Все способы измерить слона: индустриальные метрики трансформеров, ИИ-тесты, пробинг

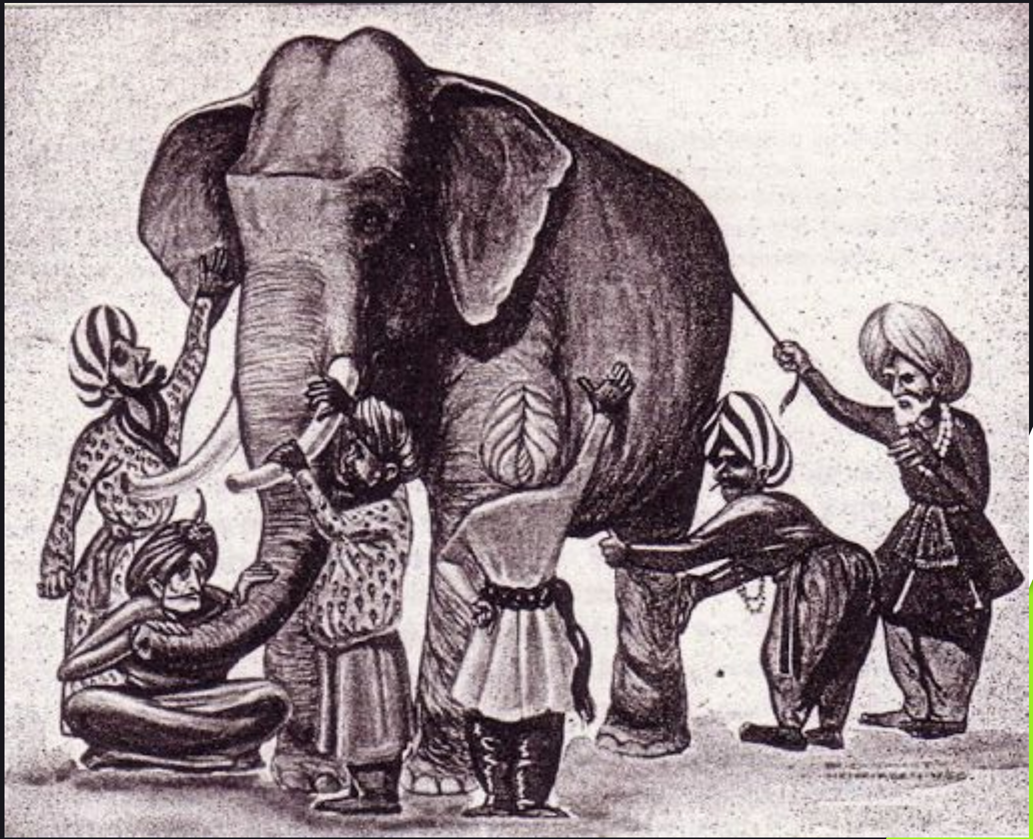
Татьяна Шаврина
AGI NLP, Sberdevices

Model Zoo

BERT, GPT-3...

pretrained models

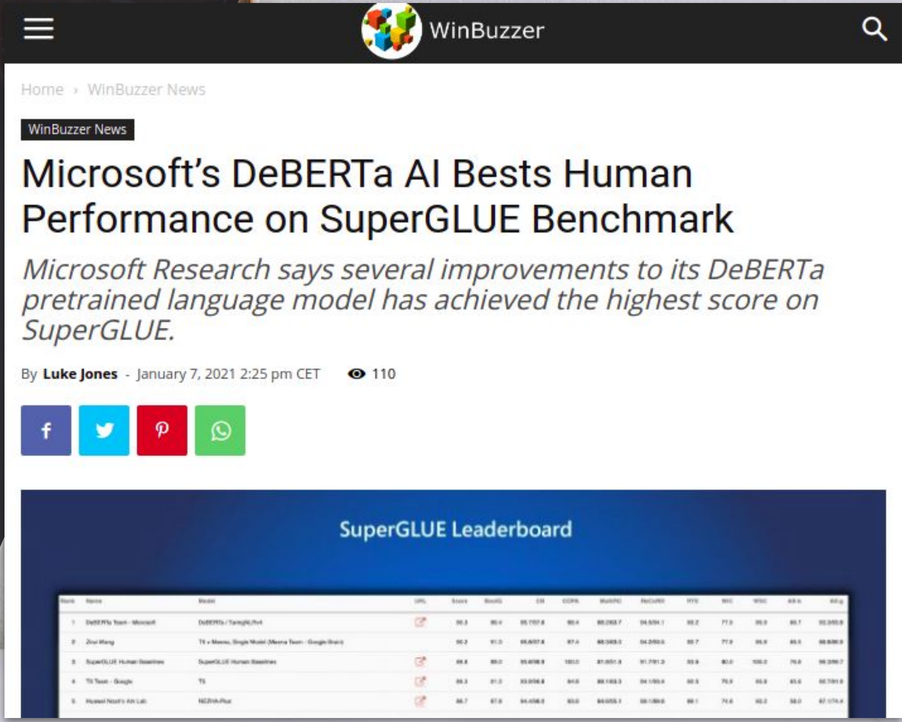
universal abilities to
recreate human skills



Transformers are all we need

SOTA results with transformers:

- Open-Domain Question Answering
- Sentiment Classification
- Machine Translation
- Text Generation
- Named Entity Recognition
- Reading Comprehension
- General Language Understanding
- and much more...



