Skoltech
Skolkovo Institute of Science and Technology

Methods for Detoxification of Texts for the Russian Language

Daryna Dementieva ‡ , Daniil Moskovskiy ‡ , Varvara Logacheva ‡ , David Dale ‡,
Olga Kozlova † , Nikita Semenov † , and Alexander Panchenko ‡

‡ Skolkovo Institute of Science and Technology, Moscow, Russia
† Mobile TeleSystems (MTS), Moscow, Russia

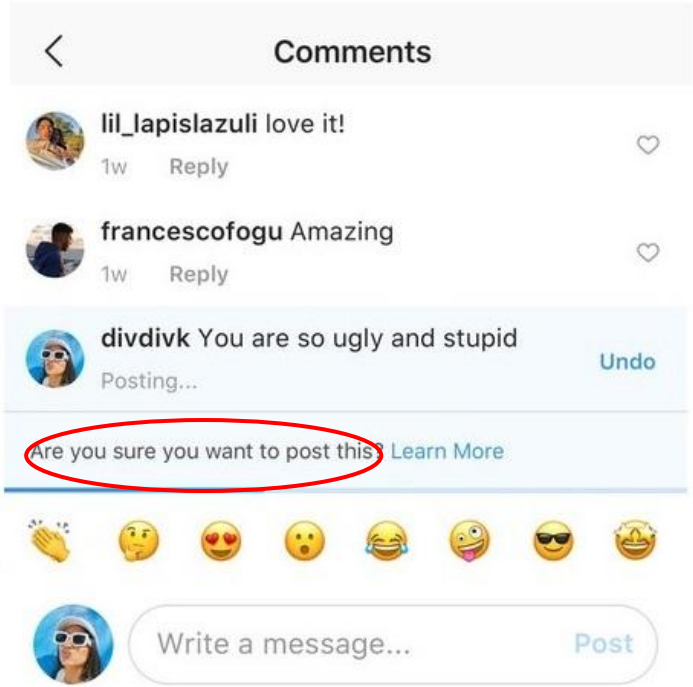
June, 2021

Motivation

Skoltech
Skolkovo Institute of Science and Technology

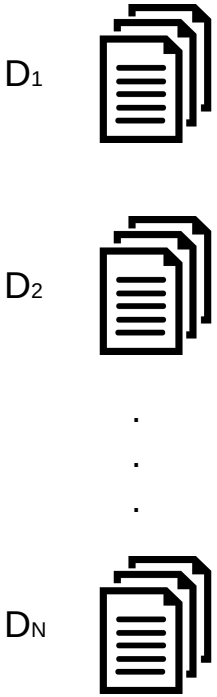
Chatbots: rewrite toxic answers before showing them to user

Social media: suggestions to users



Images: https://i.imgur.com/iVof3D4.jpg, https://www.theguardian.com/technology/2019/jul/09/instagram-bullying-new-feature-do-you-want-to-post-this

Task Formulation



Find: $f_\Theta : X \rightarrow Y$,
where x in style s_x and y in style s_y.
Measure of style: classifier $g(x) \rightarrow s_i$

Task Formulation

Skoltech
Skolkovo Institute of Science and Technology

Rewrite the text to:

- save the text content
- eliminate toxicity

These niggers are beggars → These black people are poor

You are a fucking idiot if you do this → I don't think your solution is well thought.

Stupid peace of shit stop deleting my stuff asshole → Stop deleting my stuff.

Previous Work

Skoltech
Skolkovo Institute of Science and Technology

Style transfer examples

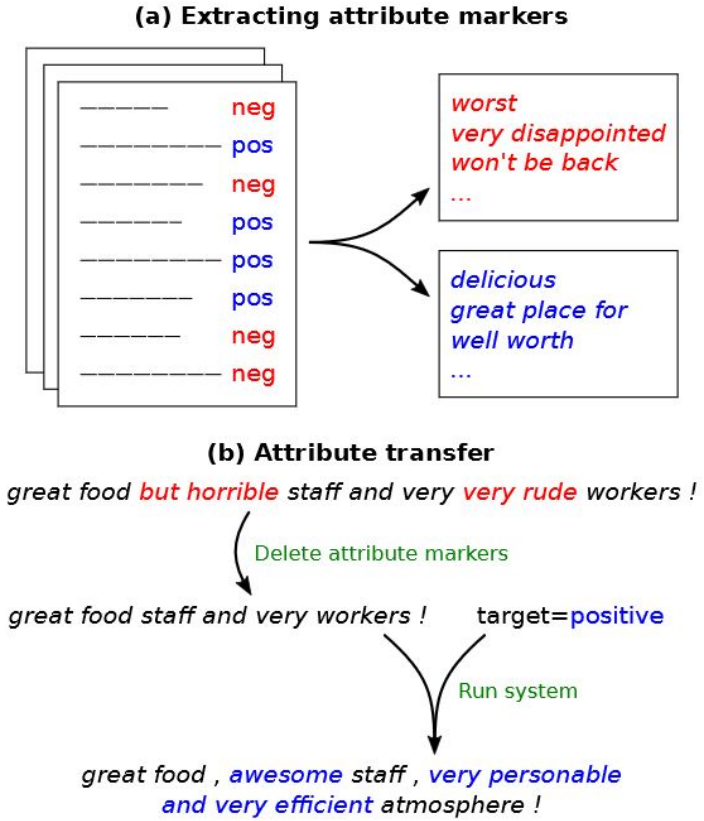
Negative → Positive	this is the worst game i have come across in a long time.	this is the best game i have come across in a long time.
Negative → Positive	we sit down and we got some really slow and lazy service	we sit down and we got some great and quick service .
Factual → Romantic	two dogs play by a tree	two dogs play by a tree, enjoying the happiness of childhood
Politics → Entertainment	how do you publish a song?	how do you handle a war?
Male → Female	Gotta say that beard makes you look like a Viking...	Gotta say that hair makes you look like a Mermaid

