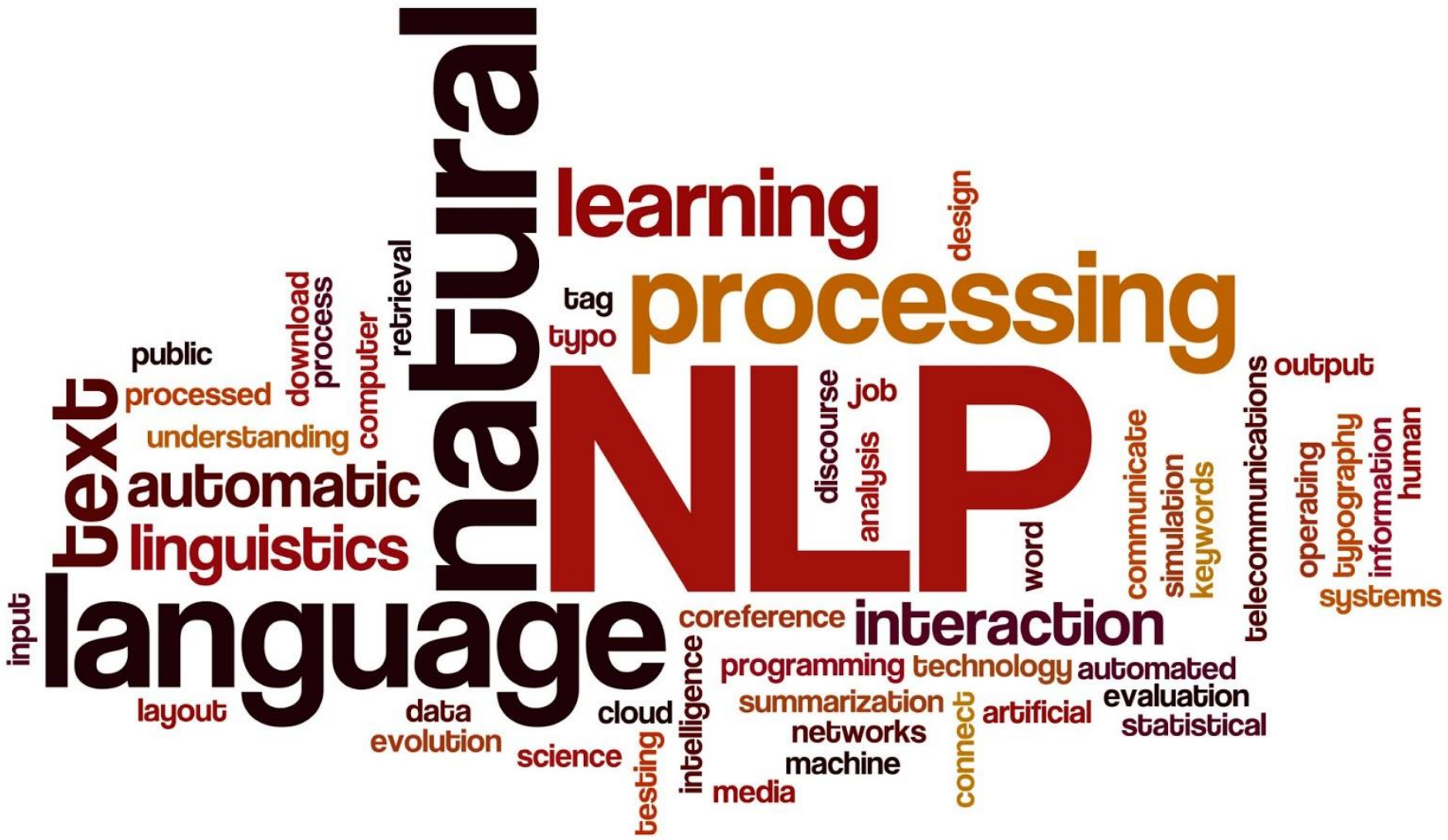
DaNetQA – ответ на вопрос наличия информации в тексте(бинарная классификация)

Выбор темы

Почему именно DaNetQA

Об NLP


natural
learning
processing
NLP
language
text
automatic
linguistics
interaction
public
processed
understanding
download
process
computer
retrieval
tag
typo
design
discourse
analysis
job
word
communicate
simulation
keywords
telecommunications
output
operating
typography
information
human
systems
input
layout
data
evolution
science
cloud
testing
intelligence
coreference
programming
technology
automated
evaluation
statistical
summarization
networks
machine
media
artificial
connect