First Models on Russian SuperGLUE

* More information about speed sc

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQ
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
2	RuGPT3XL few-shot	sberdevices	i	**0.535**	0.096	0.302 / 0.418	0.676	0.74 / 0.546	0.573	0.565	0.649	0.59
3	MT5 Large	AGI NLP	i	**0.528**	0.061	0.366 / 0.454	0.504	0.844 / 0.543	0.561	0.633	0.669	0.657
4	RuBERT plain	DeepPavlov	i	**0.521**	0.191	0.367 / 0.463	0.574	0.711 / 0.324	0.642	0.726	0.669	0.639
5	RuGPT3Large	sberdevices	i	**0.505**	0.231	0.417 / 0.484	0.584	0.729 / 0.333	0.654	0.647	0.636	0.604
6	RuBERT conversational	DeepPavlov	i	**0.5**	0.178	0.452 / 0.484	0.508	0.687 / 0.278	0.64	0.729	0.669	0.606
7	Multilingual Bert	DeepPavlov	i	**0.495**	0.189	0.367 / 0.445	0.528	0.639 / 0.239	0.617	0.69	0.669	0.624
8	heuristic majority	ling_ling	i	**0.468**	0.147	0.4 / 0.438	0.478	0.671 / 0.237	0.549	0.595	0.669	0.642
9	RuGPT3Medium	sberdevices	i	**0.468**	0.01	0.372 / 0.461	0.598	0.706 / 0.308	0.505	0.642	0.669	0.634
10	RuGPT3Small	sberdevices	i	**0.438**	-0.013	0.356 / 0.473	0.562	0.653 / 0.221	0.488	0.57	0.669	0.61
11	Baseline TF-IDF1.1	AGI NLP	i	**0.434**	0.06	0.301 / 0.441	0.486	0.587 / 0.242	0.471	0.57	0.662	0.621

NTI AI Hackaton

Искусственный интеллект

01. Задача

Реши задачи по NLP лучше других и докажи, что достоин забрать главный приз

02. Чат

Общаемся, обсуждаем новости, задаем вопросы организаторам в Телеграмм чате

03. Рейтинг

rubert_conv_dp_notlower	rubert_sen_dp_lower
германии	германии
хамас	рф
ташкента	ташкента
сми	heckler & koch
франк-вальтер штайнмайер	франк-вальтер штайнмайер
хафтар	халифа хафтар
россии	хафтара
мазиной	мазиной

```
о один из наших методов ансамблирования
S. уважаемые программисты, не делайте сразу фейспа
гласен, это сложно назвать нормальным ансамблем
 оно повышает точность и использует несколько моде
 ап = []
```

Загрузка обученных нами моделей

```
[ ]: # with open('/content/drive/MyDrive/New_models/bert_f
     #     bert_xquad_notlower = pickle.load(f)
     # with open('/content/drive/MyDrive/New_models/distil
     #     distilbert_notlower = pickle.load(f)
     with open('/content/drive/MyDrive/dpmlbert_cased_lowe
         bert_dp_lower = pickle.load(f)
     with open('/content/drive/MyDrive/dprubertconv_cased_
         rubert_conv_dp_lower = pickle.load(f)
     with open('/content/drive/MyDrive/dprubertconv_cased_
         rubert_conv_dp_notlower = pickle.load(f)
     with open('/content/drive/MyDrive/model_rubert_low.pk
         rubert_lower = pickle.load(f)
     with open('/content/drive/MyDrive/model_rubert_senten
         rubert_sen_lower = pickle.load(f)
     with open('/content/drive/MyDrive/model_rubert_no.pkl
         rubert_sen_dp_lower = pickle.load(f)

     with open('/content/drive/MyDrive/finalized_model_ber
         bert_fin = pickle.load(f)
```

А дальше начинается сущий ад и куча методов ансамблирования предиктов бертов, которые мы придумали

```
[ ]: # получение предиктов каждого берта
```

NTI AI Hackaton

03. Рейтинг

Обр. хакатон | Отборочный | **Финал**

Место	Имя	Результат
1	Avengers Ensemble	0.9313
2	Братва рвется в топ	0.8847
3	Спутник-V	0.8753
4	Почему Берт выдаёт единички	0.8693
5	{team_name}	0.8693
6	RuGpT3	0.864
7	Arima	0.8573
8	The AI Gang	0.8533
9	Ninja Turtles	0.8447
10	NTI: Become chelovek	0.8333