Previous Work

Skoltech
Skolkovo Institute of Science and Technology

Delete, Retrieve, Generate:

- Text is a combination of attributes of content and style.
- Attributes are words.
- We need to separate words into style and content attributes
- Replace the initial style attributes with the attributes of opposite style - change the style and keep the content



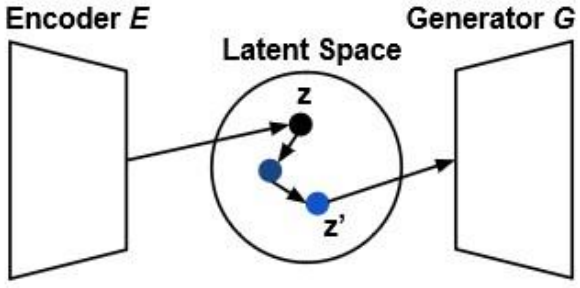
J.Li et al. (2018) Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer

Previous Work

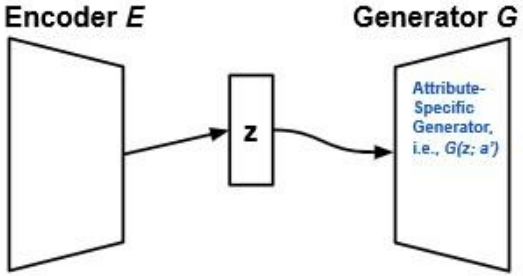
Skoltech
Skolkovo Institute of Science and Technology

Disentanglement

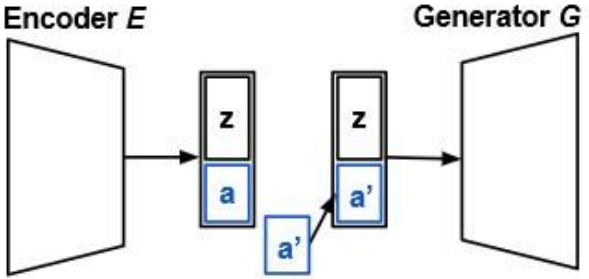
- Encode a text ***x*** with the style ***α*** to a latent representation ***z***
- Manipulate ***z*** to remove the style ***α***
- Decode ***z*** into ***y*** with the style ***α'***



edit latent representation



perform style-specific generation



split content and style latent representations

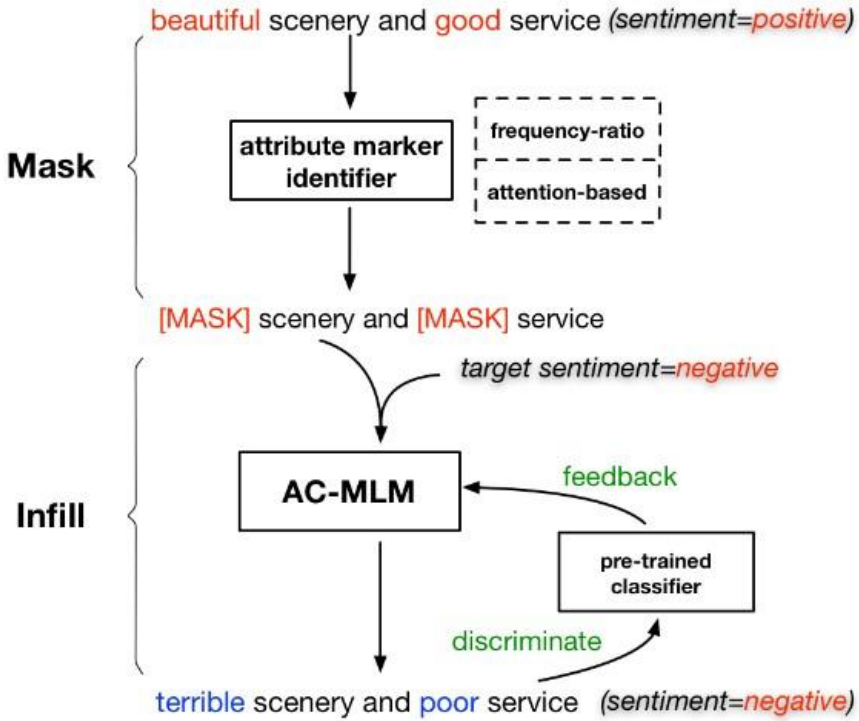
D.Jin et al. (2020) Deep Learning for Text Style Transfer: A Survey