Датасет

0	Вднх - это выставочный центр?	территории Останкинского района Северо-Восточного административного округа города Москвы. Переход на станцию ВДНХ Калужско-Рижской линии. Названа в честь Всероссийского выставочного центра — названия ВДНХ с 1992 по 2014 год. 20 ноября 2004 года линия монорельса начала работать в «экскурсионном режиме» и перевезла первых пассажиров .	True	0
1	Вднх - это выставочный центр?	Вы́ставка достиже́ний наро́дного хозя́йства , в 1959—1991 годах — Вы́ставка достиже́ний наро́дного хозя́йства СССР , в 1992—2014 годах — Всероссийский вы́ставочный центр) — выставочный комплекс в Останкинском районе Северо-Восточного административного округа города Москвы, второй по величине выставочный комплекс в городе. Входит в 50 крупнейших выставочных центров мира. Ежегодно ВДНХ посещают 30 млн гостей. 1 августа 2019 года выставка отпраздновала 80-летний юбилей. Территориально ВДНХ объединена с парком «Останкино» и Главным ботаническим садом , их общая площадь составляет почти 700 га: 240,2 га — площадь ВДНХ, 75,6 га — площадь парка «Останкино», 361 га — площадь ГБС, 9,5 га музейно-выставочный центр «Рабочий и колхозница» и площадь перед аркой Главного входа. На территории Выставки расположено множество шедевров архитектуры — 49 объектов ВДНХ признаны памятниками культурного наследия.	True	1
2	Был ли джиган в black star?	Вместе с этим треком они выступили на церемонии вручения наград MTV RMA — это был первый выход Джигана на большую сцену. В 2007 году Джиган стал официальным артистом лейбла Black Star Inc., выпустил первый песню и клип «Одноклассница» — совместная работа с Тимати. В марте 2011 года появилась совместная с Юлей Савичевой композиция «Отпусти». Песня поднялась в радиочарте «Tophit» до 8 места и заняла первое место чарта Weekly Audience Choice Top Hit . Количество просмотров клипа на YouTube превысило 17 миллионов.	True	2
3	Xiaomi конкурент apple?	Xiaomi — китайская компания, основанная в 2010 году и занимающаяся выпуском электроники и бытовой техники, а также разработкой программного обеспечения. Основной продукцией компании являются смартфоны, первый из которых был выпущен в 2011 году. В настоящее время Xiaomi производит широкий ассортимент смартфонов в разных ценовых сегментах и является одним из крупнейших производителей смартфонов в мире. Так, в третьем квартале 2014 года эта компания заняла наивысшее для себя третье место в мире по поставкам смартфонов, набрав 5,2 % в штучном выражении и уступив лишь Samsung и Apple . За весь 2017 год компания заняла по продажам первое место в Китае и второе место в Индии. В данном списке приводятся все смартфоны, когда-либо выпущенные компанией Xiaomi.	True	3
4	Был ли автомат калашникова в вов?	Отметив некоторые недостатки и в целом удачную конструкцию, специалисты ГАУ не рекомендовали принимать ПП Калашникова на вооружение по технологическим причинам. Заключение гласило: С 1942 года Калашников работал на Центральном научно-исследовательском полигоне стрелкового и миномётного вооружения ГАУ РККА. Здесь в 1944 году он создал опытный образец самозарядного карабина, который, хотя и не вышел в серийное производство, частично послужил прототипом для создания автомата. С 1945 года Михаил Калашников начал разработку автоматического оружия под промежуточный патрон 7,62×39 образца 1943 года. Автомат Калашникова победил в конкурсе 1947 года и был принят на вооружение.	False	4
...	...	...	...	...
1744	Разрешен ли такой вид ловли акул в настоящее время?	Для человека они потенциально полезны в медицине и применяются в качестве пищи. Исторически вылов акул производился в относительно небольших масштабах и не составлял проблем для восстановления их численности. Однако возросший с 80-х годов XX века промысел поставил под угрозу многие виды[46]. Одна из причин роста популярности акул в качестве объекта промысла — это их плавники. Суп из акульих плавников считается деликатесом, и плавник по стоимости выше акульего мяса. Это привело к негуманному способу охоты за плавниками, которые добывают, срезая их с живой рыбы, а саму акулу при этом выбрасывая обратно в море. В настоящее время	True	1745