First Models on Russian SuperGLUE

* More information about speed sc

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	Avengers Ensemble	i	**0.679**	0.0	0.406 / 0.546	0.908	0.941 / 0.819	0.871	0.587	0.545	0.917	0.92 / 0.924
3	RuGPT3XL few-shot	sberdevices	i	**0.535**	0.096	0.302 / 0.418	0.676	0.74 / 0.546	0.573	0.565	0.649	0.59	0.67 / 0.665
4	MT5 Large	AGI NLP	i	**0.528**	0.061	0.366 / 0.454	0.504	0.844 / 0.543	0.561	0.633	0.669	0.657	0.57 / 0.562
5	RuBERT plain	DeepPavlov	i	**0.521**	0.191	0.367 / 0.463	0.574	0.711 / 0.324	0.642	0.726	0.669	0.639	0.32 / 0.314
6	RuGPT3Large	sberdevices	i	**0.505**	0.231	0.417 / 0.484	0.584	0.729 / 0.333	0.654	0.647	0.636	0.604	0.21 / 0.202
7	RuBERT conversational	DeepPavlov	i	**0.5**	0.178	0.452 / 0.484	0.508	0.687 / 0.278	0.64	0.729	0.669	0.606	0.22 / 0.218
8	Multilingual Bert	DeepPavlov	i	**0.495**	0.189	0.367 / 0.445	0.528	0.639 / 0.239	0.617	0.69	0.669	0.624	0.29 / 0.29
9	heuristic majority	ling_ling	i	**0.468**	0.147	0.4 / 0.438	0.478	0.671 / 0.237	0.549	0.595	0.669	0.642	0.26 / 0.257
10	RuGPT3Medium	sberdevices	i	**0.468**	0.01	0.372 / 0.461	0.598	0.706 / 0.308	0.505	0.642	0.669	0.634	0.23 / 0.224
11	RuGPT3Small	sberdevices	i	**0.438**	-0.013	0.356 / 0.473	0.562	0.653 / 0.221	0.488	0.57	0.669	0.61	0.21 / 0.204
12	Baseline TF-IDF1.1	AGI NLP	i	**0.434**	0.06	0.301 / 0.441	0.486	0.587 / 0.242	0.471	0.57	0.662	0.621	0.26 / 0.252
13	Random weighted	ling_ling	i	**0.385**	0.0	0.319 / 0.374	0.48	0.45 / 0.071	0.483	0.528	0.597	0.52	0.25 / 0.247
14	majority_class	ling_ling	i	**0.374**	0.0	0.217 / 0.484	0.498	0.0 / 0.0	0.513	0.587	0.669	0.503	0.25 / 0.247

BERTology

BERTology
Ways to look into the black void

What does Bertology do?

- accessing all the hidden-states of BERT/GPT/GPT-2,
- accessing all the attention weights for each head of BERT/GPT/GPT-2, 3...
- retrieving heads output values and gradients to be able to compute head importance score
- probing! evaluate layer representations

What does Bertology do?

- accessing all the hidden-states of BERT/GPT/GPT-2,
- accessing all the attention weights for each head of BERT/GPT/GPT-2, 3...
- retrieving heads output values and gradients to be able to compute head importance score
- probing! evaluate layer representations

Docs » BERTology

BERTology

There is a growing field of study concerned with investigating the inner working of large-scale transformers like BERT (that some call "BER
field are:

- BERT Rediscovers the Classical NLP Pipeline by Ian Tenney, Dipanjan Das, Ellie Pavlick: https://arxiv.org/abs/1905.05950
- Are Sixteen Heads Really Better than One? by Paul Michel, Omer Levy, Graham Neubig: https://arxiv.org/abs/1905.10650
- What Does BERT Look At? An Analysis of BERT's Attention by Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning: ht

In order to help this new field develop, we have included a few additional features in the BERT/GPT/GPT-2 models to help people access th
from the great work of Paul Michel (https://arxiv.org/abs/1905.10650):

Но что находится внутри русских BERTов?

