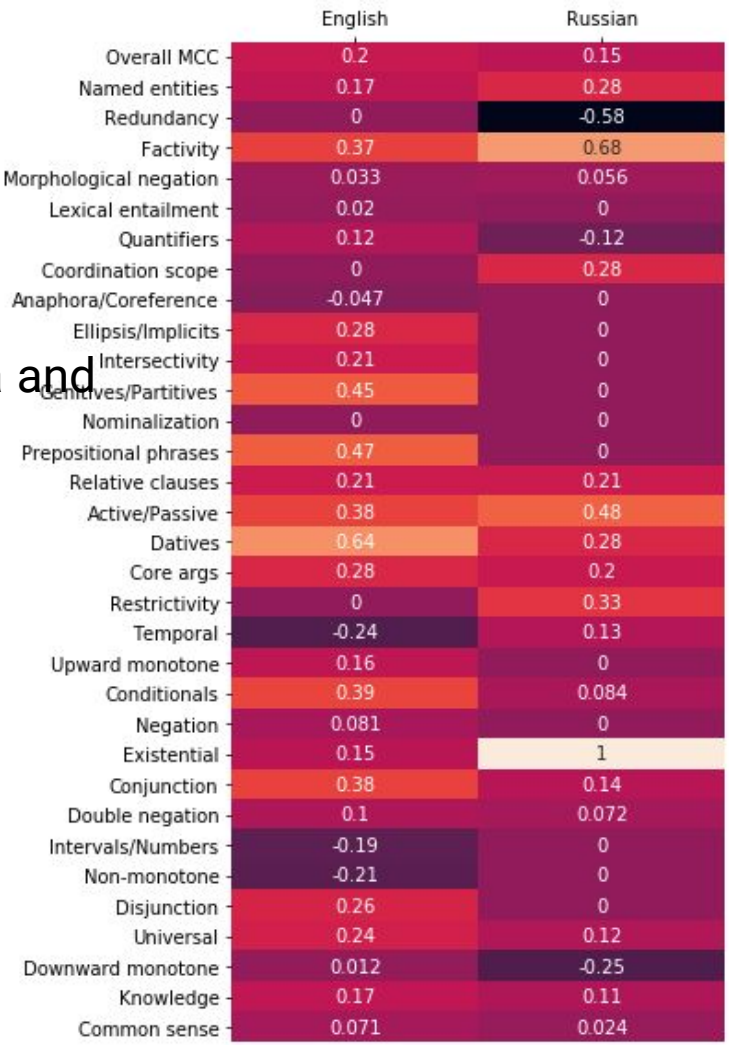
Comparison to SuperGLUE

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!



Sources of typology - WALS.info

russian superglue

Home | Features | Chapters | Languages | References | Authors

Feature 126A: 'When' Clauses

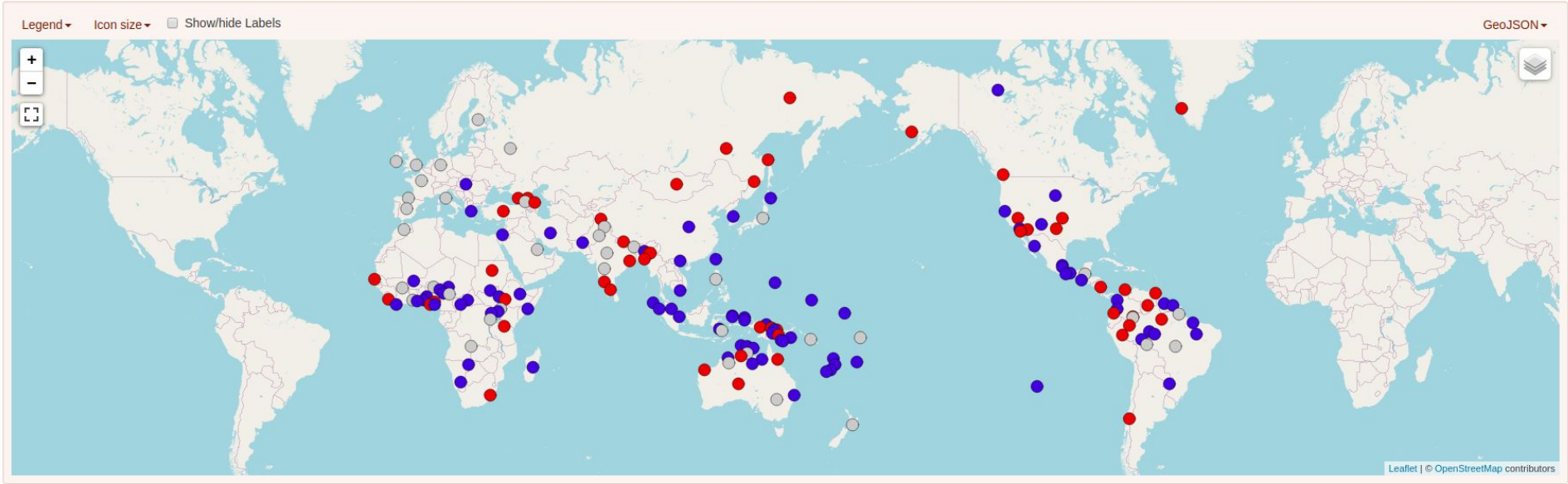
This feature is described in the text of chapter 126 'When' Clauses by Sonia Cristofaro cite