Previous Work

Skoltech
Skolkovo Institute of Science and Technology

Mask & Infill:

- Identify words associated with style and mask them.
- Fill in the blanks with BERT fine-tuned on class-informed data.



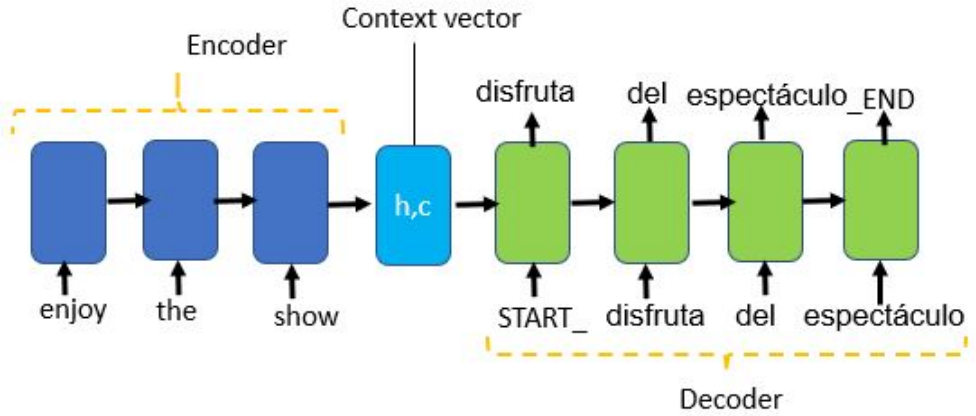
X.Wu et al. (2019) "Mask and Infill" : Applying Masked Language Model to Sentiment Transfer

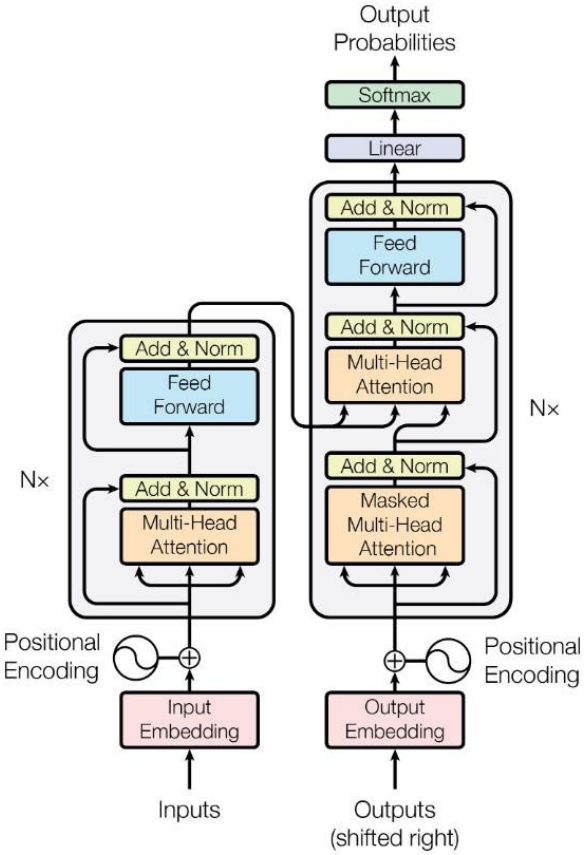
Previous Work

Skoltech
Skolkovo Institute of Science and Technology

Parallel data available: existing architectures

- style transfer as a sequence-to-sequence task (e.g. Machine Translation)
- apply standard architectures (Transformer, LSTM)





Images: https://towardsdatascience.com/implementing-neural-machine-translation-using-keras-8312e4844eb8,
A.Vaswani et al. (2017) Attention Is All You Need