🤗 🤔

RuSentEval framework

probing Russian models

RuSentEval - First Russian Probing

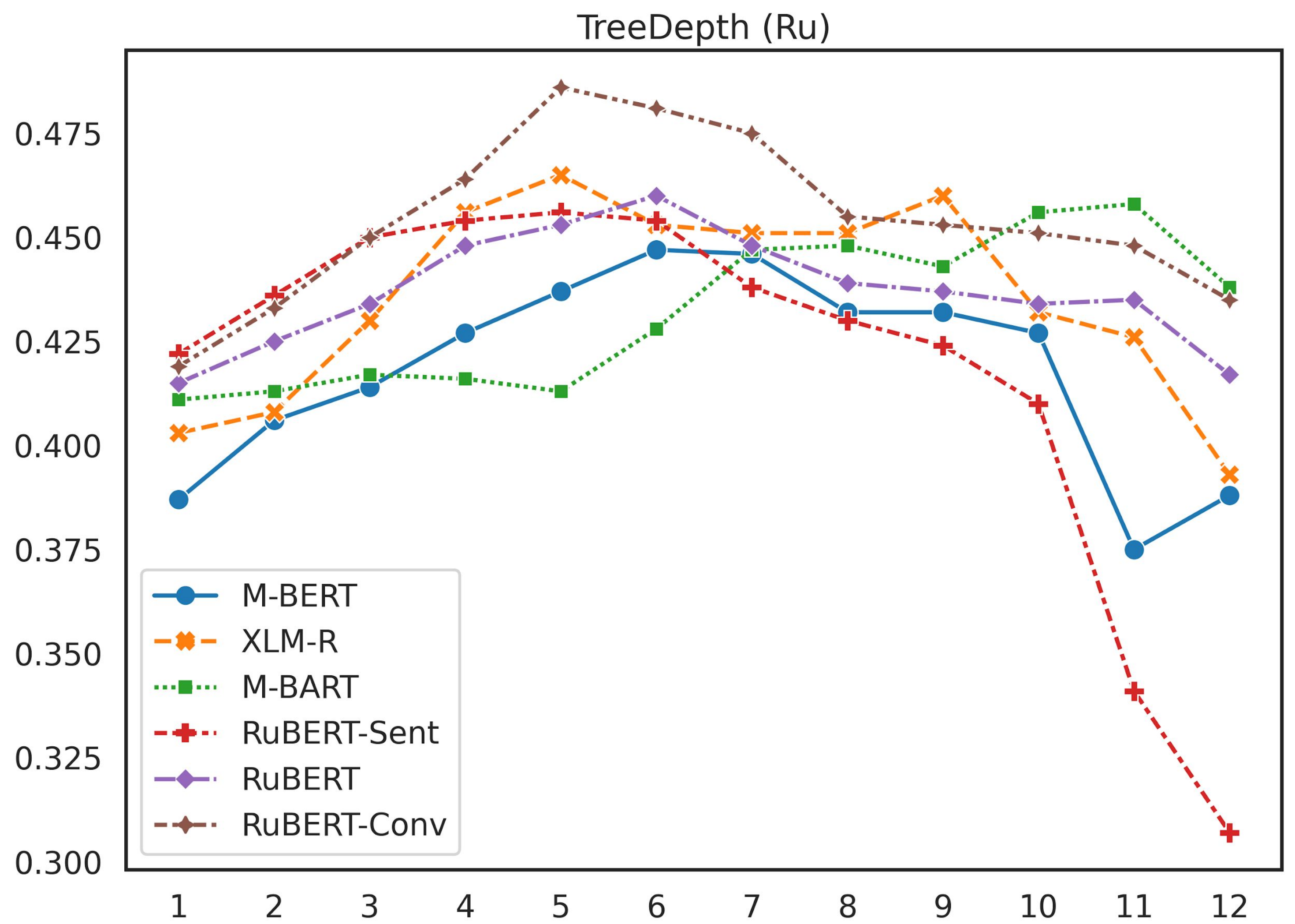
BERT-like models - source of embeddings:

