

TATIANA SHAVRINA
R&D TEAM, NLP



NLP: обзор инструментов для разработчиков

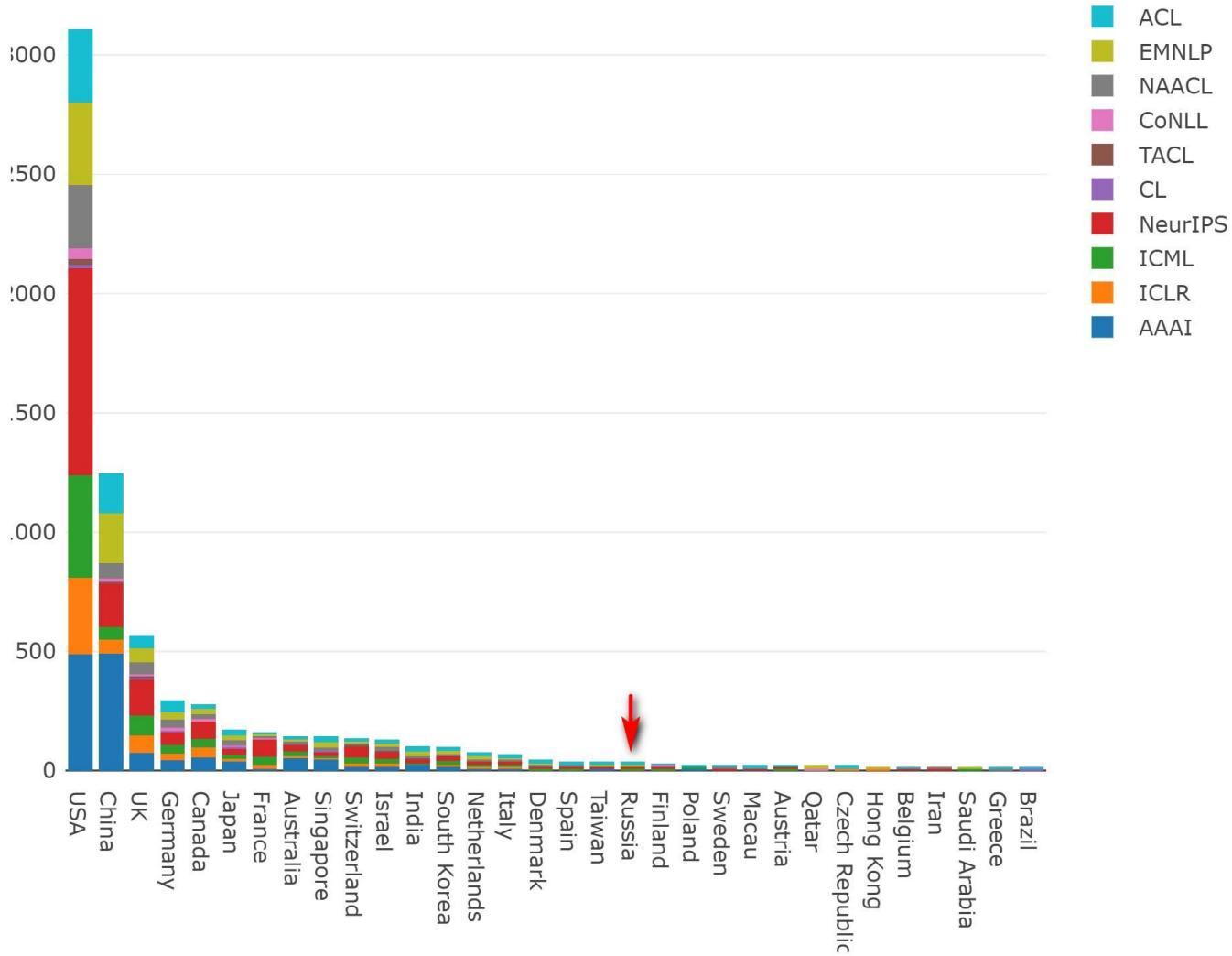
От сбора данных до универсальных
моделей, решающих ЕГЭ

NLP tools

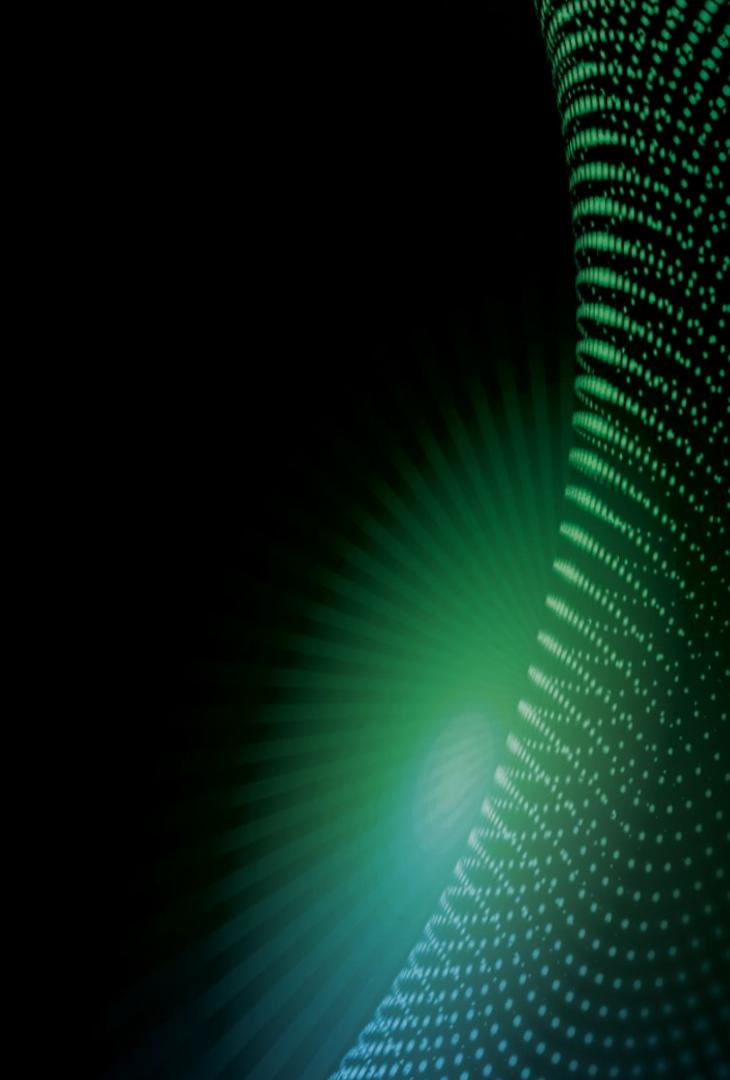
- Русский vs английский
- препроцессинг: SOTA pipeline, основные библиотеки
- универсальные модели и трансформеры
- сбор данных
- интерпретация и оценка моделей