Baselines

Skoltech
Skolkovo Institute of Science and Technology

Duplicate

какой долбаеб такое сделал → какой долбаеб такое сделал

Delete

какой долбаеб такое сделал → какой ~~долбаеб~~ такое сделал

Retrieve

какой долбаеб такое сделал → я сама делала такой керпичь.

detoxGPT

Skoltech
Skolkovo Institute of Science and Technology

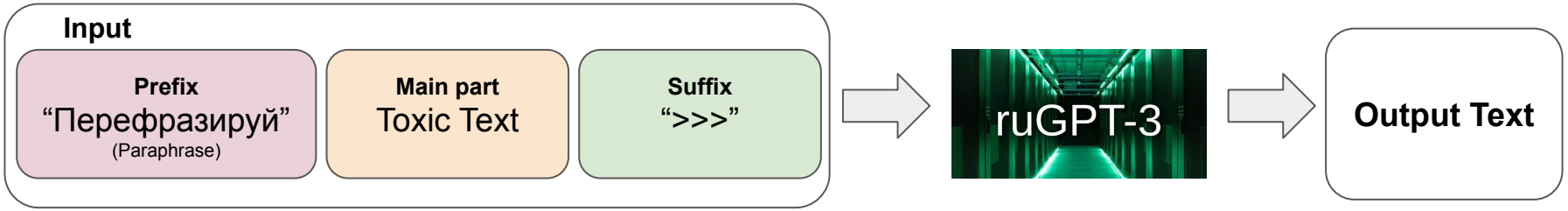
https://github.com/sberbank-ai/ru-gpts

ruGPT3XL, ruGPT3Large, ruGPT3Medium, ruGPT3Small and ruGPT2Large

detoxGPT

Skoltech
Skolkovo Institute of Science and Technology

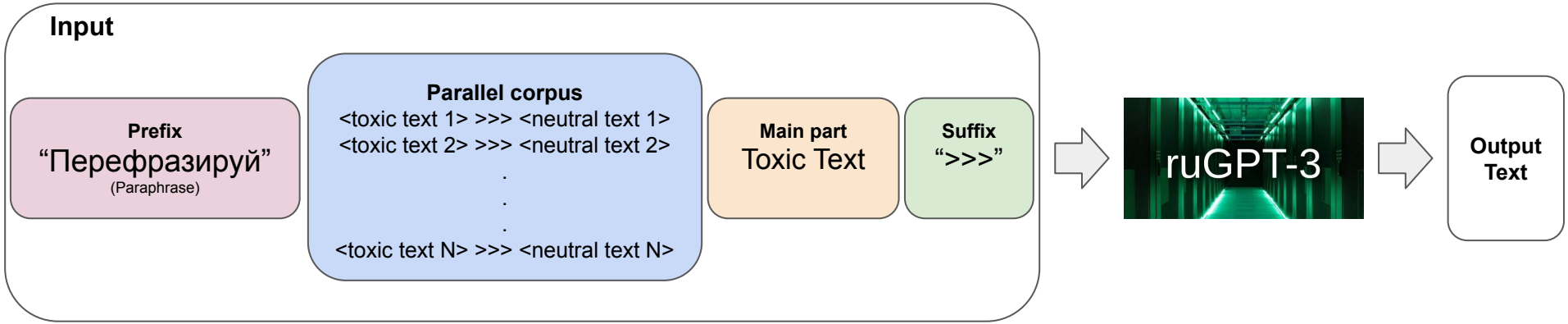
zero-shot detoxGPT



detoxGPT

Skoltech
Skolkovo Institute of Science and Technology

few-shot detoxGPT



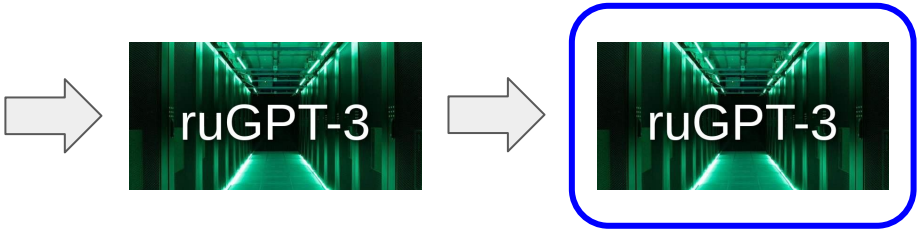
detoxGPT

Skoltech
Skolkovo Institute of Science and Technology

fine-tuned detoxGPT

fine-tuning

Parallel corpus
<toxic text 1> >>> <neutral text 1>
<toxic text 2> >>> <neutral text 2>
.
.
.
<toxic text N> >>> <neutral text N>