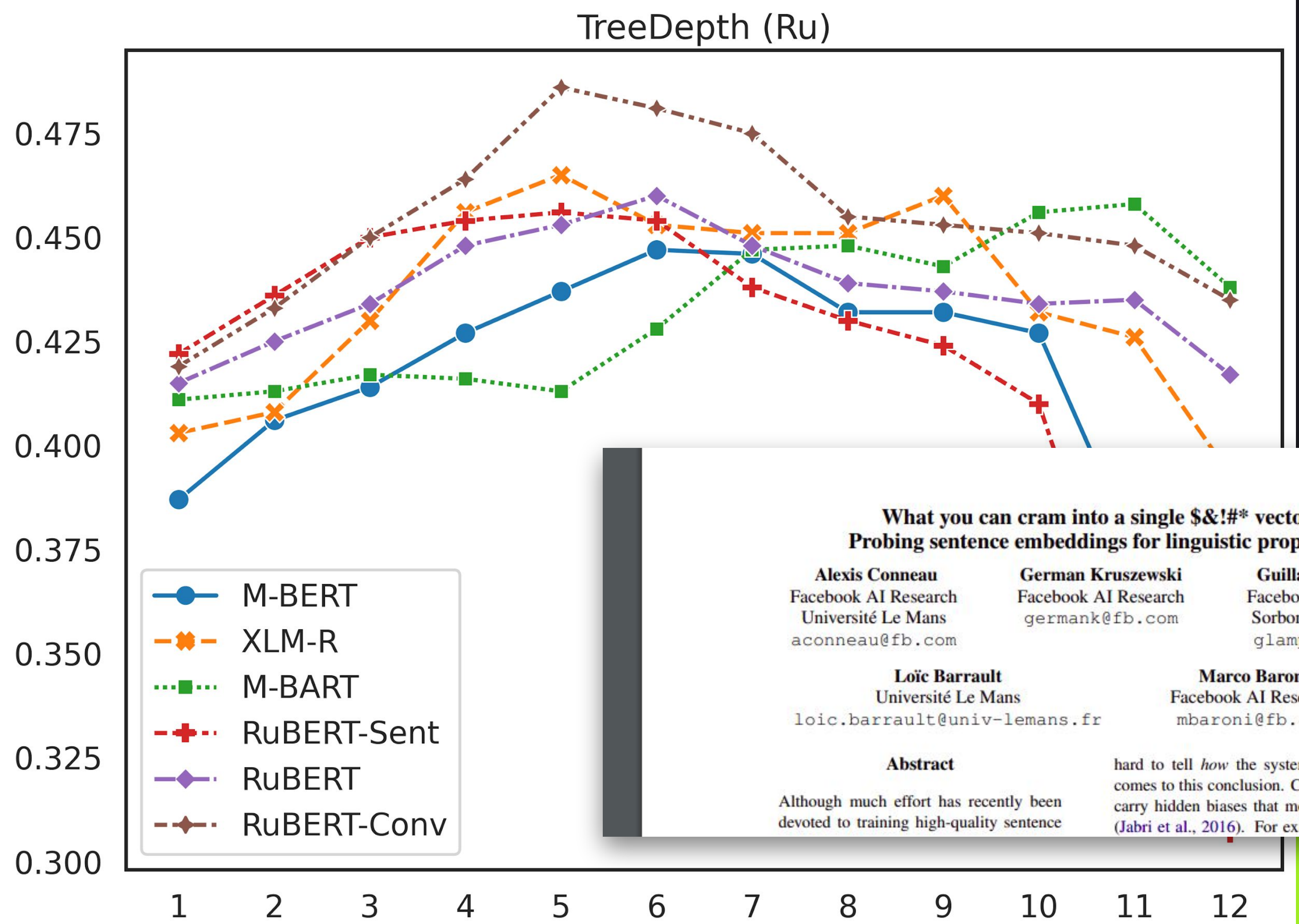
word embeddings, sentence embeddings

How do we distinguish the good embeddings from the bad?
Probing!

Let’s use annotated Russian sentences and get their embeddings from different layers from the model

- + a simple classifier on top
- + sentence annotation on embeddings
- + bad classification quality = no info in embeddings = bad embeddings


TreeDepth (Ru)
0.475
0.450
0.425
0.400
0.375
0.350
0.325
0.300
1
2
3
4
5
6
7
8
9
10
11
12
M-BERT
XLM-R
M-BART
RuBERT-Sent
RuBERT
RuBERT-Conv



What you can cram into a single $&!#* vector:
Probing sentence embeddings for linguistic properties

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

German Kruszewski
Facebook AI Research
germank@fb.com

Guillaume Lample
Facebook AI Research
Sorbonne Universités
glample@fb.com

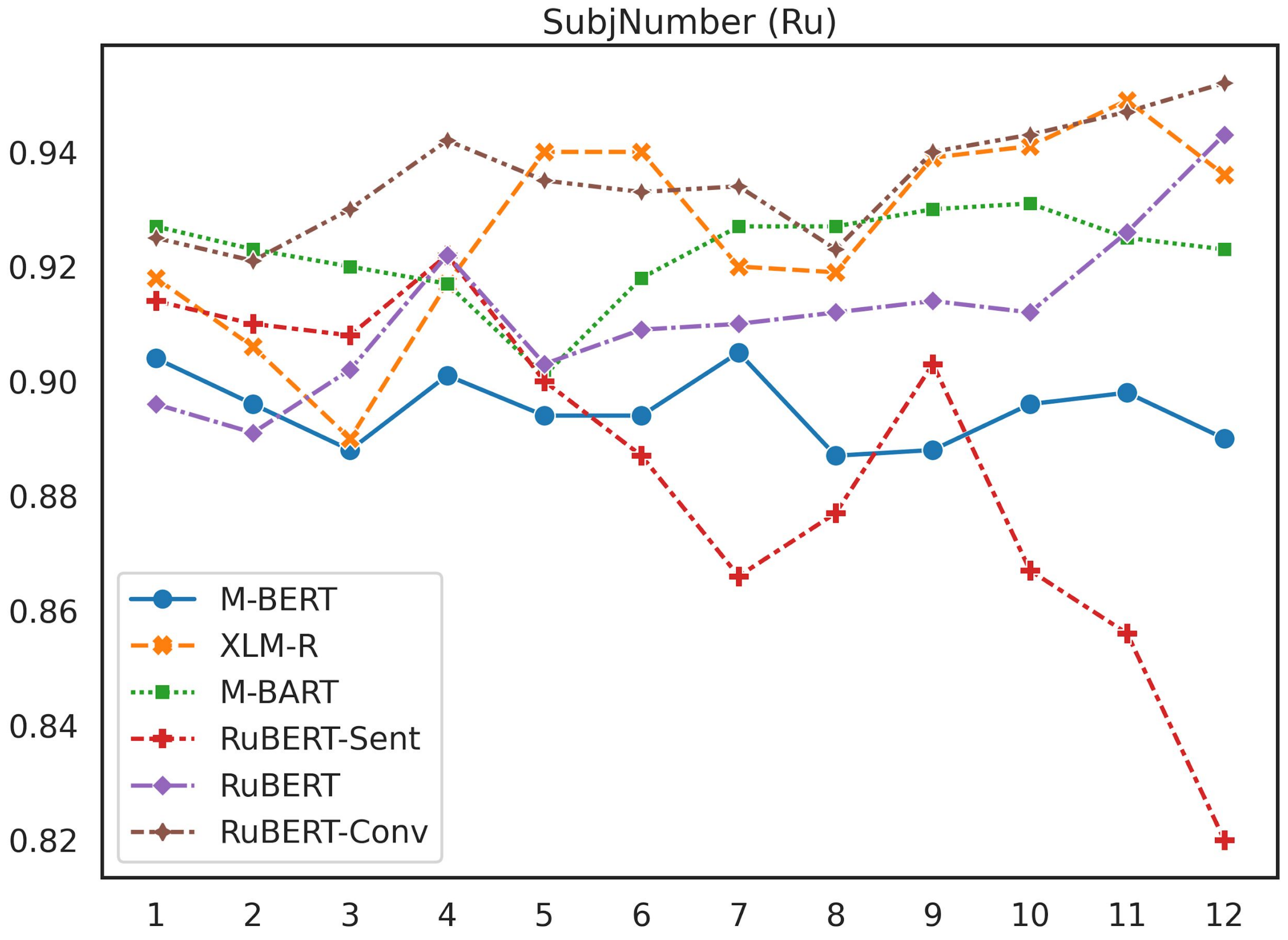
Loïc Barrault
Université Le Mans
loic.barrault@univ-lemans.fr