Russian vs English





Preprocessing & Libraries



Pipelines

Standard: nltk, spacy, TextBlob, mystem, pymorphy, malt parser

SOTA:

Sentence tokenization - [/pypi.org/project/rusenttokenize/](https://pypi.org/project/rusenttokenize/)

Morphology - [/pypi.org/project/rnnmorph/](https://pypi.org/project/rnnmorph/)

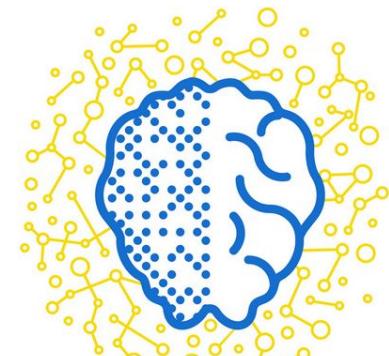
Syntax:

UDPipe - pypi.org/project/ufal.udpipe/

TurkuNLP Russian

DeepPavlov
morpho_ru_syntagrus_pymorphy
ru_syntagrus_joint_parsing

ссылка на демо в [CodaLab](#)



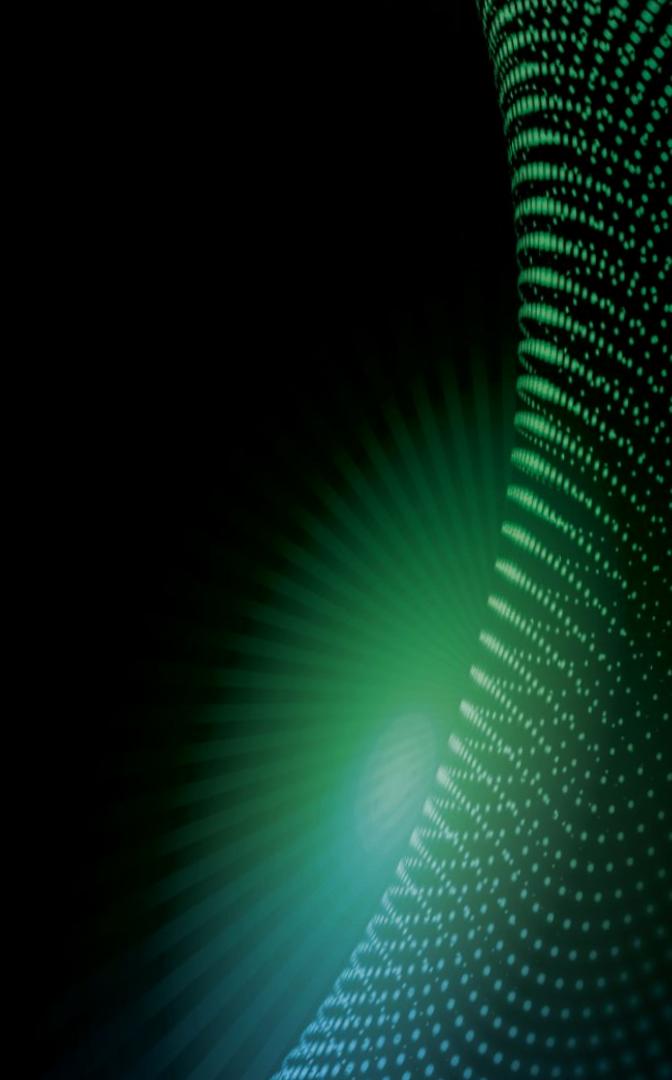


Deep
Pavlov

Не только чат-боты!

- Sentiment
- NER
- SQuAD
- BERT, ELMo
- классификаторы:
 - оскорблений
 - намерения
 - информативное/неинформационное сообщение
- диалоговые агенты
- чит-чат на общие темы
-

Universal Models & Transformers



Universal NLP Models

Как мы к этому шли?