True/False – 58/42

План решения задачи

	id	label	alpha	text
1600	0	1	a	выявлена ли корреляция данных кена юка с данными публикуемыми стерном и стюартом с момента представления концепции eva широким профессиональным и научным кругам большое количество исследований было направлено на выявления значимых корректировок eva в 1997 году кен юк провёл масштабное исследование сравнив собственные расчётные данные eva 1000 крупнейших компаний с данными публикуемыми стерном и стюартом несмотря на небольшое количество корректировок в разработанной им методике результаты расчёта eva оказались сильно коррелированными с расчётами разработчиков модели кроме того множество других авторов предложило свои методики корректировки бухгалтерских данных большинство из которых основано на корректировке капитальных эквивалентов
1601	1	0	a	занимало ли понятие цивилизация центральное место в сочинениях вико вольтера и гердера в xix веке европейские историки получив первые сведения о восточных обществах пришли к выводу что между обществами находящимися на стадии цивилизационного развития могут существовать качественные различия что позволило им говорить не об одной цивилизации а о нескольких цивилизациях впрочем представления о культурных различиях между европейской и неевропейскими культурами появились ещё раньше к примеру российский исследователь и н ионов трактует заявления итальянского философа джамбаттисты вико 1668—1744 о том что император китайский в высшей степени культурен как зародыш представлений о существовании особой китайской цивилизации а значит и о вероятной множественности цивилизаций тем не менее ни в его работах ни в сочинениях вольтера и иоганна готфрида гердера выражавших идеи родственные идеям вико понятие цивилизация не было главенствующим а понятие локальная цивилизация не употреблялось вовсе1
1602	2	0	a	признают ли в российской федерации чуп в качестве самостоятельной организации в большинстве стран снг существуют также частные унитарные предприятия чуп не наделённые правом собственности на закреплённое за ним имущество имущество является неделимым и не может быть распределено по вкладам паям долям акциям и находится в общей совместной собственности его членов физических лиц одного физического лица или одного юридического лица к таковым относят крестьянские фермерские хозяйства индивидуальные семейные и дочерние предприятия в российской федерации в качестве самостоятельных организаций за исключением дочерних предприятий таковые не признаются а руководители таких организаций являются индивидуальными предпринимателями что создаёт имущественные и организационные трудности у индивидуального предпринимателя фактически предприятия так например отсутствует право частной собственности на предприятие как на имущественный комплекс поскольку предприятие предполагает дополнительные хозяйственные отношения чего нет при индивидуальном предпринимательстве нет чёткого регламента положения членов в предприятии распределения прибыли и ответственности между ними и многих других аспектов
1603	3	1	a	питается ли медведями амурский тигр амурские тигры и бурые медведи представляют довольно серьёзную опасность друг для друга существует много достоверных свидетельств агрессивного взаимодействия между ними бурые и гималайские медведи составляют 5—8 пищевого рациона амурского тигра тем самым занимая в нем третье место есть многочисленные сообщения о том что тигры убивают медвежат и даже нападают на взрослых медведей в основном этим промышляют взрослые самцы медведи в свою очередь способны иногда отнимать добычу у тигров и иногда даже нападать на тигриц и молодых самцов в голодное время малайские медведи губачи и обитающие на юге гималайские медведи являясь весьма агрессивными животными временами отгоняют тигров от добычи хотя чаще случается обратное коегде тигры также как и на севере целеустремлённо охотятся на этих медведей6
1604	4	1	a	является ли обратимой реакция нуклеофильного замещения для реакций нуклеофильного замещения у sp2гибридного ацильного атома углерода реализируется двухстадийный механизм присоединенияотщепления в первой стадии нуклеофильный агент присоединяется к карбоновой кислоте или её производному с образованием заряженного для анионного нуклеофильного агента или незаряженного для нейтрального тетраэдрического интермедианта во второй стадии от этого интермедианта отщепляется в виде аниона или нейтральной молекулы уходящая группа z и образуется конечный продукт присоединения реакция обратима однако если z и nu сильно различаются по своей основности и нуклеофильности она становится необратимой

Актуальность проблемы

model	F-1	Accuracy
Neural networks [11]	79.82	76.65
Classifier + linguistic features [11]	81.10	77.39
Machine Translation + Semantic similarity [6]	78.51	81.41
BERT multilingual	85.48 ± 0.19	81.66 ± 0.38
RuBERT	$\mathbf{87.73} \pm 0.26$	$\mathbf{84.99} \pm 0.35$

We leverage two BERT-derived models as baseline. Multilingual BERT (MultiBERT), released by (Devlin et al., 2019), is a single language model pre-trained from monolingual corpora in 104 languages, Russian texts being a part of training data. MultiBERT uses a shared vocabulary for all languages. The capabilities of MultiBERT for zeroshot cross-lingual tasks have been recently studied by (Pires et al., 2019). Russian BERT (RuBERT) was trained on large-scale corpus of news and Wikipedia in Russian. To alleviate the training all weights except sub-word embeddings were borrowed from MultiBERT. The sub-word vocabulary was obtained from the same training corpus and the new mono-lingual embeddings were transformed from the multi-lingual ones. This allowed to incorporate longer Russian sub-word units into the vocabulary. This model is part of DeepPavlov framework (Kuratov and Arkhipov, 2019).

-Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language Yuri Kuratov , Mikhail Arkhipov , Neural Networks and Deep Learning Lab, Moscow Institute of Physics and Technology, May 2019

-SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis Pavel Efimov1∗ , Andrey Chertok2 , Leonid Boytsov, Pavel Braslavski3,4 1Saint Petersburg State University, Saint Petersburg, Russia 2Sberbank, Moscow, Russia 3Ural Federal University, Yekaterinburg, Russia 4 JetBrains Research, Saint Petersburg, Russia, May 2020