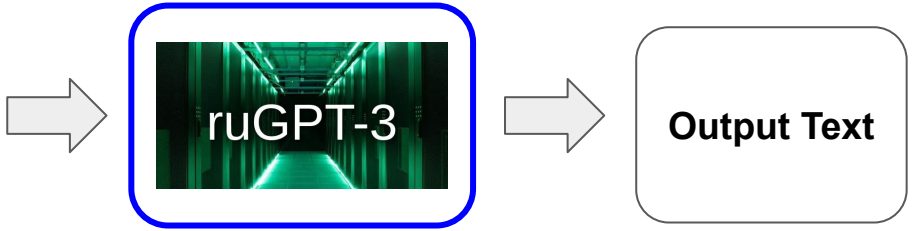


inference

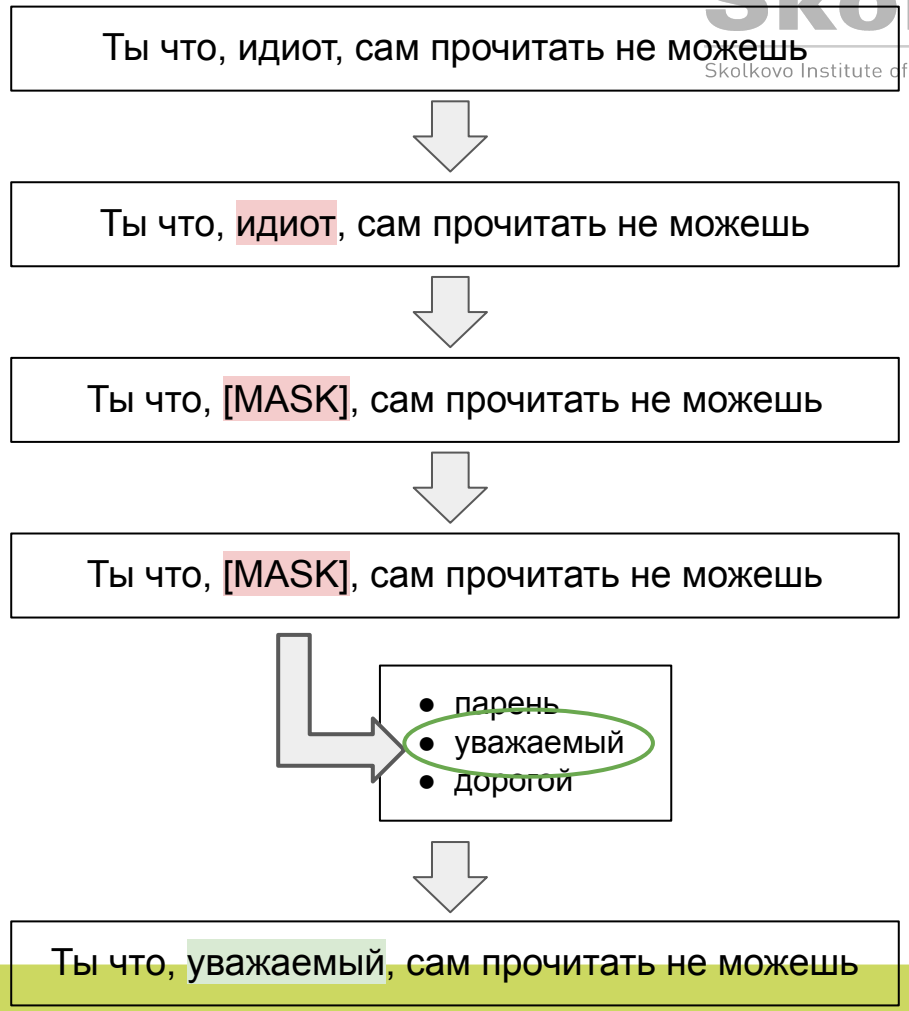
Input

Main part
Toxic Text

Suffix
“>>>”