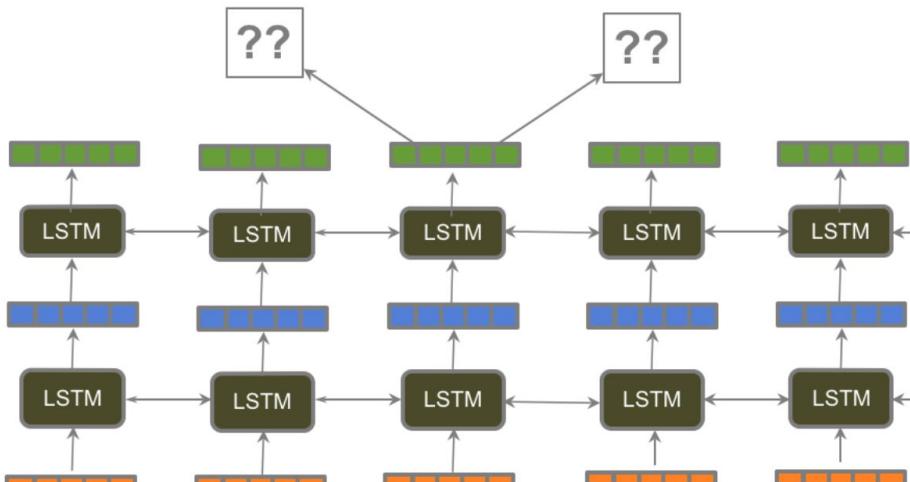
2013 - word2vec

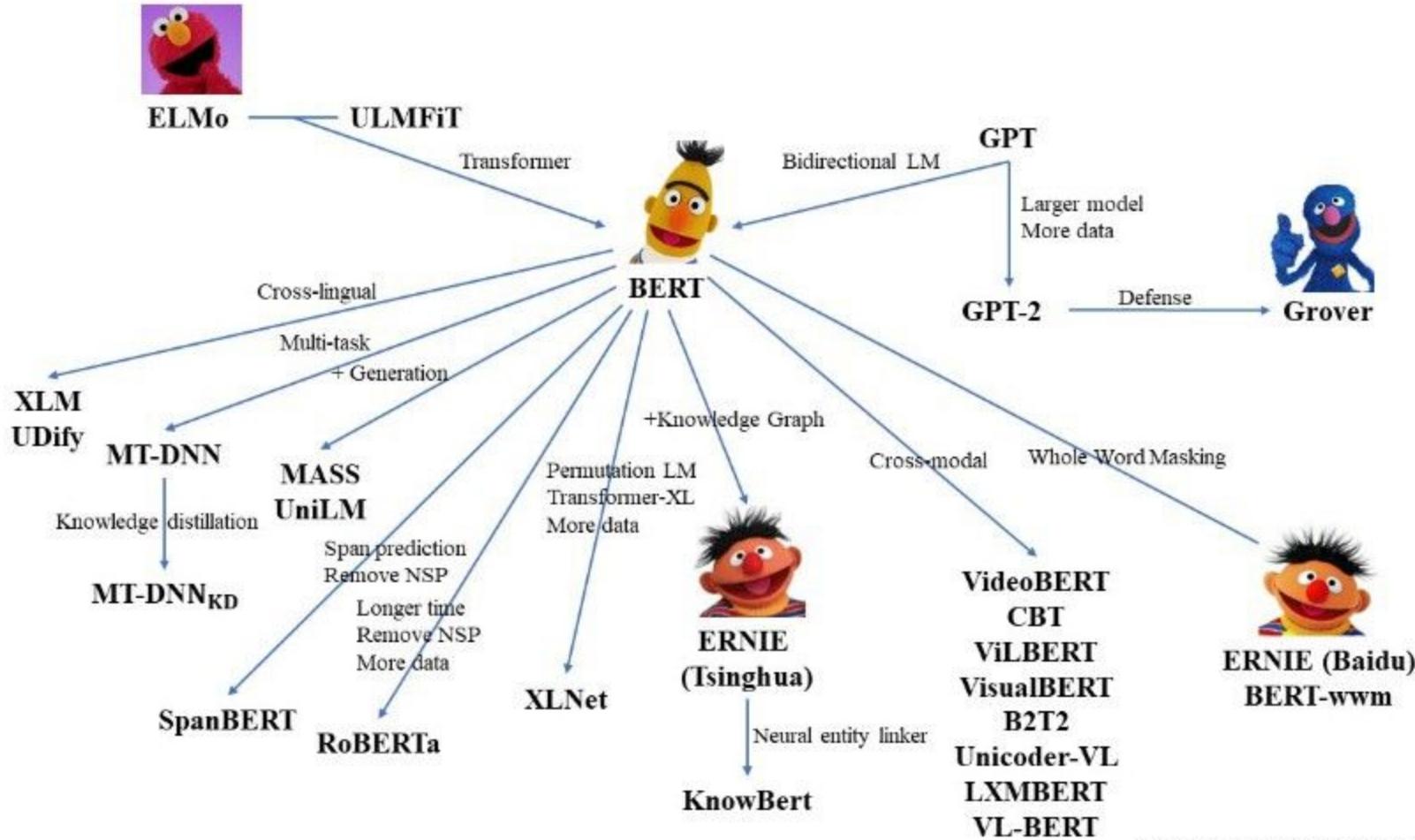
2016 - fastText

2017 - Attention

2018 - BERT, ULMFit, ElMo...

NLP's ImageNet moment has arrived





Universal NLP Models

Modern universal transformers have their pros and cons:

Auto-regressive models (ELMo, GPT)

- + Free of artificial noise in the data
- - No bidirectional context

Denoising auto-encoding (BERT)

- + Can make independent predictions
- - Trained on noisy data with {mask}
- + Have a natural context on the right and left

How to combine the advantages of both approaches?

- XLNet

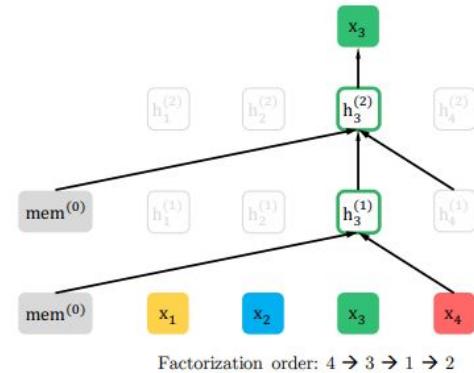
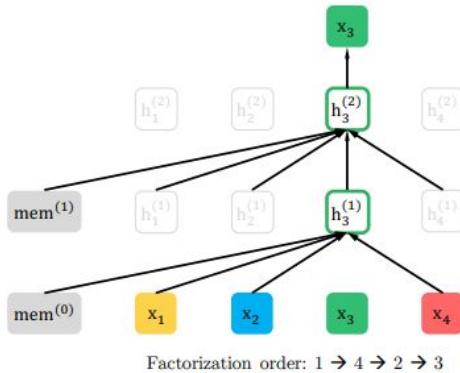
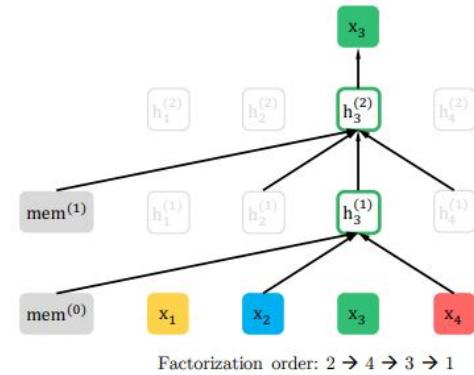
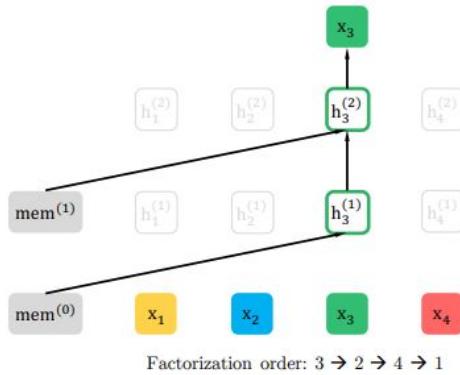
Universal NLP Models