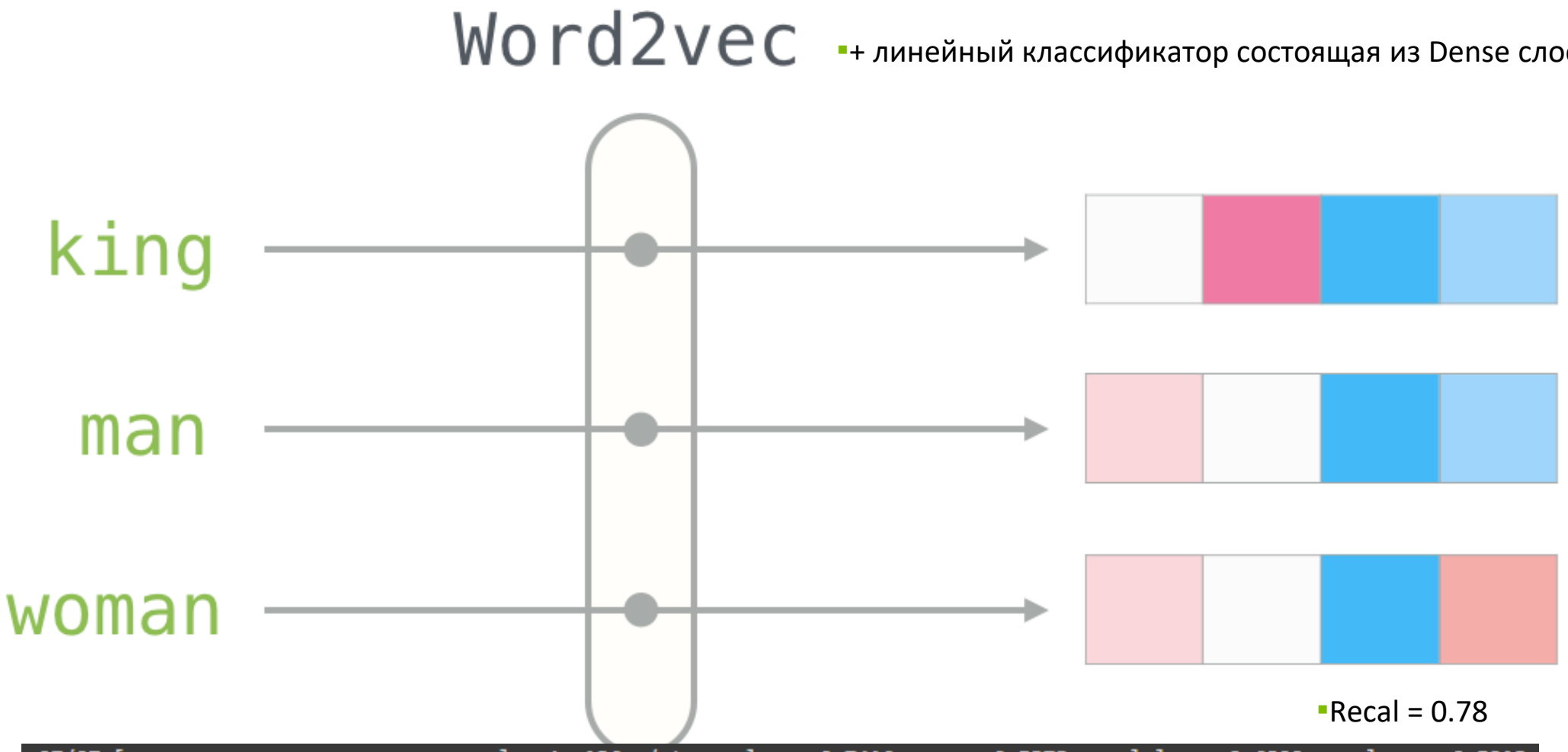
Цель

Dataset	Metrics	RuBERT	MultiBERT	TF-IDF	Human
LiDiRus	MCC	0.186	0.157	0.059	**0.626**
RCB	F_1/Acc.	0.432/0.468	0.383/0.429	0.45	**0.68/0.702**
PARus	Acc	0.61	0.588	0.48	**0.982**
MuSeRC	F_1/EM	0.656/0.256	0.626/0.253	0.589/0.244	**0.806/0.42**
TERRa	Acc	0.639	0.62	0.47	**0.92**
RUSSE	Acc	**0.894**	0.84	0.66	0.747
RWSD	Acc	0.675	0.675	0.66	**0.84**
DaNetQA	Acc	0.749	0.79	0.68	**0.879**
RuCoS	F_1/EM	0.255/0.251	0.371/0.367	0.256/0.251	**0.93/0.924**
Average		0.546	0.542	0.461	0.802

Table 2: Results of the human benchmark and the baseline models. MCC stands for Matthews Correlation Coefficient; Acc - Accuracy; EM - Exact Match.

- RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark Tatiana Shavrina1,2, Alena Fenogenova1 , Anton Emelyanov1,3, Denis Shevelev1 , Ekaterina Artemova2 , Valentin Malykh4 , Vladislav Mikhailov1,2, Maria Tikhonova1,2, Andrey Chertok1 and Andrey Evlampiev1 1Sberbank / Moscow, Russia 2National Research University Higher School of Economics / Moscow, Russia 3Moscow Institute of Physics and Technology / Moscow, Russia – Nov, 2021