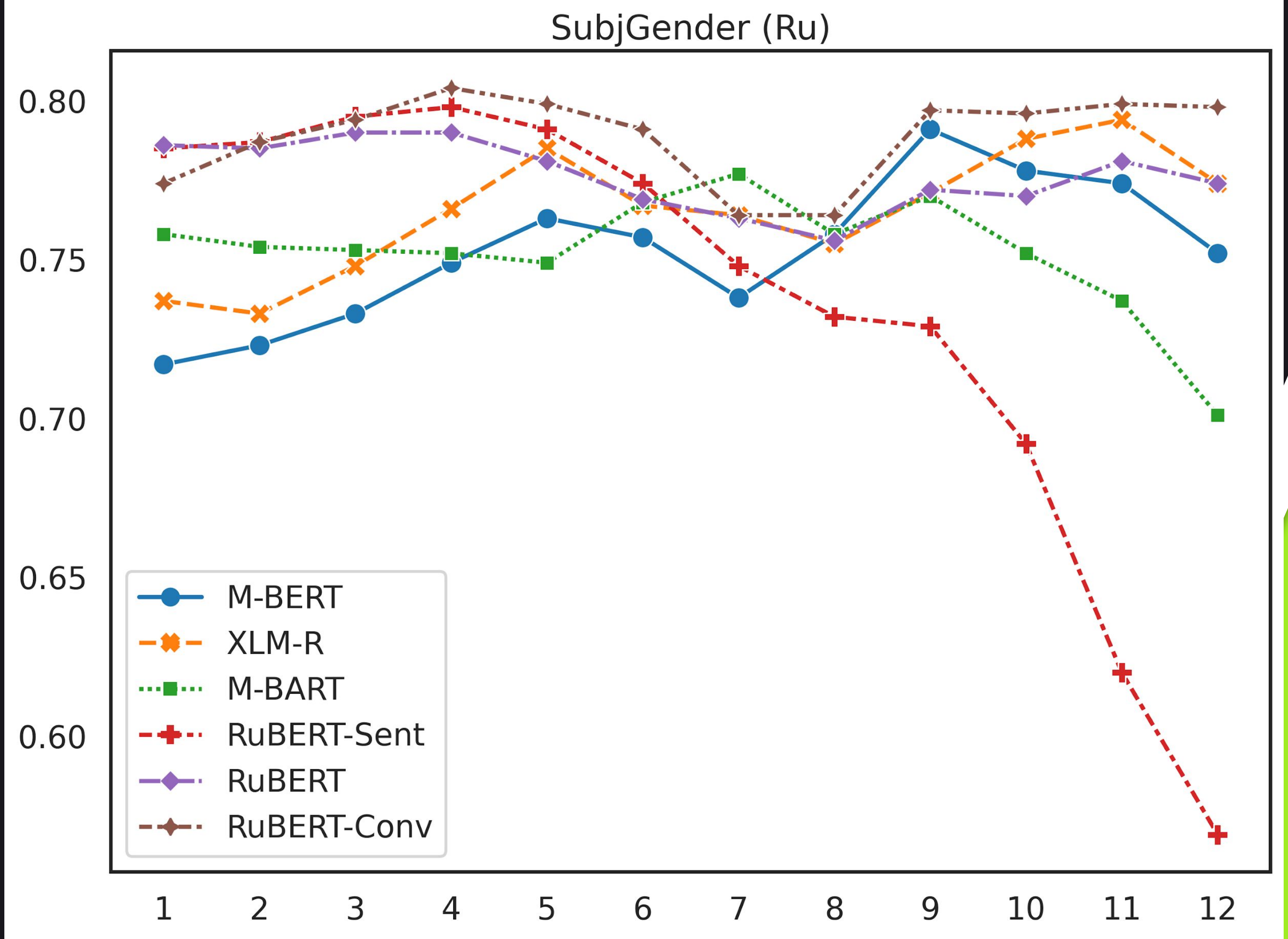
Marco Baroni
Facebook AI Research
mbaroni@fb.com

Abstract

Although much effort has recently been devoted to training high-quality sentence

hard to tell *how* the system (or even a human) comes to this conclusion. Complex tasks can also carry hidden biases that models might lock onto (Jabri et al., 2016). For example, Lai and Hock-


SubjNumber (Ru)
0.94
0.92
0.90
0.88
0.86
0.84
0.82
1
2
3
4
5
6
7
8
9
10
11
12
M-BERT
XLM-R
M-BART
RuBERT-Sent
RuBERT
RuBERT-Conv