2.2 Objective: Permutation Language Modeling

- XLNet

Permutation Language
Modeling (PLM),

combining autoregressive +
bidirectional approaches



DeepPavlov Pretrained Embeddings

Pre-trained embeddings

- BERT
- ELMo
- fastText
- AutoML

MODELS

- BERT-based models
- Context Question Answering

- Classification

- Morphological Tagger

- Named Entity Recognition

- Neural Ranking

- Slot filling

- Spelling Correction

- Syntactic parsing

- Model usage

- Joint model usage

- Model architecture

- Model quality

- TF-IDF Ranking

- Popularity Ranking

- Knowledge Base Question answering

SKILLS

- Goal-Oriented Dialogue Bot

- Open-Domain Question Answering

- Sequence-To-Sequence Dialogue Bot

- Frequently Asked Questions Answering

- AIML

- Rasa

- ...

[Docs](#) » Pre-trained embeddings

[!\[\]\(6b01ea4b684d505bfddfcab4dd0e2003_img.jpg\) Edit on GitHub](#)

Pre-trained embeddings

BERT

We are publishing several pre-trained BERT models:

- RuBERT for Russian language
- Slavic BERT for Bulgarian, Czech, Polish, and Russian
- Conversational BERT for informal English
- and Conversational BERT for informal Russian

Description of these models is available in the [BERT section](#) of the docs.

License

The pre-trained models are distributed under the [License Apache 2.0](#).

Downloads

The models can be run with the original [BERT repo](#) code. The download links are:

Description	Model parameters	Download link
RuBERT	vocab size = 120K, parameters = 180M, size = 632MB	[rubert_cased_L-12_H-768_A-12]
Slavic BERT	vocab size = 120K, parameters = 180M, size = 632MB	[bg_cs_pl_ru_cased_L-12_H-768_A-12]
Conversational BERT	vocab size = 30K, parameters = 110M, size = 385MB	[conversational_cased_L-12_H-768_A]
Conversational RuBERT	vocab size = 120K, parameters = 180M, size = 630MB	[conversational_cased_L-12_H-768_A]

ELMo

We are publishing [Russian language ELMo embeddings model](#) for tensorflow-hub and [LM model](#) for training and fine-tuning ELMo

Порфириевич - GPT-2

Универсальные модели способны **запускать развивающийся по внутренним законам самоорганизующийся интеллект, способный за пределами привычной логики следовать своим собственным.**



Порфириевич

написано с помощью нейронной сети
<https://porfirevich.ru>

https://github.com/mgrankin/ru_transformers

RusVectores Pretrained Embeddings

RusVectores Похожие слова Визуализации Калькулятор Различные операции Модели О проекте Контакты RU/EN

Модели

Все модели можно скачать и свободно использовать на условиях лицензии CC-BY (жирным выделены модели, доступные для использования в веб-интерфейсе).