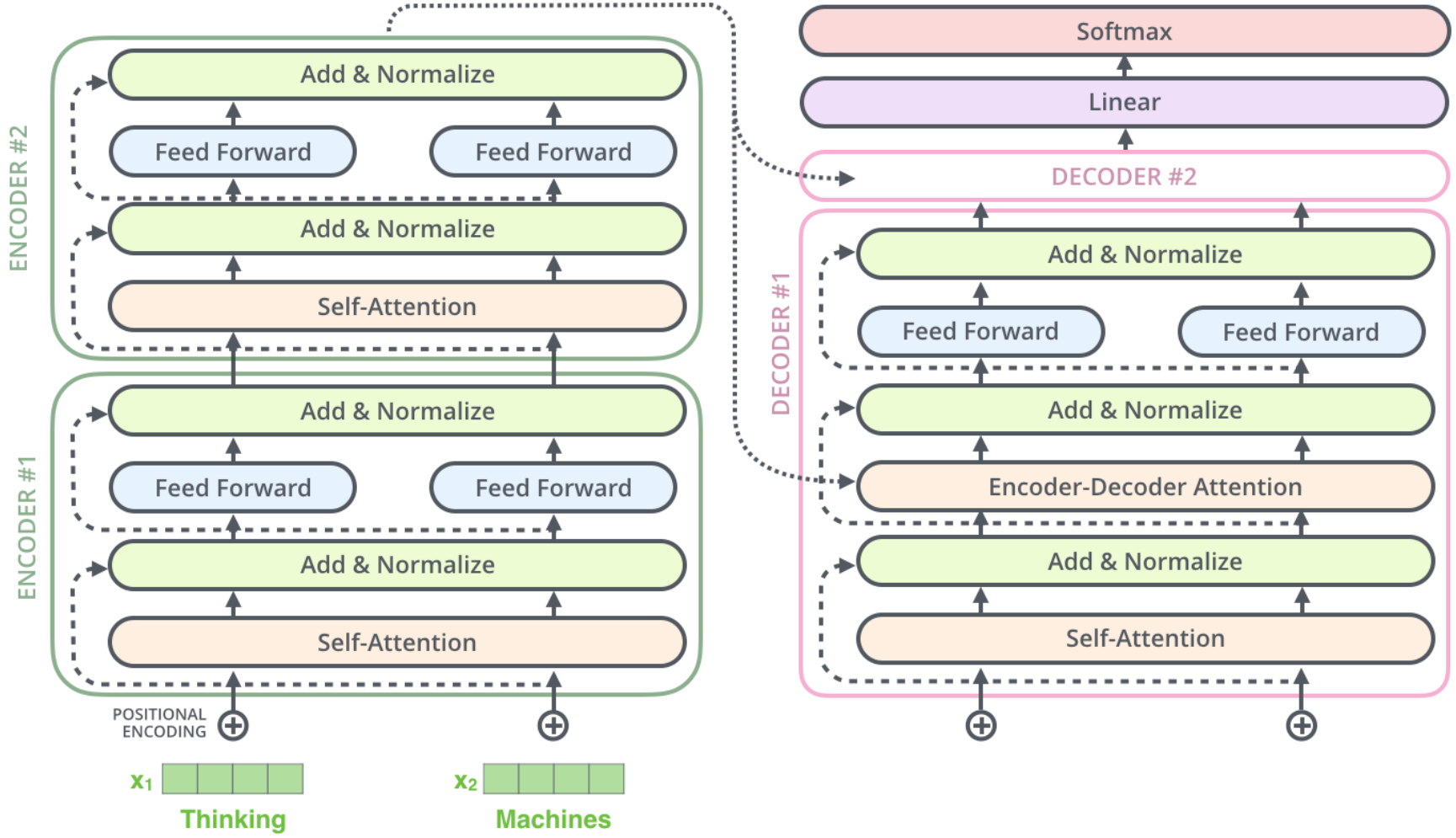
Лучшие существующие решения



- + линейный классификатор состоящая из Dense слоёв
- Recal = 0.78

```
27/27 [==============================] - 4s 136ms/step - loss: 0.7416 - acc: 0.5378 - val_loss: 0.6802 - val_acc: 0.5913
```

BERT


Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

Источники

Transfer learning from Hugging

Rubert large model whole word masking finetuned on SQuAD

Pretrained model using a masked language modeling (MLM) objective. Fine tuned on Russian QA datasets

Used QA Datasets

SQuAD + SberQuAD

SberQuAD original paper is here! Recommend to read!

```
tokenizer = AutoTokenizer.from_pretrained("cointegrated/rubert-tiny")

model = AutoModelForSequenceClassification.from_pretrained("cointegrated/rubert-tiny")
model.to(device)
```

Модель

cointegrated / **rubert-tiny** like 6

Fill-Mask | PyTorch | Transformers | ru | en | mit | bert | pretraining | russian | embeddings | masked-lm | tiny | AutoNLP Compatible | Infinity Compatible

Model card | Files and versions | Train | Deploy | Use in Transformers

This is a very small distilled version of the bert-base-multilingual-cased model for Russian and English (45 MB, 12M parameters).

This model is useful if you want to fine-tune it for a relatively simple Russian task (e.g. NER or sentiment classification), and you care more about speed and size than about accuracy. It is approximately x10 smaller and faster than a base-sized BERT. Its [CLS] embeddings can be used as a sentence representation aligned between Russian and English.

It was trained on the Yandex Translate corpus, OPUS-100 and Tatoeba, using MLM loss (distilled from bert-base-multilingual-cased), translation ranking loss, and [CLS] embeddings distilled from LaBSE, rubert-base-cased-sentence, Laser and USE.

There is a more detailed description in Russian.

NEW Select AutoNLP in the "Train" menu to fine-tune this model automatically.

