

# Monolingual and Cross-lingual Text Detoxification

Alexander Panchenko

17.03.2023

Skoltech, AIRI

# Structure of the talk

- Monolingual text detoxification (prior results)
- Cross-lingual text detoxification (new results)

# About me

- Making internationally recognised R&D in the field of NLP/CL since 2008



- 2019-...: **Skoltech**, Assistant → Associate Professor



- 2017-2019: **University of Hamburg**, Postdoc in NLP



- 2015-2017: **TU Darmstadt**, Postdoc in NLP



- 2013-2015: **Startup in SNA**, Research Engineer in NLP



- 2009-2013: **Université catholique de Louvain**, PhD in NLP



- 2008: **Xerox Research Centre Europe**, Research Intern NLP



- 2007: **Université catholique de Louvain**, Research Intern NLP



- 2002-2008: **Bauman MSTU**, Computer Science, Engineer



# About me

- **Experience with publications in pretty much all TOP int'l conferences & journals in NLP:**
  - ACL, COLI, EMNLP, NAACL, EACL, ECIR, Natural Language Engineering, ...
- **Best papers:** Representation Learning for NLP and SemEval workshops.
- **Editor/chair/area chair:**
  - Cambridge Natural Language Engineering (NLE)
  - Springer LNCS/CCIS: AIST conference
  - EACL-2021, 2023
  - Various \*ACL workshops and more
- **Program Committee (PC) member** in pretty much all top conferences in NLP:
  - ACL, NAACL, EMNLP, ...



CAMBRIDGE  
UNIVERSITY PRESS



Springer



Association for  
Computational Linguistics

# Acknowledgement for collaborators and colleagues

Daryna Dementieva, David Dale, Varvara Logacheva,  
Daniil Moskovskiy, Irina Krotova, Daniil Moskovskiy, Anton Voronov,  
Anton Markov, Sergey Ustyantsev, Olga Kozlova, Alena Fenogenova,  
Tatiana Shavrina, Irina Nikishina, Nikita Semenov



David



Daryna



Varvara

Special acknowledgement for slides: Daryna Dementieva

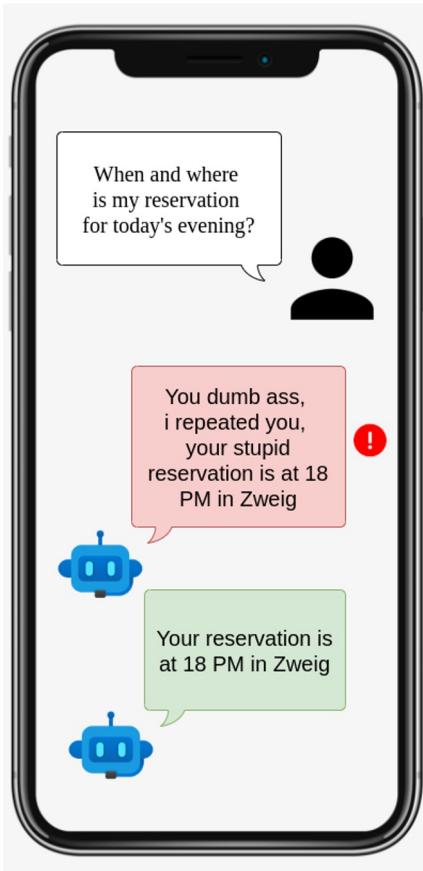
Project was conducted within the framework of **MTS-Skoltech joint laboratory**:



# Based on several publications including top venues, such as ACL and EMNLP, Master and PhD theses

- Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., Semenov, N., and Panchenko, A. (2022): **ParaDetox: Detoxification with Parallel Data**. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.
- Dale, D., Voronov, A., Dementieva, D., Logacheva, V., Kozlova, O., Semenov, N. and Panchenko, A. (2021): **Text Detoxification using Large Pre-trained Neural Models**. In Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021). Punta Cana, Dominican Republic.
- Dementieva, D.; Moskovskiy, D.; Logacheva, V.; Dale, D.; Kozlova, O.; Semenov, N.; Panchenko, A. **Methods for Detoxification of Texts for the Russian Language**. Multimodal Technol. Interact. 2021, 5, 54.
- Logacheva, V., Dementieva, D., Krotova, I., Fenogenova, A., Nikishina, I., Shavrina, T., and Panchenko, A. (2022): **RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora**. In Proceedings of the Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”. Moscow, Russia (Online)
- Moskovskiy, D., Dementieva, D., and Panchenko, A. (2022): **Exploring Cross-lingual Text Detoxification with Large Multilingual Language Models**. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 346–354, Dublin, Ireland. Association for Computational Linguistics.
- Logacheva, V., Dementieva, D., Krotova, I., Fenogenova, A., Nikishina, I., Shavrina, T., and Panchenko, A. (2022): **A Study on Manual and Automatic Evaluation for Text Style Transfer: The Case of Detoxification**. In Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 90–101, Dublin, Ireland. Association for Computational Linguistics.
- Dementieva, D., Ustyantsev, S., Dale, D., Kozlova, O., Semenov, N., Panchenko, A., and Logacheva, V. (2021): **Crowdsourcing of Parallel Corpora: the Case of Style Transfer for Detoxification**. Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021). Copenhagen, Denmark

# Problem: toxicity of automatic generation by neural models



Chatbots trained on open data (for example, user comments) sometimes begin to offend users. This is frustrating for the user and ultimately damages the company's reputation.