You may combine this feature with another one. Start typing the feature name or number in the field below.

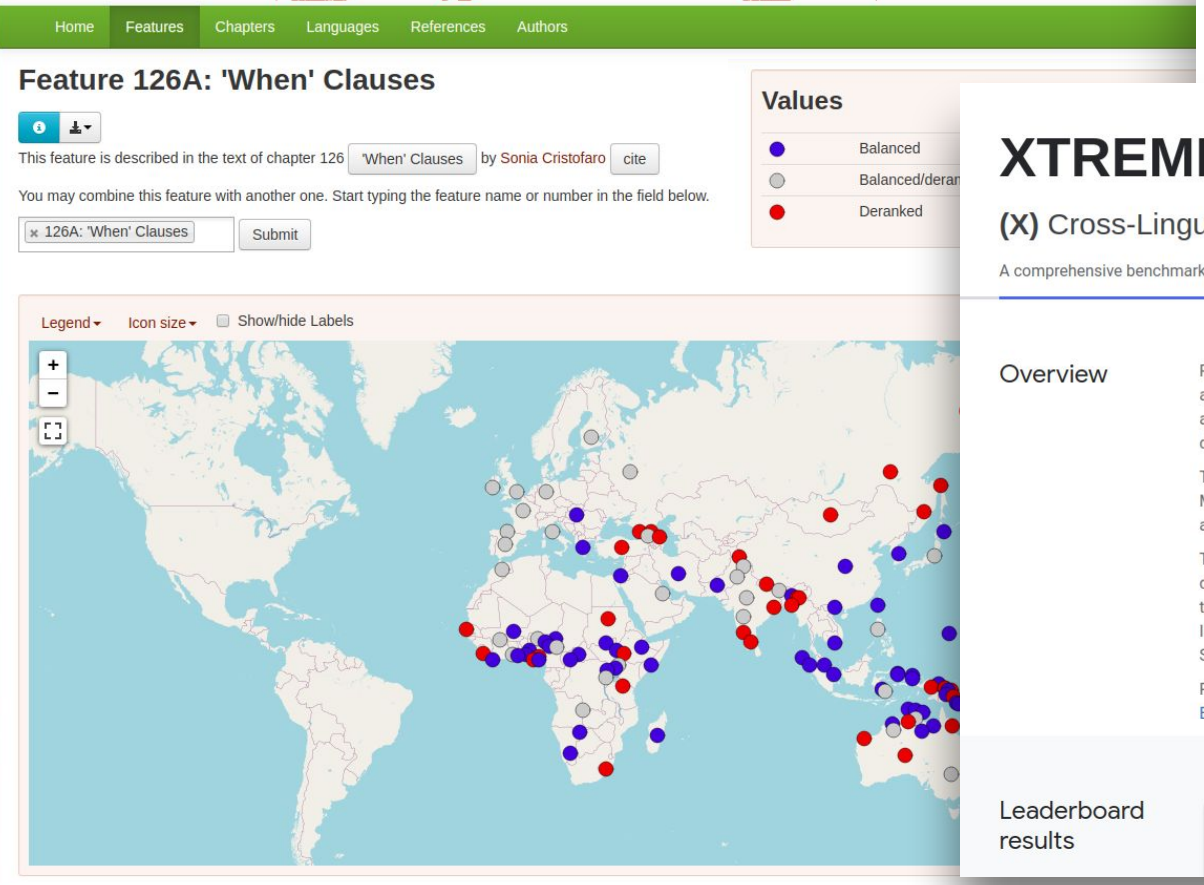
× 126A: 'When' Clauses Submit

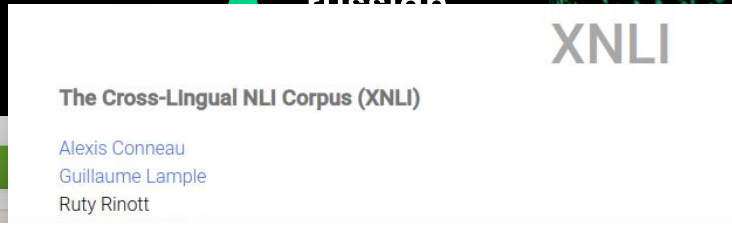
Values

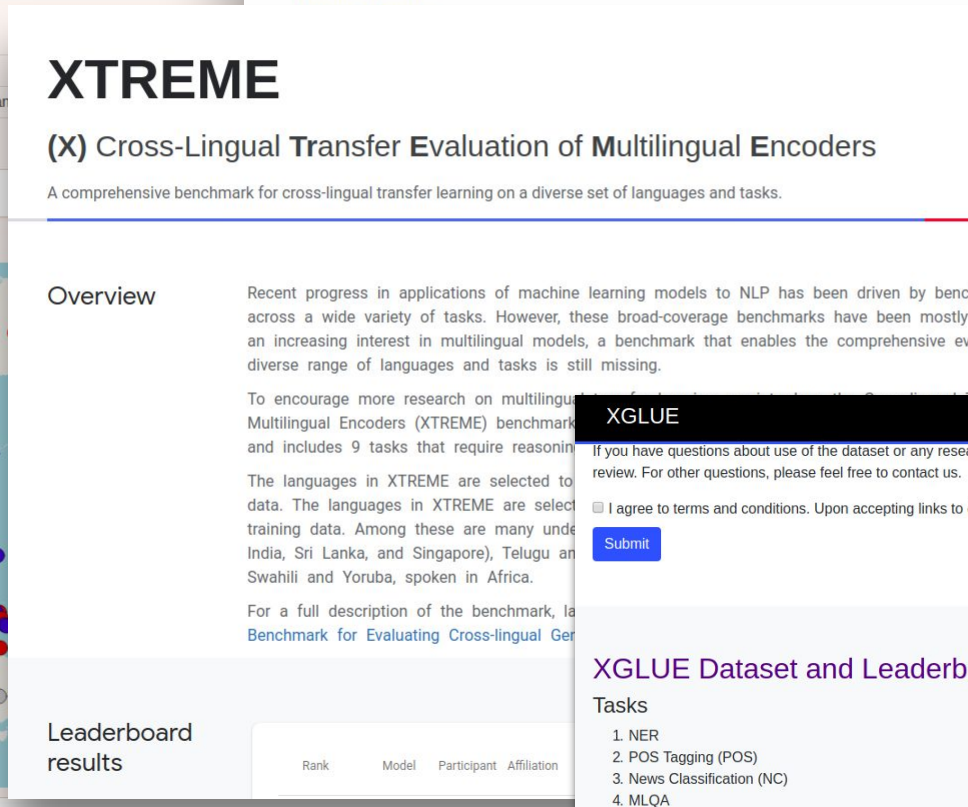
●	Balanced	84
●	Balanced/deranked	39
●	Deranked	51



Sources of typology







Diagnosing Model Stability with Linguistically-Aware Adversarial Attacks

russian superglue

33 linguistic categories included are _not_ universal

Do these categories affect the quality of the models equally?

Linguistically-Aware Adversarial Attacks:

original sentence + sentence piece with additional category not changing the meaning