SubjGender (Ru)
0.80
0.75
0.70
0.65
0.60
1
2
3
4
5
6
7
8
9
10
11
12
M-BERT
XLM-R
M-BART
RuBERT-Sent
RuBERT
RuBERT-Conv

Multilingual Models

Probing Task	Language	M-BERT	LABSE	XLM-R	MiniLM	M-BART
Nshift	Ru	84.8 [8]	82.6 [5]	**86.9 [9]**	80.5 [9]	78.6 [12]
	En	81.8 [10]	84.4 [5]	**85.7 [10]**	79.3 [8]	83.8 [12]
ObjNumber	Ru	82.8 [6]	82.5 [2]	**83.7 [10]**	77.8 [10]	81.5 [7]
	En	**86.2 [6]**	85.4 [3]	86.0 [8]	85.2 [6]	85.9 [9]
SentLen	Ru	91.3 [2]	93.3 [1]	94.5 [2]	94.1 [2]	**96.2 [4]**
	En	96.3 [2]	96.6 [1]	95.8 [2]	96.1 [3]	**97.3 [3]**
SubjNumber	Ru	90.5 [7]	92.9 [3]	**94.9 [11]**	94.2 [12]	93.1 [10]
	En	87.8 [7]	**90.7 [12]**	86.9 [10]	85.6 [6]	87.3 [9]
Tense	Ru	99.5 [8]	**99.8 [5]**	**99.8 [5]**	98.2 [7]	99.6 [7]
	En	88.9 [8]	88.8 [6]	88.8 [9]	87.3 [5]	**89.1 [9]**
TreeDepth	Ru	44.7 [6]	46.1 [4]	**46.5 [5]**	44.8 [7]	45.8 [11]
	En	41.2 [5]	**42.7 [5]**	41.8 [7]	40.9 [7]	41.2 [12]
WC	Ru	84.8 [2]	85.8 [1]	82.6 [1]	72.8 [1]	**88.0 [1]**
	En	92.6 [1]	93.7 [1]	89.8 [1]	82.3 [1]	**93.8 [1]**

Table 1: Results of Logistic Regression classifier for each encoder over the shared English and Russian tasks. Languages: **Ru**=Russian, **En**=English.

github.com/RussianNLP/rusenteval

RussianNLP / **rusenteval**

Unwatch

Code | Issues 1 | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

main · 1 branch · 0 tags · Go to file · Add file · Code

vmkhlv Update README.md · 7868207 · 5 hours ago · **15** commits

data	added tasks	4 months ago
images	Add files via upload	19 hours ago
probing	added rusenteval code	6 days ago
README.md	Update README.md	5 hours ago
install_tools.sh	added rusenteval code	6 days ago
requirements.txt	added rusenteval code	6 days ago

About

No description, website, or topics provided.

Readme

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Contributors 3

TatianaShavrina Tatiana Shavrina

vmkhlv Vlad Mikhailov

artemovae Katya Artemova

README.md

RuSentEval

Linguistic Source, Encoder Force!

RuSentEval is an evaluation toolkit for sentence embeddings for Russian.

In this repo you can find the data and scripts to run an evaluation of the quality of sentence embeddings.

RuSentEval, an enhanced set of 14 probing tasks for Russian, including ones that have not been explored yet. We