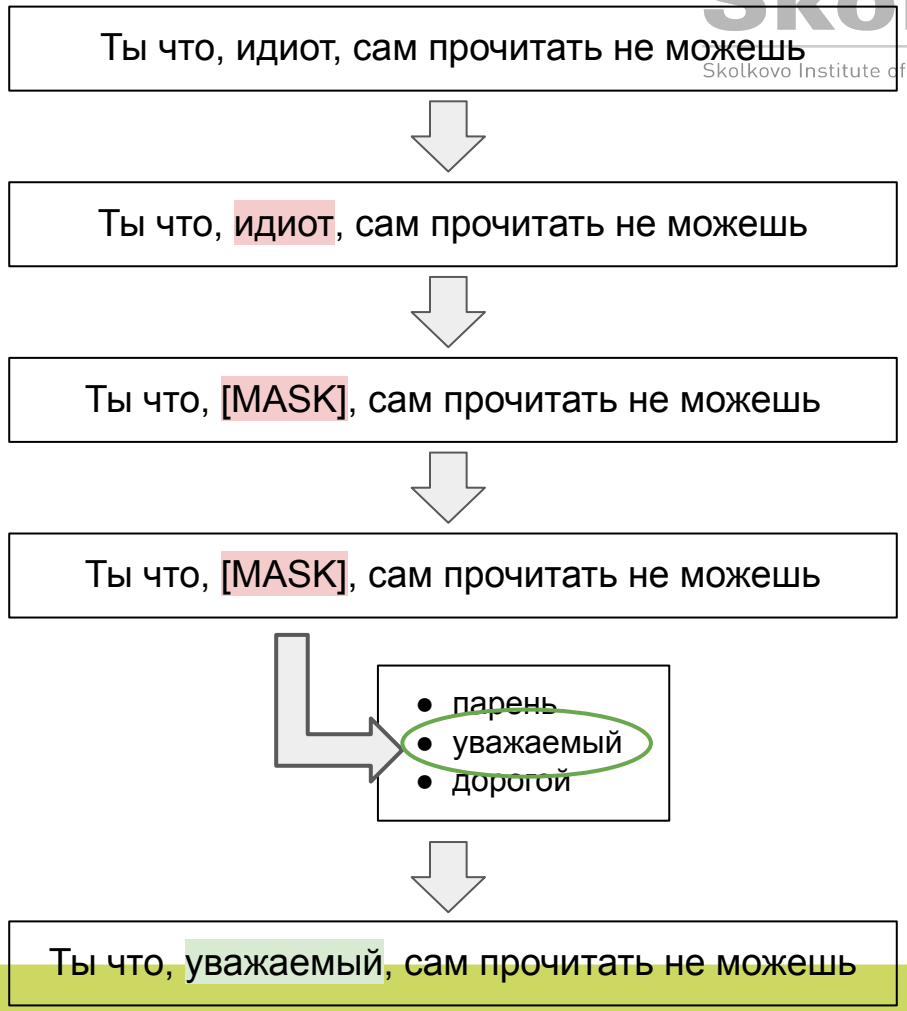
condBERT



condBERT

Setups:
- zero-shot;
- fine-tuned;

Models:
- RuBERT
- Geotrend;



Datasets

Skoltech
Skolkovo Institute of Science and Technology

train (non-parallel)

	toxic	**neutral**	**total**
RuToxic	31,407	131,780	163,187

train (parallel): 200 randomly selected samples

test (non-parallel): 10,000 randomly selected samples

Evaluation

Skoltech
Skolkovo Institute of Science and Technology

There are no reference answers, so we cannot compute BLEU.

Objectives of style transfer models:

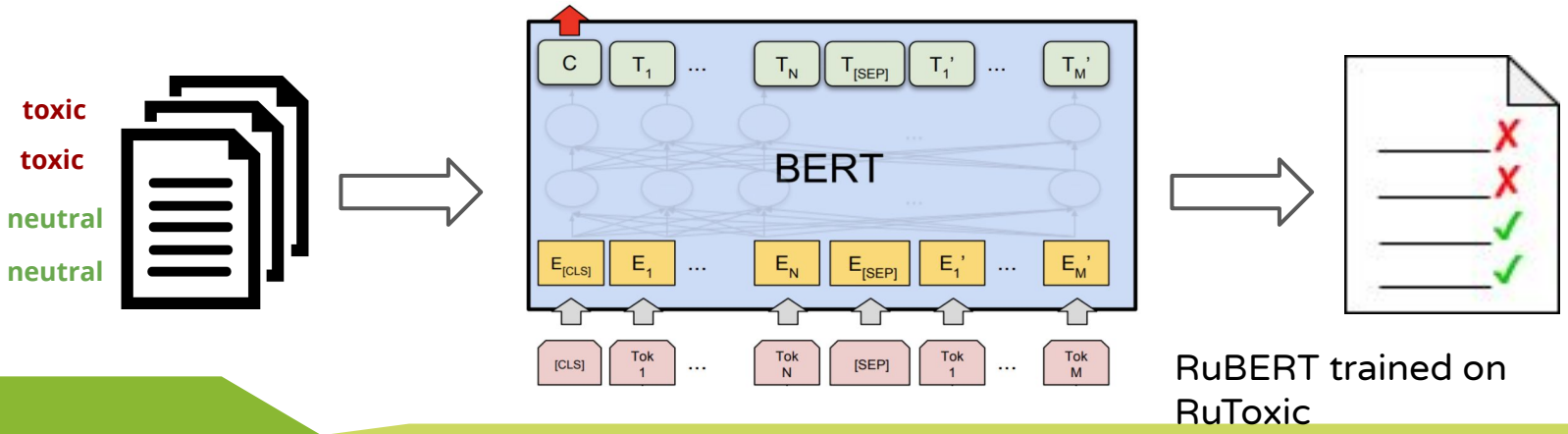
- change the style;
- save the content;
- generate a fluent sentence.