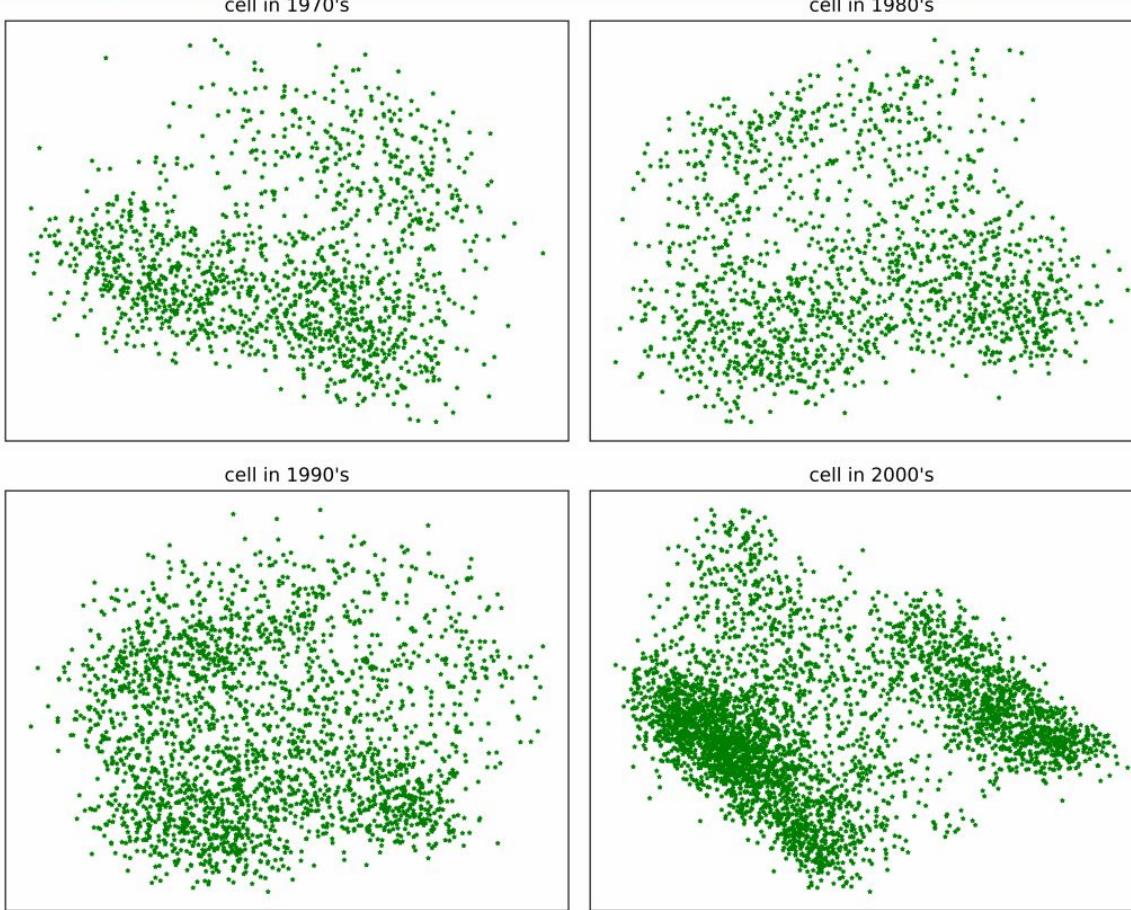
Контекстуализированные модели

Постоянный идентификатор ▾▼	Скачать ▾▼	Алгоритм ▾▼	Корпус ▾▼	Размер корпуса ▾▼	Таргет ▾▼	Размерность вектора ▾▼	RUSSE'18 ▾▼	ParaPhraser ▾▼	Дата создания ▾▼
tayga_lemmas_elmo_2048_2019	1.7 Гбайт	ELMo	Тайга (леммы)	почти 5 миллиардов слов	Нет	2048	0.93	0.54	Декабрь 2019
ruwikiruscorpora_tokens_elmo_1024_2019	197 Мбайт	ELMo	НКРЯ и Википедия за декабрь 2018 (токены)	989 миллионов слов	Нет	1024	0.88	0.55	Август 2019
ruwikiruscorpora_lemmas_elmo_1024_2019	197 Мбайт	ELMo	НКРЯ и Википедия за декабрь 2018 (леммы)	989 миллионов слов	Нет	1024	0.91	0.57	Август 2019

Статические модели

Таблицу можно (и нужно) пролистывать по горизонтали!

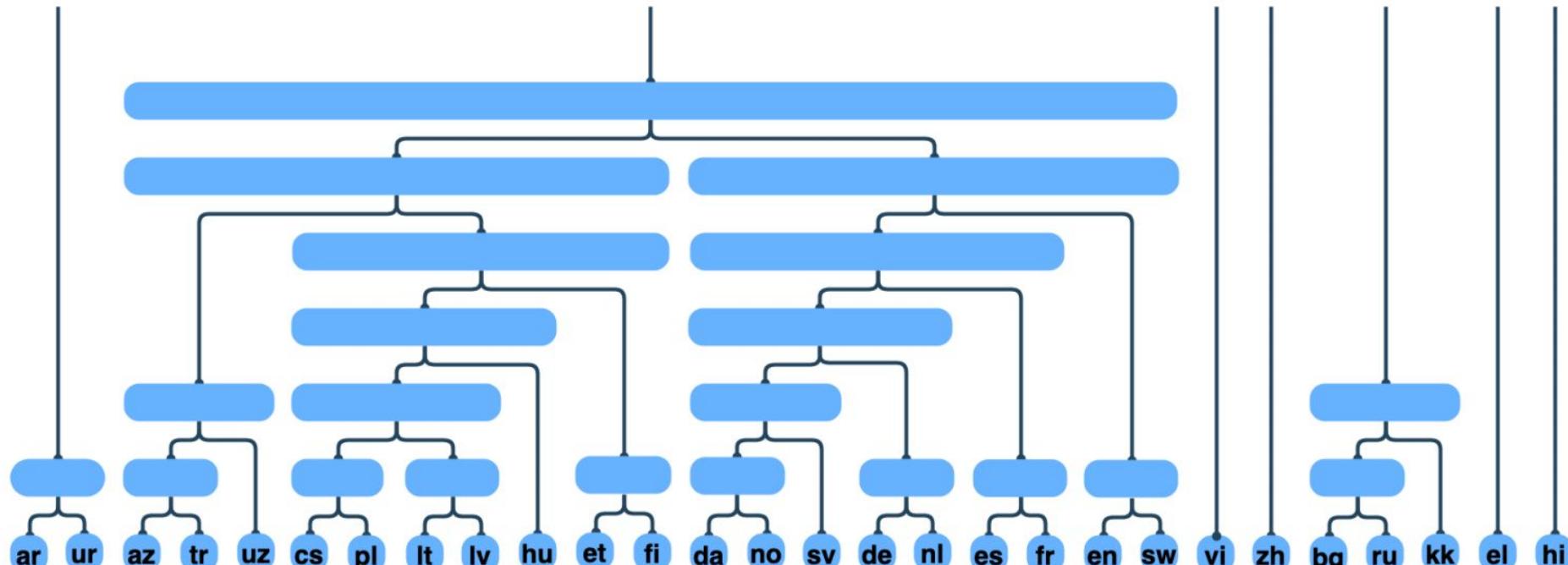
Постоянный идентификатор ▾▼	Скачать ▾▼	Корpus ▾▼	Размер корпуса ▾▼	Объём словаря ▾▼	Частотный порог ▾▼	Таргет ▾▼	Алгоритм ▾▼	Размерность вектора ▾▼
ruscorpora_upos_cbow_300_20_2019	462 Мбайт	НКРЯ	270 миллионов слов	189 193	5 (потолок словаря 250K)	Universal Tags	Continuous Bag-of-Words	300
ruwikiruscorpora_upos_skipgram_300_2_2019	608 Мбайт	НКРЯ и Википедия за декабрь 2018	788 миллионов слов	248 978	5 (потолок словаря 250K)	Universal Tags	Continuous Skipgram	300



ELMo representations of each occurrence of the word '*cell*' in 4 decades:
actual semantic shift. Diversity significantly increased in 2000s.

Jasdeep Singh (2019)

BERT is Not an Interlingua and the Bias of Tokenization



(a) Agglomerative clustering of languages based on subword overlap, generated from using the BERT tokenizer to tokenize the first 15 thousand examples from XNLI and our translated data.