Downloads last month
10,968

Hosted inference API

Fill-Mask | Mask token: [MASK] | Examples

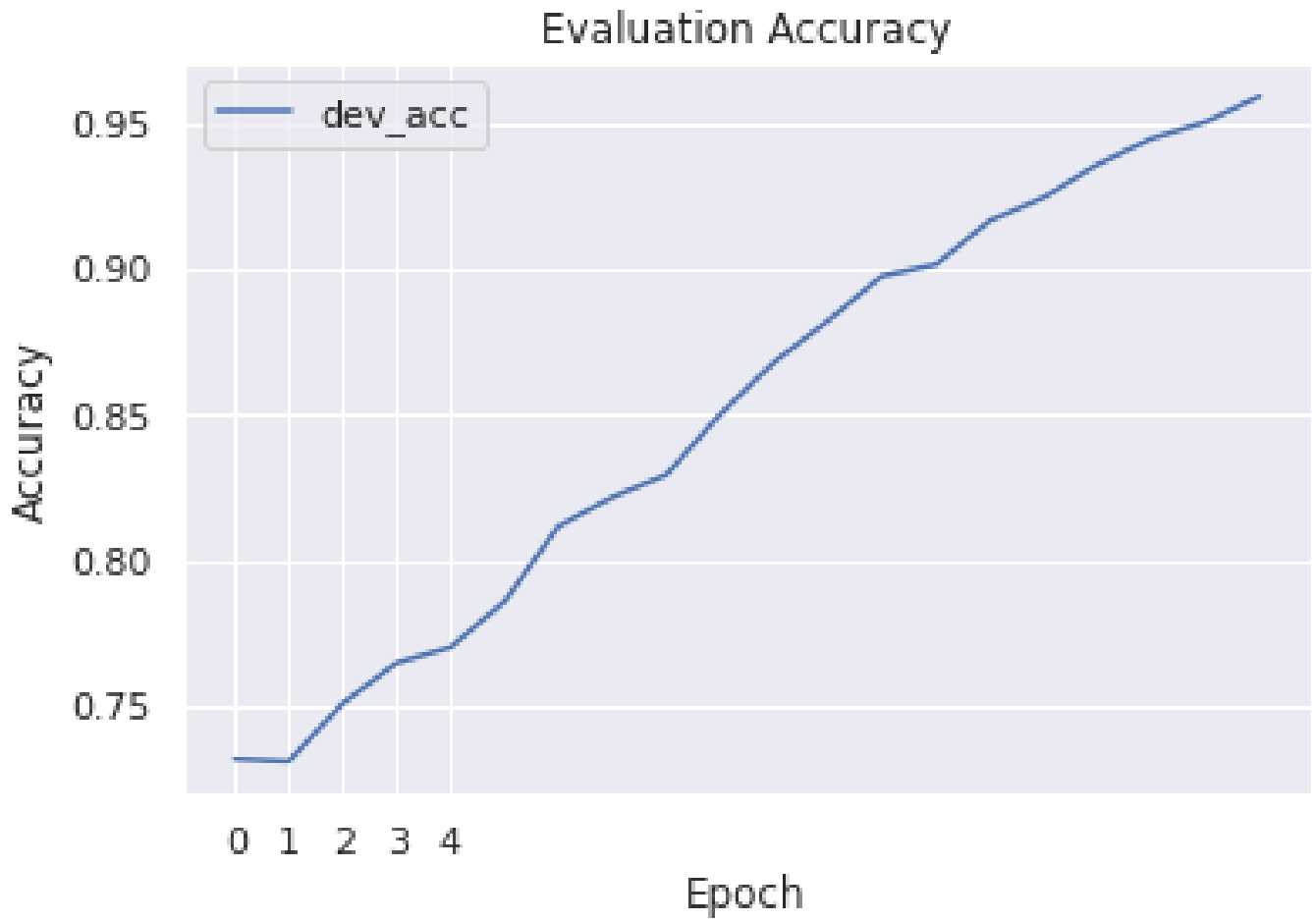
Миниатюрная модель для [MASK] разных задач.

Compute

This model can be loaded on the Inference API on-demand.