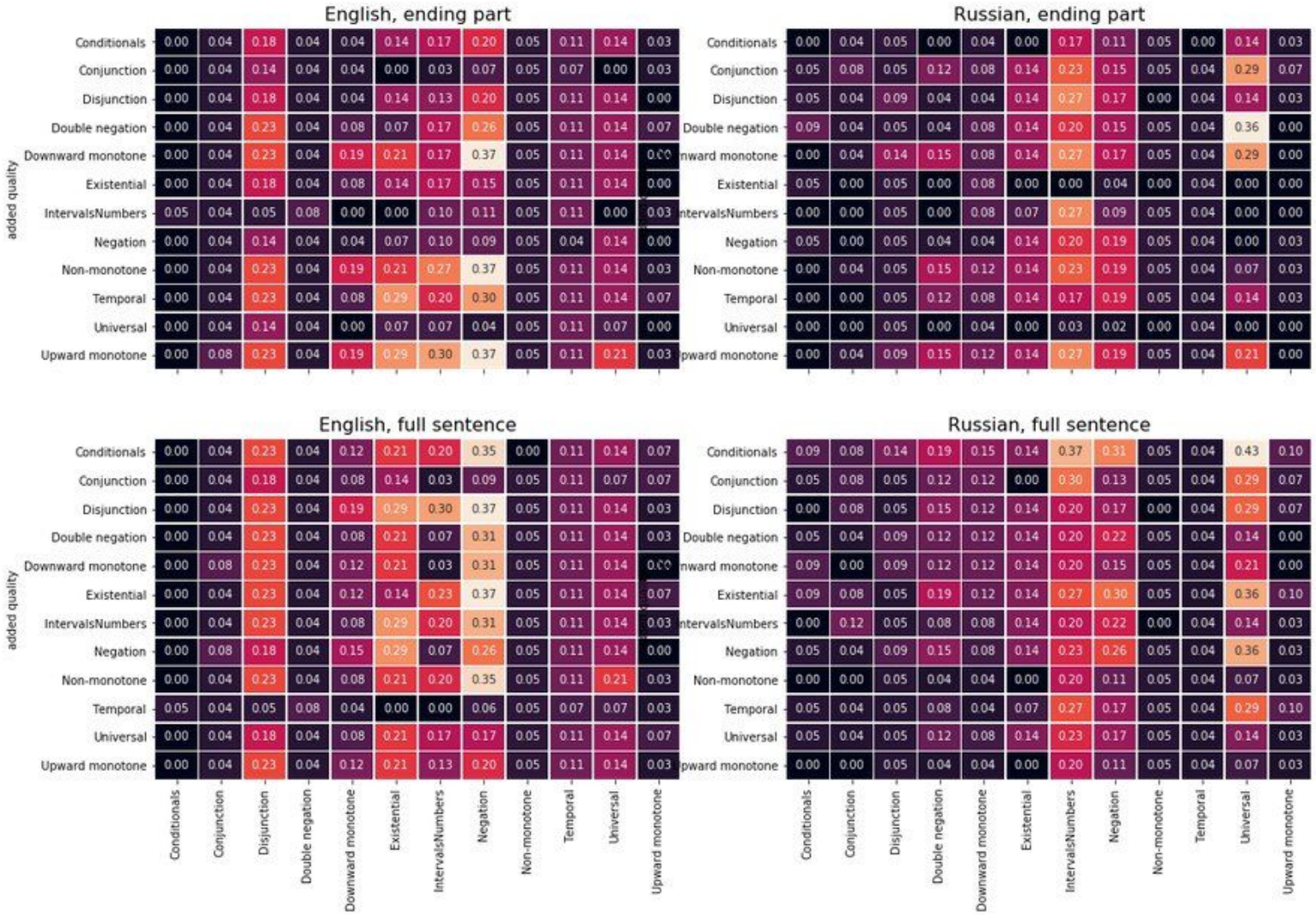
testing on NLI task

\+ PrepPhrase

The market is about to get harder, but possible to navigate. +”on Sunday”

”On Sunday”+ The market is about to get harder, but not impossible to navigate.

Label: entailment


English, ending part
Russian, ending part
English, full sentence
Russian, full sentence
added quality
Conditionals
Conjunction
Disjunction
Double negation
Downward monotone
Existential
IntervalsNumbers
Negation
Non-monotone
Temporal
Universal
Upward monotone

Conclusion

- First benchmark on *General Language Understanding* evaluation for Russian +platforms + new tasks
- Are we on our way to multilingual models transferring knowledge and abilities?
- Texts should be collected more diversely to be a better source for AI

Exploring language-specific categories affecting stability:

	Common categories	Specific for English	Specific for Russian
Logic	Negation, IntervalsNumbers, Universal	Temporal, Disjunction, Existential	
Lexical semantics	Factivity, Quantifiers	Named entities	
Predicate argument structure	Relative clauses, EllipsisImplicits	Coordination scope, Prepositional phrases	AnaphoraCoreference, Datives
Knowledge	Common sense	World knowledge	

See you on the leaderboard!

Join us at russiansuperglue.com

russian
superglue