Как мы решали ЕГЭ

AI Journey 2019: решить экзамен по русскому языку

26 заданий - с открытыми и закрытыми вариантами ответов (59% от оценки)
+ сочинение по тексту (41% от оценки)

Типы заданий:

- орфография
- логика
- семантика
- пунктуация
- орфоэпия (ударения)
- морфология и синтаксис
- составление / генерация текста



Как мы решали ЕГЭ

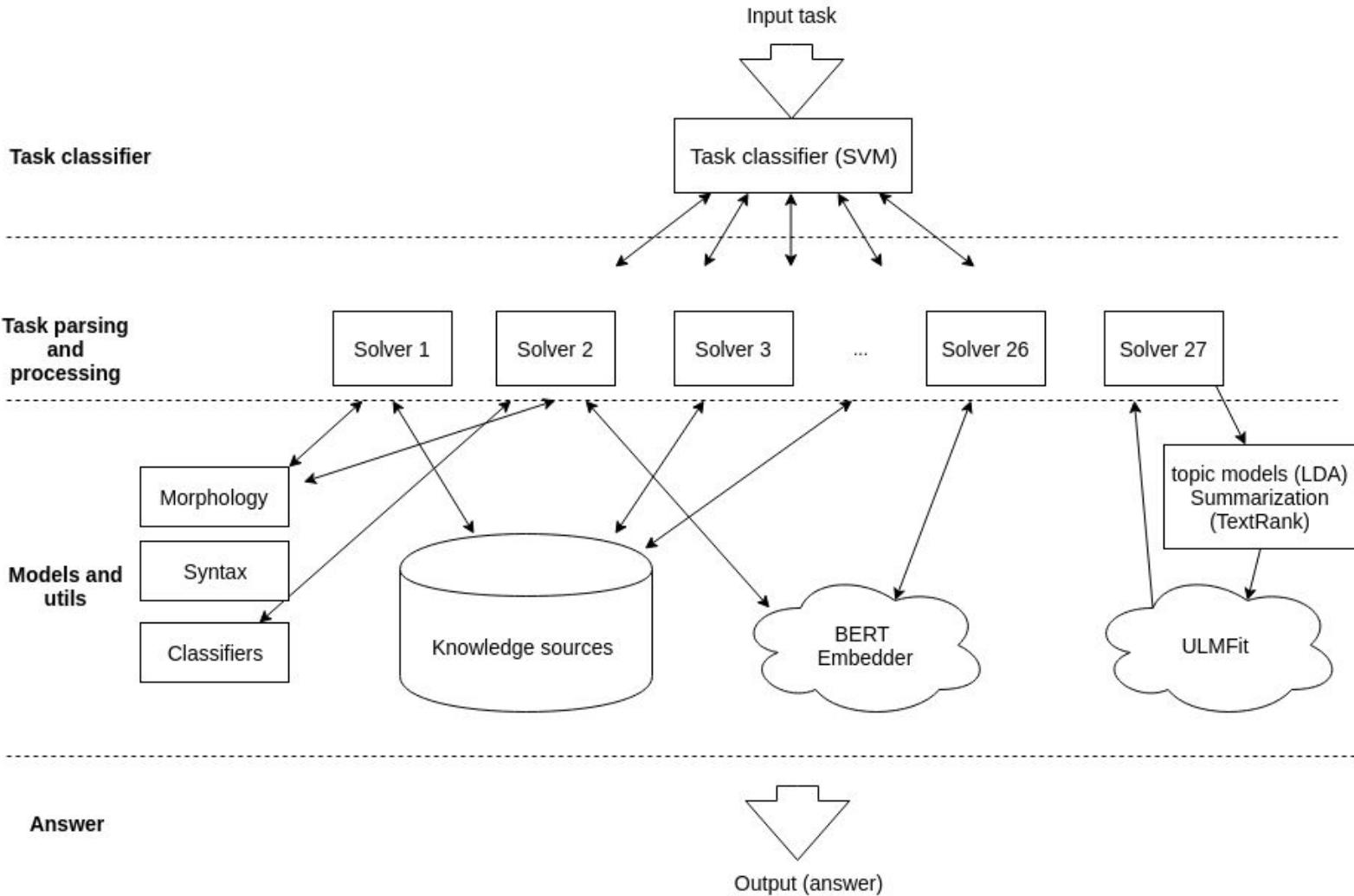
№ 9. Укажите варианты ответов, в которых во всех словах одного ряда пропущена безударная чередующаяся гласная корня. Запишите номера ответов.

- 1) зап..рать, р..стение, прил..гательное
- 2) сп..раль, заст..лить, к..мфорт
- 3) б..режок, ф..рмат, затв..рдеть
- 4) предв..рительный, прид..рожный, зам..чать
- 5) тв..рительный, з..рница, пл..вец