JSON Output | Maximize

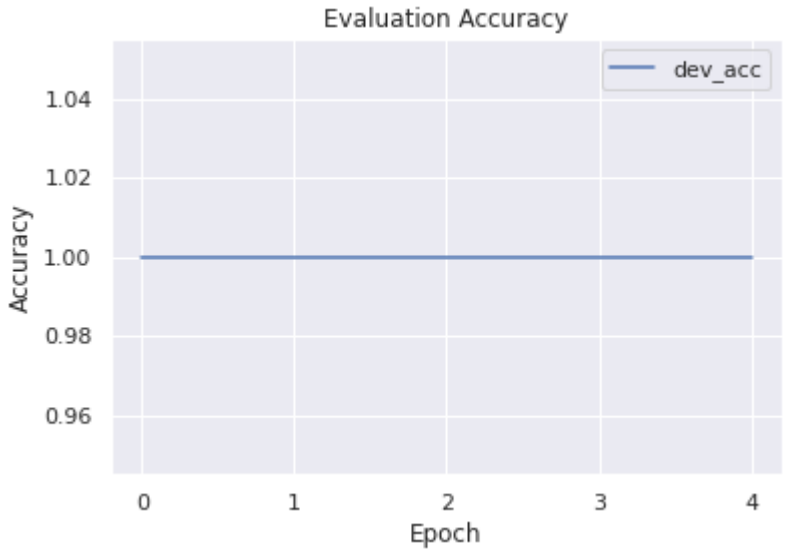
Результаты Transfer learning

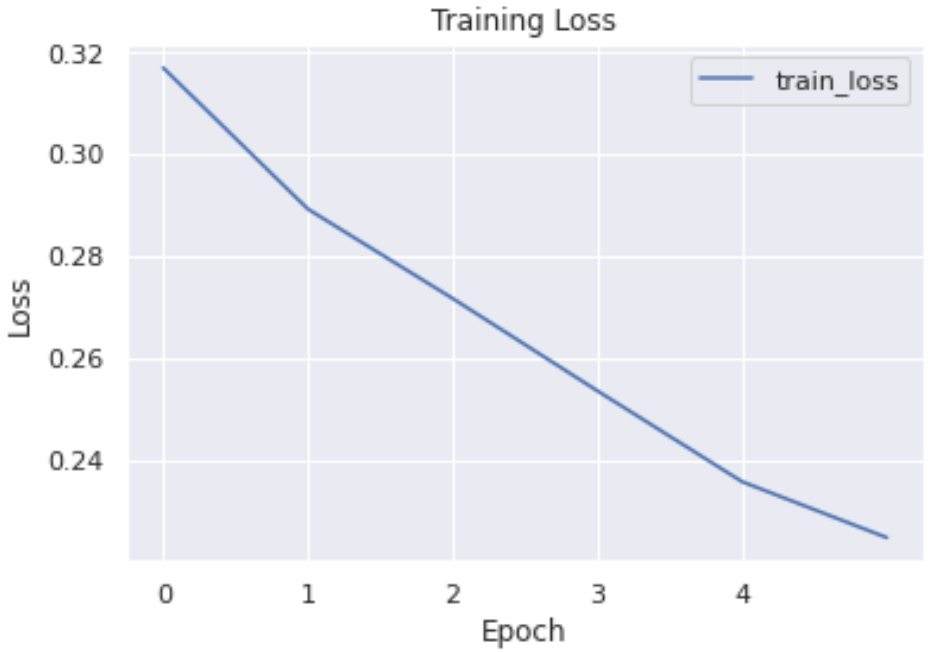


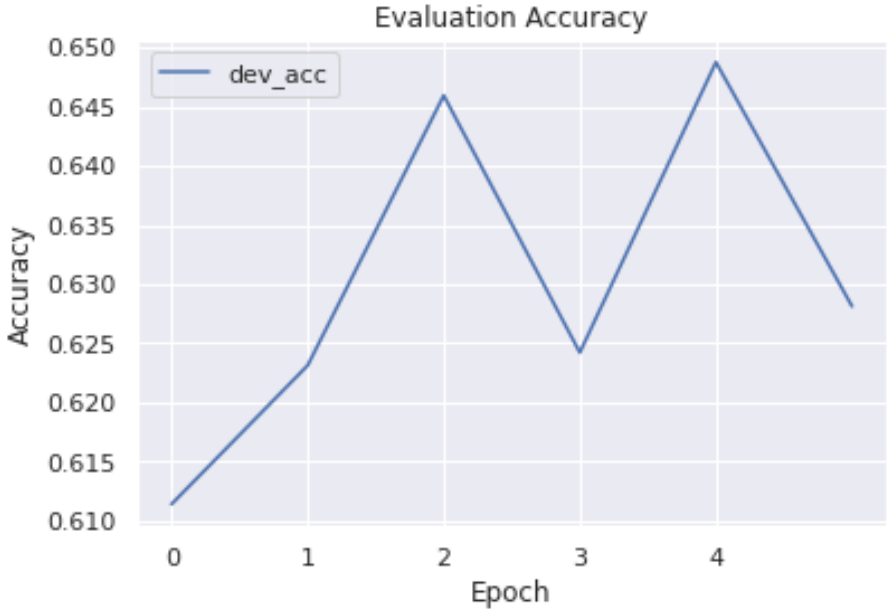
3 часа

Что не успел сделать?

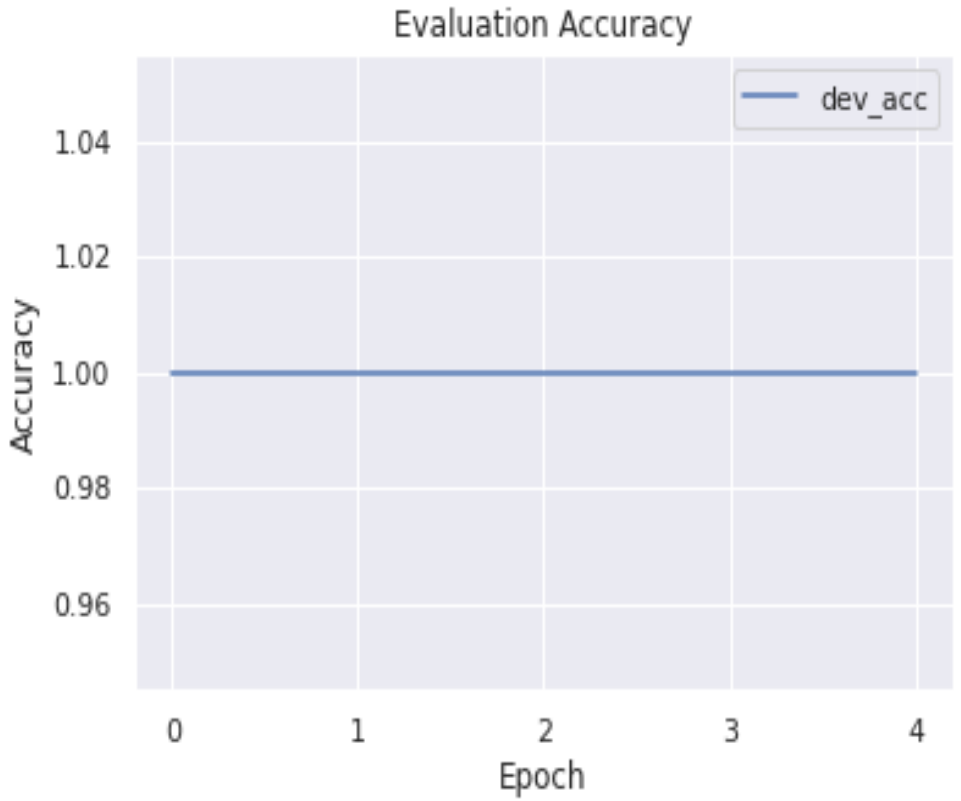
1. Заменить буквы с ударением
2. Провести анализ ошибок(визуализацию T-sne)
3. Модель обученная на большем кол-ве данных
4. Multilanguage модель