This situation can be prevented by filtering chatbot remarks and generating a polite message.

Tinkoff's chat bot offensive response as an example:

<https://meduza.io/news/2019/11/26/chat-bot-tinkoff-banka-posovetoval-klientke-otrezal-paltsy-v-banke-eto-ob-yasnili-obucheniem-na-otkrytyh-dannyh>

# Text Style Transfer - Task Formulation

Rewrite the text to:

- eliminate toxicity;
- save the text content as much as it is possible;
- results should be grammatically correct;

These n\*\*\*ers are b\*\*gars → These black people are poor

You are a f\*\*king \*\*iot if you do this → I don't think your solution is well thought.

Stupid peace of s\*\*t stop deleting my stuff a\*\*\*ole → Stop deleting my stuff.

# Automatic Evaluation

**Style Transfer Accuracy** (STA)

**Content Similarity** (SIM)

**Fluency** (FL)

**Joint score** (J)

# Automatic Evaluation: English

## Style Transfer Accuracy (STA)

RoBERTa-based Toxicity classifier fine-tuned on Jigsaw Dataset

## Content Similarity (SIM)

cosine similarity based on LaBSE embeddings

## Fluency (FL)

RoBERTa-based language acceptability classifier fine-tuned on CoLa dataset

## Joint score (J)

+ BLEU

$$\mathbf{J} = \frac{1}{n} \sum_{i=1}^n \mathbf{STA}(x_i) \cdot \mathbf{SIM}(x_i) \cdot \mathbf{FL}(x_i)$$

# Automatic Evaluation: Russian

## Style Transfer Accuracy (STA)

BERT Toxicity classifier fine-tuned on Odnoclassniki+Pikabu Dataset

## Content Similarity (SIM)

cosine similarity based on LaBSE embeddings

## Fluency (FL)

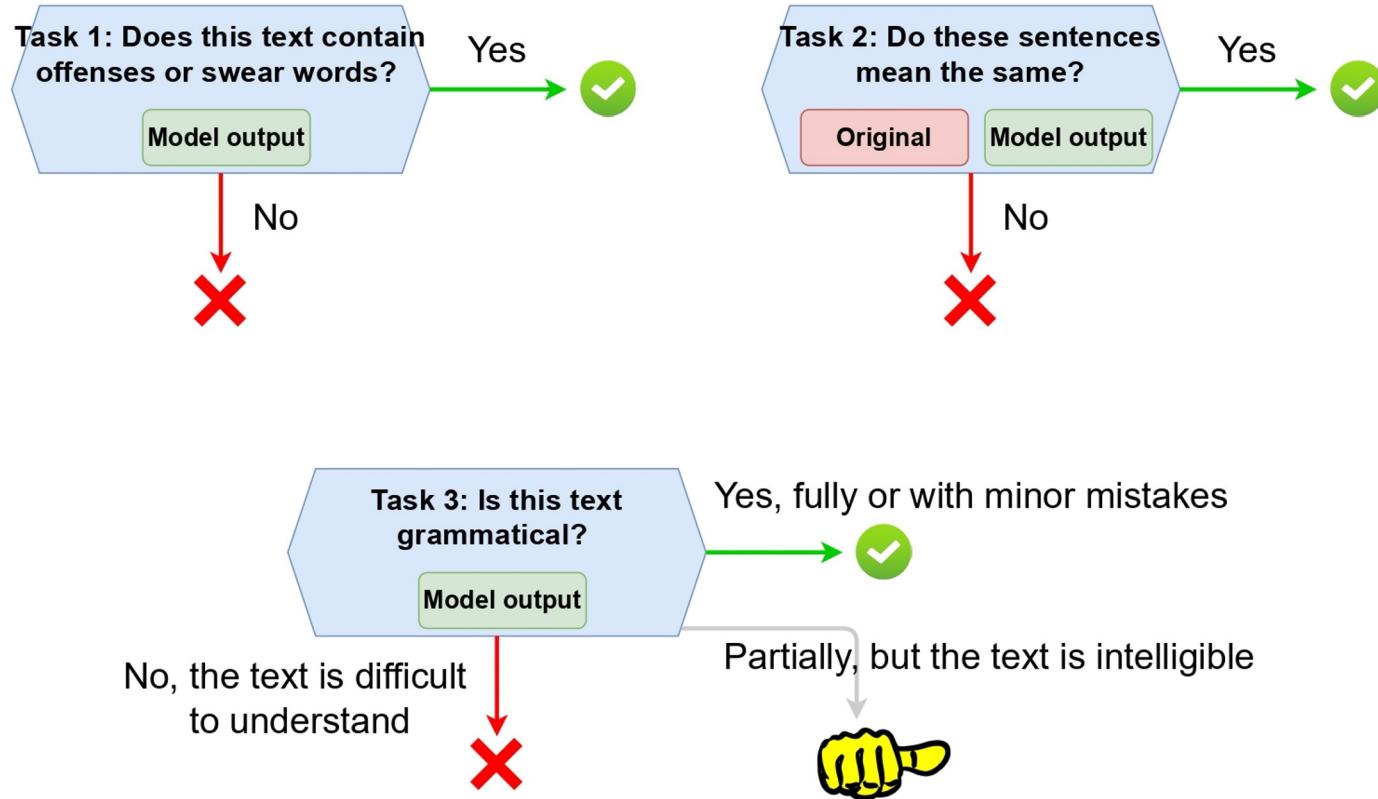
BERT language acceptability classifier fine-tuned on pre-collected dataset