MOROCCO framework

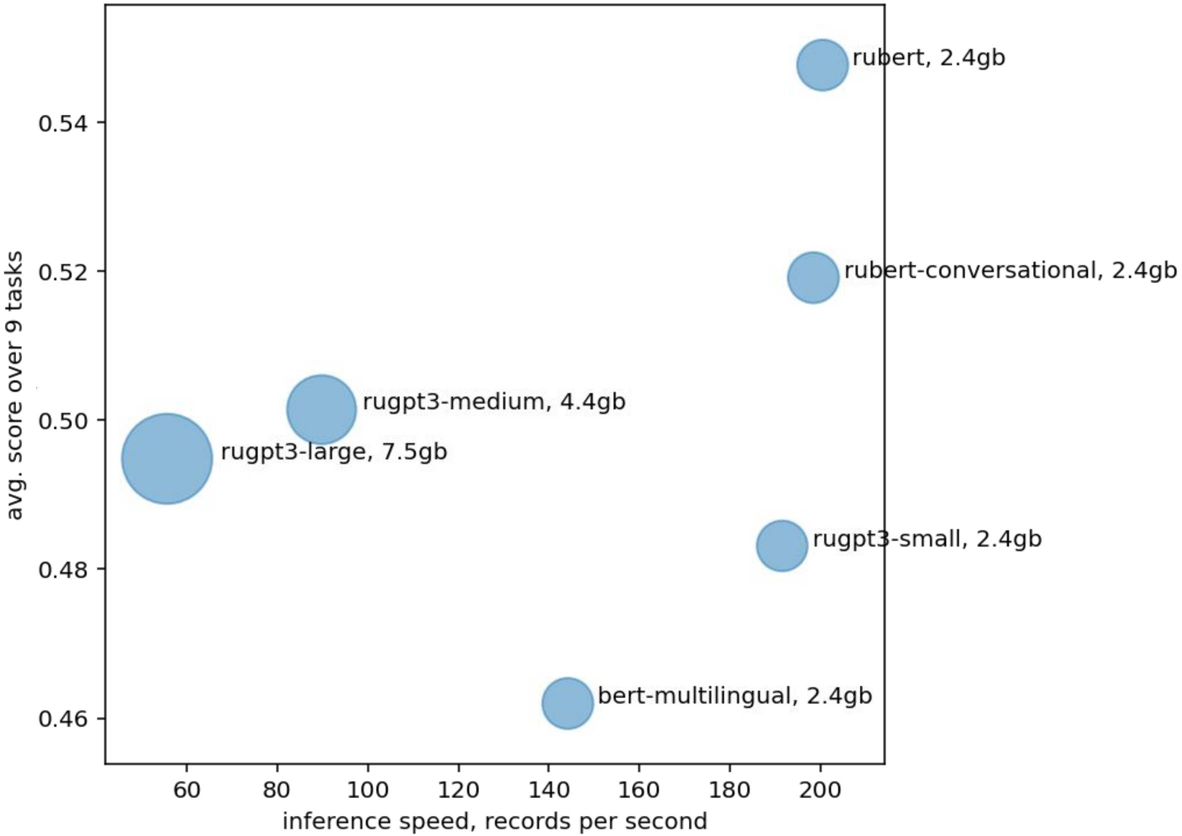
MOdel ResOurCe COnsumption

MOROCCO idea

Lets evaluate models by GPU RAM usage + inference speed + Russian SuperGLUE score

- Model results on GLUE are unstable, depend on random seed
- Smaller models have higher inference speed. `rugpt3-small` processes ~200 records per second while `rugpt3-large` — ~60 records/second.
- `bert-multilingual` is a bit slower then `rubert*` due to worse Russian tokenizer. `bert-multilingual` splits text into more tokens, has to process larger batches.
- It is common that larger models show higher score but in our case `rugpt3-medium`, `rugpt3-large` perform worse then smaller `rubert*` models.
- `rugpt3-large` has more parameters then `rugpt3-medium` but is currently trained for less time and has lower score.

Russian Models by inference speed and performance



Спасибо за attention!

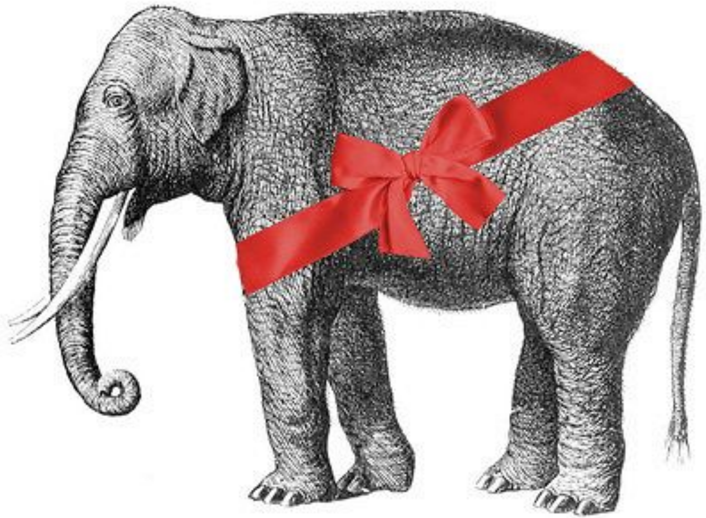
AGI tasks: github.com/RussianNLP/RussianSuperGLUE
RAM & speed: github.com/RussianNLP/MOROCCO
Probing: github.com/RussianNLP/rusenteval

@rybolos

SberDevices, HSE University, Huawei

Раздача слонов

API ruGPT-3



ruGPT-3 API
api.aicloud.sbercloud.ru/public_inference/docs

public inference api 0.0.1 OAS3

/public_inference/openapi.json

Public inference API

Servers

/public_inference

public_inference

POST `/gpt3/predict` Predict Gpt3

ruGPT-3 API
api.aicloud.sbercloud.ru/public_inference/docs

public inference api 0.0.1 OAS3

/public_inference/openapi.json

Public inference API

```
curl –location –request POST
'https://api.aicloud.sbercloud.ru/public_inference/gpt3/predict' \
–header 'Content-Type: application/json' \
–data-raw '{"text": "привет дорогой друг как твои дела"}'
```

Servers

/public_inference

public_inference

POST /gpt3/predict Predict Gpt3

ruGPT-3 API
api.aicloud.sbercloud.ru/public_inference/docs

Приходят два парфюмера в бар. Один спрашивает:
- Как ты догадался, что я левша?
- У тебя левая ноздря шире правой...

У вас сметана есть?- Нет.- А жирная?

- Фима, не называйте меня ""”моя красавица””. Это уже не в моде.
- А кто ж ты?
- Я как раз в моде.

Разговаривают два новых русских:
- Слышь, а что это у вас тут пахнет?
А ну да - налоговыми преступлениями.

Очень сложно понять, где заканчивается чёрное и начинается белое. Особенно ночью.