Evaluation

Skoltech
Skolkovo Institute of Science and Technology

Style Transfer Accuracy

- **train a style classifier** on sentences in the needed styles (same sentences as the ones used for style transfer model training)
- **run the classifier** on transferred test sentences
- **compute** the percentage of sentences for which the classifier shows the desired class (in our case: % of sentences labeled as **non-toxic**)



RuBERT trained on RuToxic

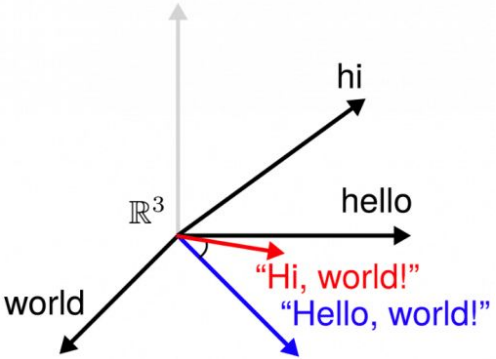
Image: J.Devlin et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Evaluation

Skoltech
Skolkovo Institute of Science and Technology

Content Preservation

- word overlap (WO): $\frac{count(x \cap y)}{count(x \cup y)}$
- BLEU: accuracy based on n-grams (1-4);
- cosine similarity (CS): between vectors of texts' embeddings.



fasttex from RusVectores

Image: https://github.com/rare-technologies/gensim/blob/develop/docs/notebooks/soft_cosine_tutorial.ipynb

Evaluation

Skoltech
Skolkovo Institute of Science and Technology

Language Quality

- **perplexity** - measures how surprised is a LM

$$2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$$

ruGPT2Large

Evaluation

Skoltech
Skolkovo Institute of Science and Technology

Aggregation Metric

$$\mathrm{GM} = (max(\mathrm{STA}, 0) \times max(\mathrm{CS}, 0) \times max(1/\mathrm{PPL}, 0))^{\frac{1}{3}}$$

Results

Skoltech
Skolkovo Institute of Science and Technology

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Results

Skoltech
Skolkovo Institute of Science and Technology

Method	Style STA↑	Content CS↑	Content WO↑	Content BLEU↑	Fluency PPL↓	Combination GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Results

Skoltech
Skolkovo Institute of Science and Technology

Style Content Fluency Combination

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06		
fine-tuned	0.51	0.70	0.05	0.0		
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.2		
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Don't change anything.
Perfectly preserves content,
doesn't change the style

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06		
fine-tuned	0.51	0.70	0.05	0.0		
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.2		
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.2		
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Delete toxic words.
Low style accuracy - toxicity is not in individual words (or we didn't find all of them).

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.2		
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.2		
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Best baseline:
- sentences selected with respect to style and similarity to the input
- low PPL (natural sentence)

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.2		
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.6		
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Meaningless output (not used in further experiments).
Fine-tuning is essential for GPT

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.2		
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

Fine-tuning helps BERT (but not too much)

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.1			0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.4			0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

BERT - slightly better style accuracy (replacements selected with respect to style)

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

BERT - better content preservation (leaves most of the sentence intact)

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned				0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov ze				0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

GPT - better fluency (generates a sentence from scratch instead of changing words)

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot	0.93	0.20	0.00	0.00	159.11	0.10 ± 0.0005
few-shot	0.17	0.70	0.05	0.06	83.38	0.11 ± 0.0009
fine-tuned	0.51	0.70	0.05	0.05	39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned	0.49	0.77	0.18	0.21	86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned					**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot					668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

GPT - better overall score due to better fluency

Results

Skoltech
Skolkovo Institute of Science and Technology

Baselines

Our GPT-based methods

Our BERT-based methods

Method	STA↑	CS↑	WO↑	BLEU↑	PPL↓	GM↑
Duplicate	0.00	1.00	1.00	1.00	146.00	0.05 ± 0.0012
Delete	0.27	0.96	0.85	0.81	263.55	0.10 ± 0.0007
Retrieve	0.91	0.85	0.07	0.09	65.74	0.22 ± 0.0010
detoxGPT-small						
zero-shot					159.11	0.10 ± 0.0005
few-shot					83.38	0.11 ± 0.0009
fine-tuned					39.48	0.20 ± 0.0011
detoxGPT-medium						
fine-tuned					86.75	0.16 ± 0.0009
detoxGPT-large						
fine-tuned	0.61	0.77	0.22	0.21	**36.92**	**0.23*** ± 0.0010
condBERT						
DeepPavlov zero-shot	0.53	0.80	0.42	0.61	668.58	0.08 ± 0.0006
DeepPavlov fine-tuned	0.52	0.86	0.51	0.53	246.68	0.12 ± 0.0007
Geotrend zero-shot	0.62	0.85	0.54	**0.64**	237.46	0.13 ± 0.0009
Geotrend fine-tuned	**0.66**	**0.86**	**0.54**	0.64	209.95	0.14 ± 0.0009