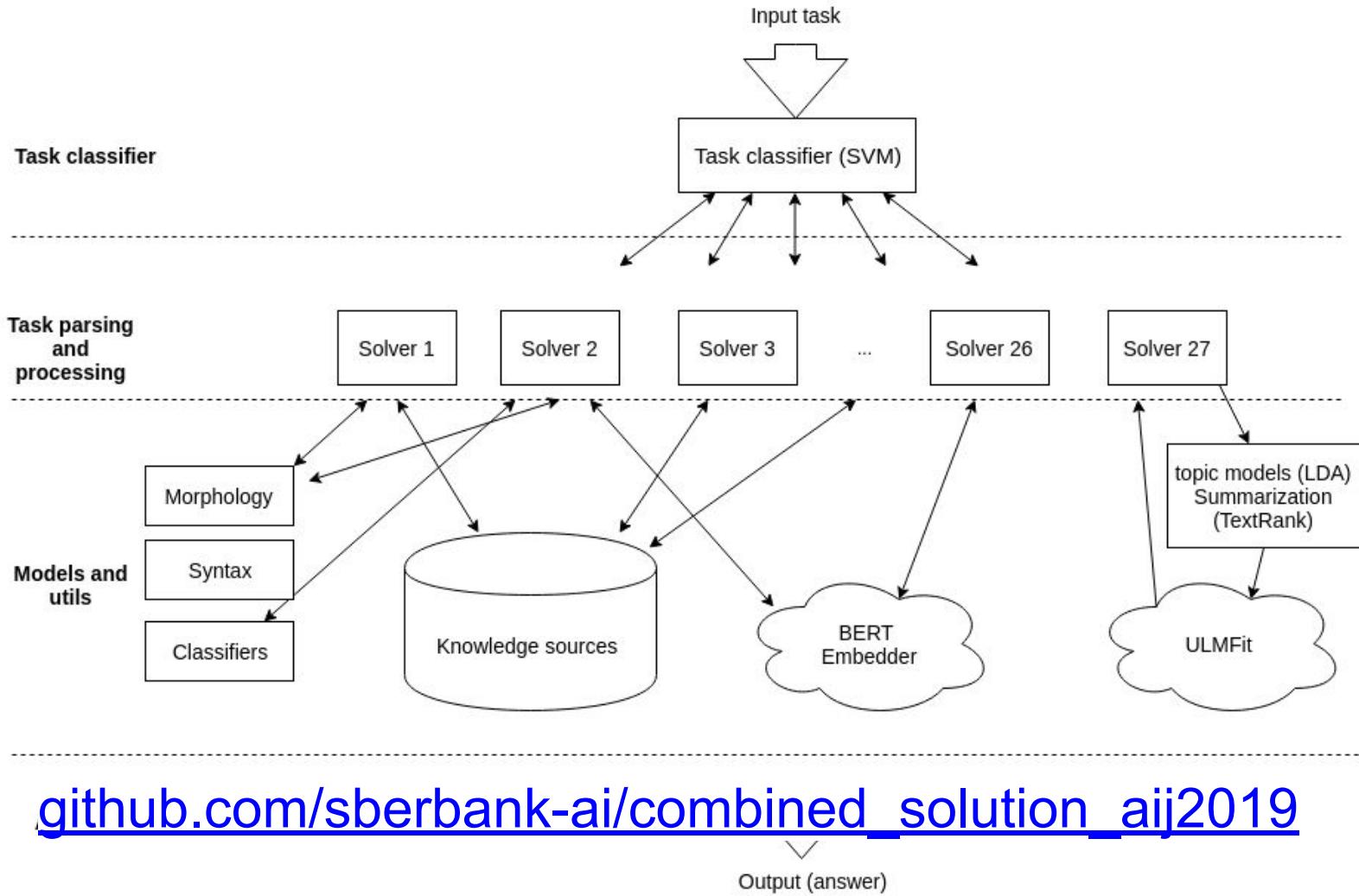
Ответ: 1, 5



BASELINE



BASELINE



contest.ai-journey.ru/ru/leaderboard

Результат решений							
#	Команда	Тест (max 34) <small>?</small>	Сочинение (max 24) <small>?</small>	Итого (max 100) <small>?</small>	Последнее решение	Попыток	
1	qbic	 	19,93	14,25	59,77	01 ноября 2019, 23:56	48
2	Bilbo Bagging	   	20,90	12,67	58,47	01 ноября 2019, 18:09	170
3	Magic City	  	14,50	16,33	55,63	01 ноября 2019, 23:47	115
4	borsden		15,77	13,33	53,40	29 октября 2019, 05:54	74
5	Ololosh AI		16,70	10,58	50,93	30 октября 2019, 21:06	150
6	nice	 	12,63	13,83	49,70	01 ноября 2019, 16:56	40
7	Niw		21,20	4,75	49,20	01 ноября 2019, 19:45	111
8	CDS_team	 	17,17	5,42	45,23	01 ноября 2019, 18:02	21
8	stickman		17,07	5,00	43,77	01 ноября 2019, 14:57	70
8	Orcs	 	13,00	8,92	43,77	01 ноября 2019, 23:58	19

Eventually
Russian AI ...
got 69 out of 100
for the exam!

Data Sources



Универсальные источники

Открытые данные соревнований, бенчмарков, статей:

- + обычно хорошо структурированные
- их слишком мало, тематические смещения

Kaggle

CodaLab

Dialogue Evaluation tracks

Большие данные:

- + volume, velocity, variety
- нужно структурировать и отбирать самому

Big Text Data

English:

- C4 - Colossal Cleaned Common Crawl www.tensorflow.org/datasets/catalog/c4
7 терабайт текстовых данных

Russian:

- Taiga Corpus - корпус для NLP-задач tatianashavrina.github.io/taiga_site
- Omnia Russica - 33 миллиарда слов из различных источников
omnia-russica.github.io

Taiga Corpus

An open-source corpus for machine learning.



Creative Commons

Wikipedia Based:

- wiki encyclopedia, wikihow, wikidata
- **wikireading** - github.com/google-research-datasets/wiki-reading
- **wikiconv** - github.com/conversationai/wikidetox/tree/master/wikiconv
- **wikidetox** - github.com/conversationai/wikidetox/
- **SberQuAD** - github.com/sberbank-ai/

CC resources:

- Open Subtitles
- Common Crawl

Interpretation & Evaluation

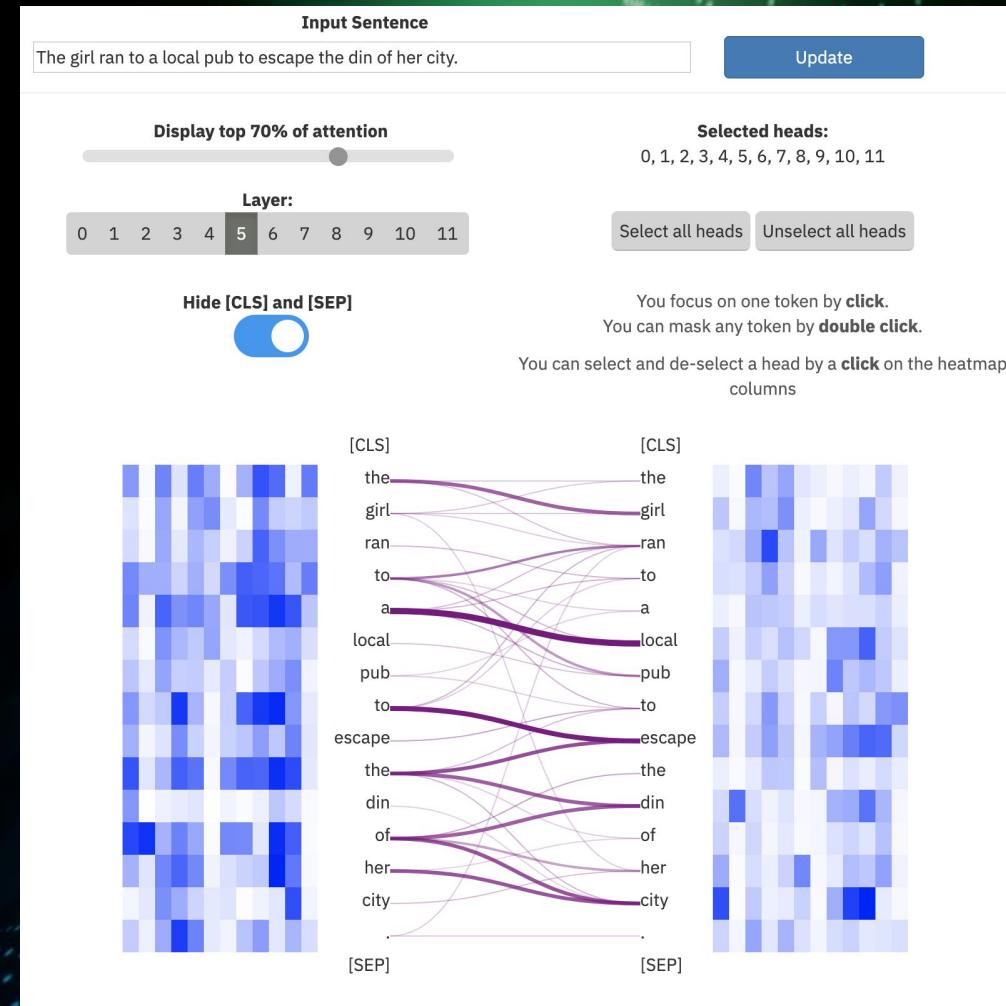


exBERT.net

[exBERT: A Visual Analysis Tool to Explain BERT's Learned Representations](#)

exBERT — это интерактивный инструмент, представленный компанией IBM, который позволяет пользователям исследовать что и как выучил трансформер в процессе создания языковой модели.

Работает с одним предложением!



Demos. AllenNLP

[AllenNLP Interpret: Explaining Predictions of NLP Models](#)

- 1) набор методов интерпретации, применимых к большинству моделей, доступных в библиотеке:
 - маскирование языковых моделей
 - анализ тональности
 - SQuAD
 - Textual Entailment
 - etc.
- 2) API для разработки новых методов интерпретации
- 3) многократно используемые веб-компоненты для визуализации результатов

Можно загружать свои модели!

[Annotate a sentence](#)[Semantic Role Labeling](#)[Named Entity Recognition](#)[Constituency Parsing](#)[Dependency Parsing](#)[Open Information Extraction](#)[Sentiment Analysis](#)[Annotate a passage](#)[Coreference Resolution](#)[Answer a question](#)[Reading Comprehension](#)[Semantic parsing](#)[WikiTableQuestions Semantic Parser](#)[Cornell NLP Semantic Parser](#)[Text to SQL \(ATIS\)](#)[QuaRel Zero](#)[Other](#)[Demo](#)[Usage](#)

Enter text or

Premise

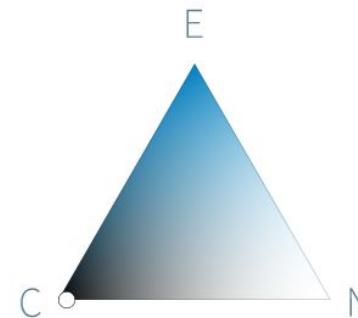
Two women are wandering along the shore drinking iced tea.

Hypothesis

Two women are sitting on a blanket near some rocks talking about politics.

Summary

It is **very likely** that the premise **contradicts** the hypothesis.

**Judgment**

Contradiction

Neutral

Probability

0%

97%

3%

Model Interpretations [What is this?](#)

SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

<https://super.gluebenchmark.com/>

Задачи SuperGLUE:

- стандартизировать оценку
- предоставить однозначную метрику для NLU моделей

SuperGLUE — средняя оценка по 8 NLU задачам:

- задачи более разнообразные (например, кореференция, QA и др.).
 - выбранные задачи тяжело оценивать, тяжело оценивать для машин, легко для человека
-
- выбраны задачи с небольшим набором данных для обучения



SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

[DOWNLOAD ALL DATA](#)

SuperGLUE Tasks

"premise": "The hamburger meat browned."

"question": "cause"

"choice1": "The cook froze it."

"choice2": "The cook grilled it."

"premise": "I decided to stay home for the night."

"question": "cause"

"choice1": "The forecast called for storms."

"choice2": "My friends urged me to go out."

Winogender Schema Diagnostics

AX-g



F1 / Accuracy

Gender Parity /
Accuracy

DOWNLOAD ALL DATA

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-mlt-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	IBM Research AI	BERT-mlt		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
-	6 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	38.0 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
		Outside Best		-	80.4	-	84.4	70.4/24.5	74.8/73.0	82.7	-	-	-	-
-	Stanford Hazy Research	Snorkel [SuperGLUE v1.9]		-	-	88.6/93.2	76.2	76.4/36.3	-	78.9	72.1	72.6	47.6	-

Click on a submission to see more information

Thank you for your attention!

Language Modeling

This demonstration uses the public 345M parameter [OpenAI GPT-2](#) language model to generate sentences.

Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:

Thank you for your

Predictions:

30.3% **support**

10.3% **continued**

5.4% **patience**

4.4% **interest**

3.2% **help**

← Undo