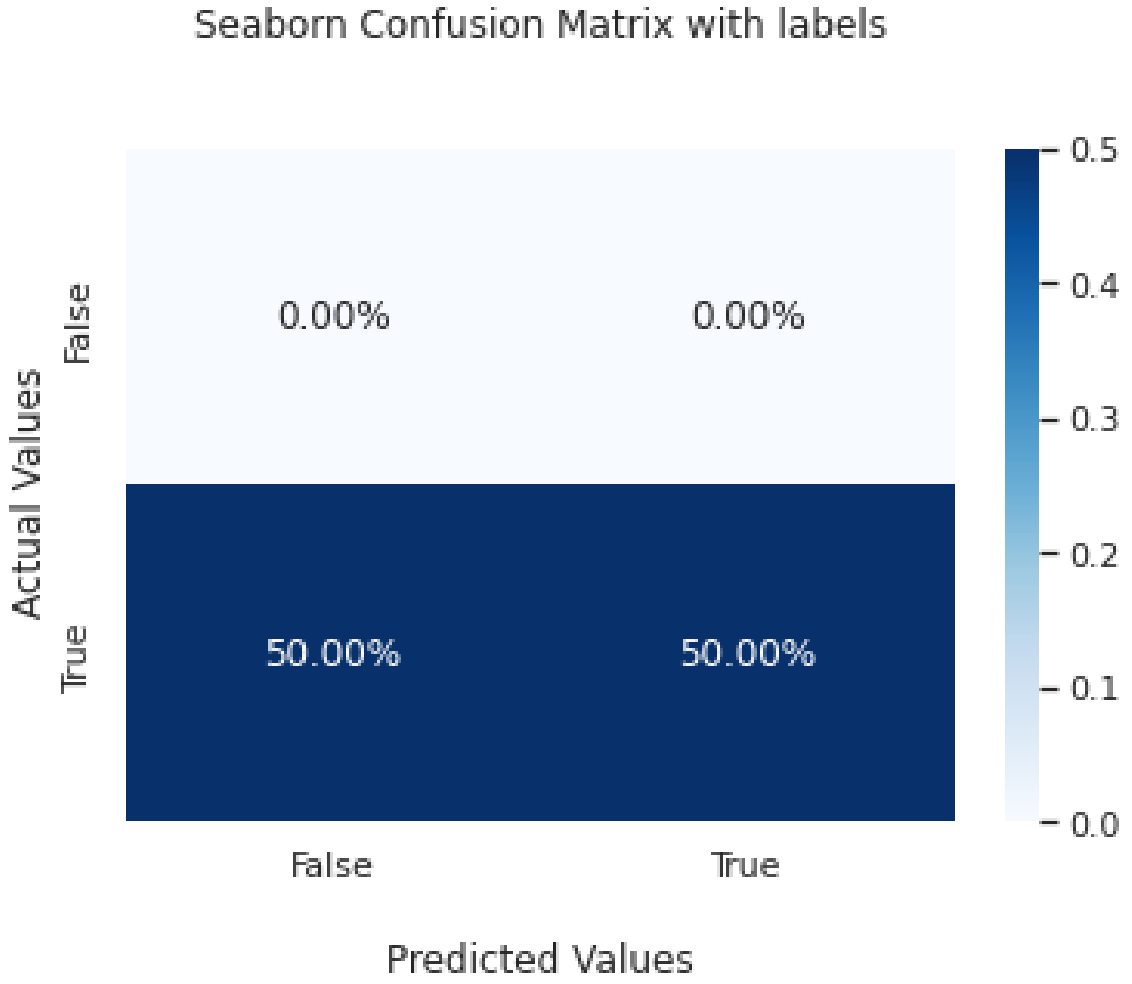






Странности


Evaluation Accuracy
dev_acc
1.04
1.02
1.00
0.98
0.96
Accuracy
0
1
2
3
4
Epoch


Seaborn Confusion Matrix with labels
Actual Values
False
True
0.00%
0.00%
50.00%
50.00%
False
True
Predicted Values
0.5
0.4
0.3
0.2
0.1
0.0

Что нужно сделать в дальнейшем?

1. Заменить буквы с ударением
2. Провести анализ ошибок(визуализацию T-sne)
3. Модель обученная на большем кол-ве данных
4. Multilanguage модель
5. Изменение гиперпараметров модели
6. Вопрос, как отдельный набор эмбеддингов

Спасибо за внимание!