GPT - closely followed by a baseline.
Caveat - content preservation is unreliable (retrieved sentences are not similar to the input)

Results

Skoltech
Skolkovo Institute of Science and Technology

Model	Sentence
Input	не дай бог моя дочь так оденется убью нахуй палкой (If, God forbid, my daughter goes out dressed like this, I'll fucking kill her with a stick)
Delete	не дай бог моя дочь так оденется убью палкой (If, God forbid, my daughter goes out dressed like this, I'll kill her with a stick)
Retrieve	не бросайте угла родного одной мы лежали больнице палате в в в те девчонкой была молодой годы (don't abandon your native corner same hospital we were ward in in in those girl was young years)
detoxGPT-small	подскажите, у меня дочь так оденется палкой. (Could you give me an advise, my daughter gets dressed as a stick.)
detoxGPT-medium	да не торопись вы так, бабушка. (Granny, don't rush, take your time.
detoxGPT-large	извините, но мне вас очень жаль, если вы начнете так одеваться. (Excuse me, but I feel sorry for you if you start getting dressed up like this.)
condBERT	не дай бог моя дочь так оденется станки палкой (God forbid, my daughter gets dressed up like this machine stick.)

Results

Skoltech
Skolkovo Institute of Science and Technology

Model	Sentence
Input	что ты читала из его произведений? дура.а если не читала,то прочти. (Which of his works have you read? You are an idiot. If you haven't read any, go read.)
Delete	что ты читала из его произведений дура.а если не читала,то прочти (Which of his works have you read you idiot. If you haven't read any, go read.)
Retrieve	одно не пойму, вы все реально думаете, что они читают что то в ок или в вк? (What I can't understand - do you all really think they read anything in Odnoklassniki or Vkontakte?)
detoxGPT-small	вы читали его произведения? не читали. (Have you read any of his works? No, you haven't)
detoxGPT-medium	его посадить, извините, но мне вас очень жаль, если для вас оскорбления - норма. (Putting him in jail, excuse me, but I feel sorry for you if insult is normal for you.)
detoxGPT-large	вы читали его произведения? недалекие. попробуйте почитать посты его друзей (Have you read any of his works? You are narrow-minded. Try reading his friends' posts.)
condBERT	что ты читала из его произведений ? пусть . а если не читала , то прочти . (Which of his works have you read? Okay. If you haven't read any, go read.)

Results

Skoltech
Skolkovo Institute of Science and Technology

Model	Sentence
Input	тебя вношу в список долбаебов. (I'm adding you to the list of buttheads.)
Delete	тебя вношу в список (I'm adding you to a list)
Retrieve	надо обнародовать этот список. чего молчать. (We should make this list public. Why keeping silence?)
detoxGPT-small	Вас вношу в список людей, нуждающихся в улучшении. (I'm adding you to the list of people who need being improved.)
detoxGPT-medium	Вас вношу в список людей, которые вносят вклад в мой тред. (I'm adding you to the list of people who contribute to this thread.)
detoxGPT-large	Вас вношу в список людей, которые не соответствуют вашим ожиданиям. (I'm adding you to the list of people who don't meet your expectations.)
condBERT	тебя вношу в список до (I'm adding you to the list of to.)

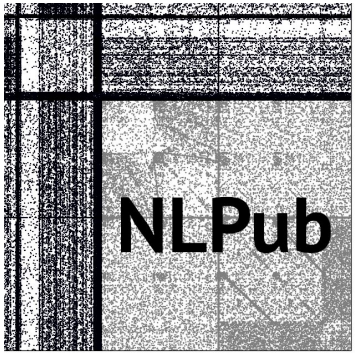
Links & Contacts

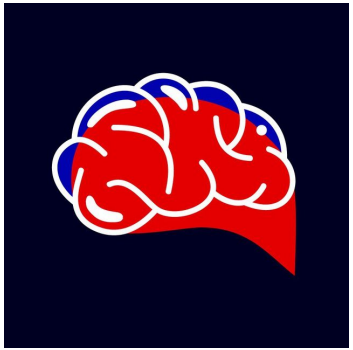
Skoltech
Skolkovo Institute of Science and Technology

Code & data:

https://github.com/skoltech-nlp/rudetoxifier

Skoltech Natural Language Processing





МТС

rusemantics@facebook

TowardsNLP@telegram

@dementyeva_ds
daryna.dementieva@skoltech.ru