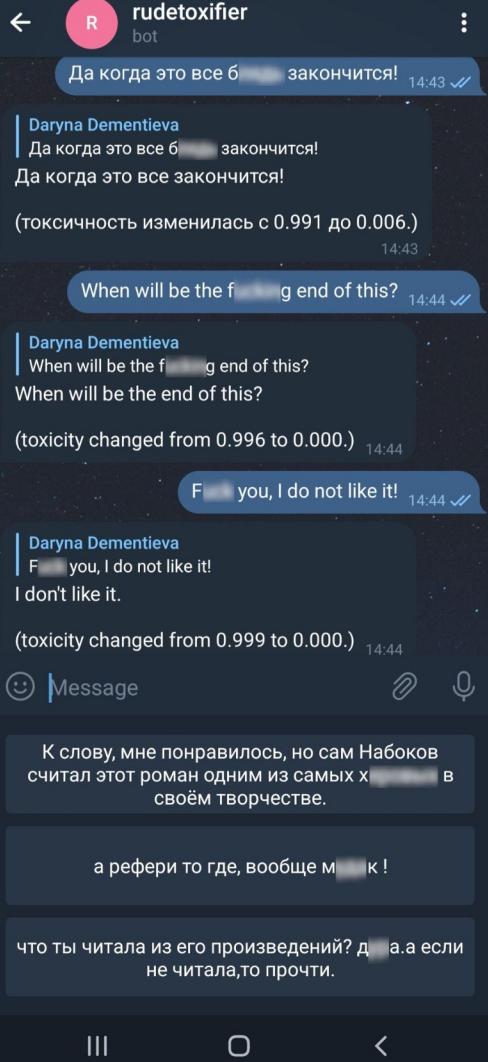
## Joint score (J)

+ ChrF1

$$\mathbf{J} = \frac{1}{n} \sum_{i=1}^n \mathbf{STA}(x_i) \cdot \mathbf{SIM}(x_i) \cdot \mathbf{FL}(x_i)$$

# Manual Evaluation





# System Demonstration

<https://detoxifier-nlp-zh.skoltech.ru>

## Empathy Guard

does he do this s... t all the time?

Randomize  
Toxicity threshold 0.5

Rewrite

Original:

**does** he do this s... t all the time?

Rewritten:

**Does** he do this all the time?

Old toxicity:

0.9980

New toxicity:

0.0000

@rudentoxifierbot. <https://t.me/rudentoxifierbot>

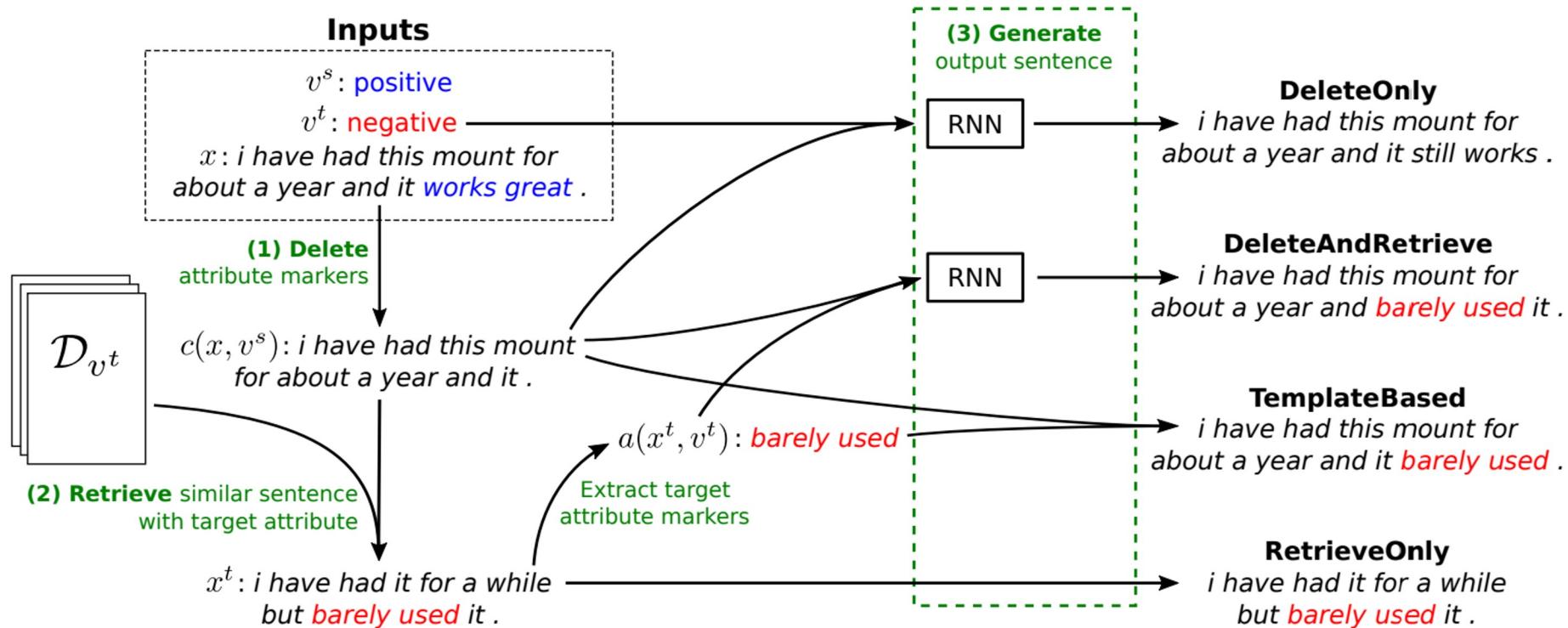
# Previous Work

**Non-parallel datasets**

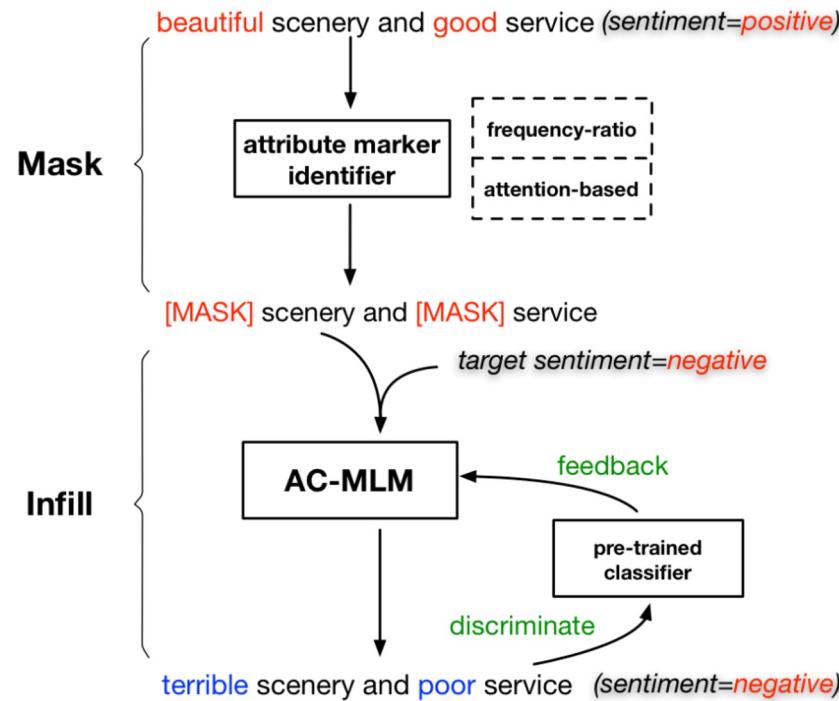
**Parallel datasets**



# Delete, Retrieve, Generate



# Mask & Fill



# condBERT

Conditional BERT Model  
for TST

- MWE support
- No finetuning for MLM

EMNLP  
2021

I hate reading this s\*\*t.



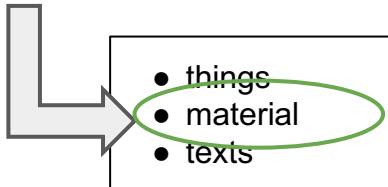
I hate reading this s\*\*t.



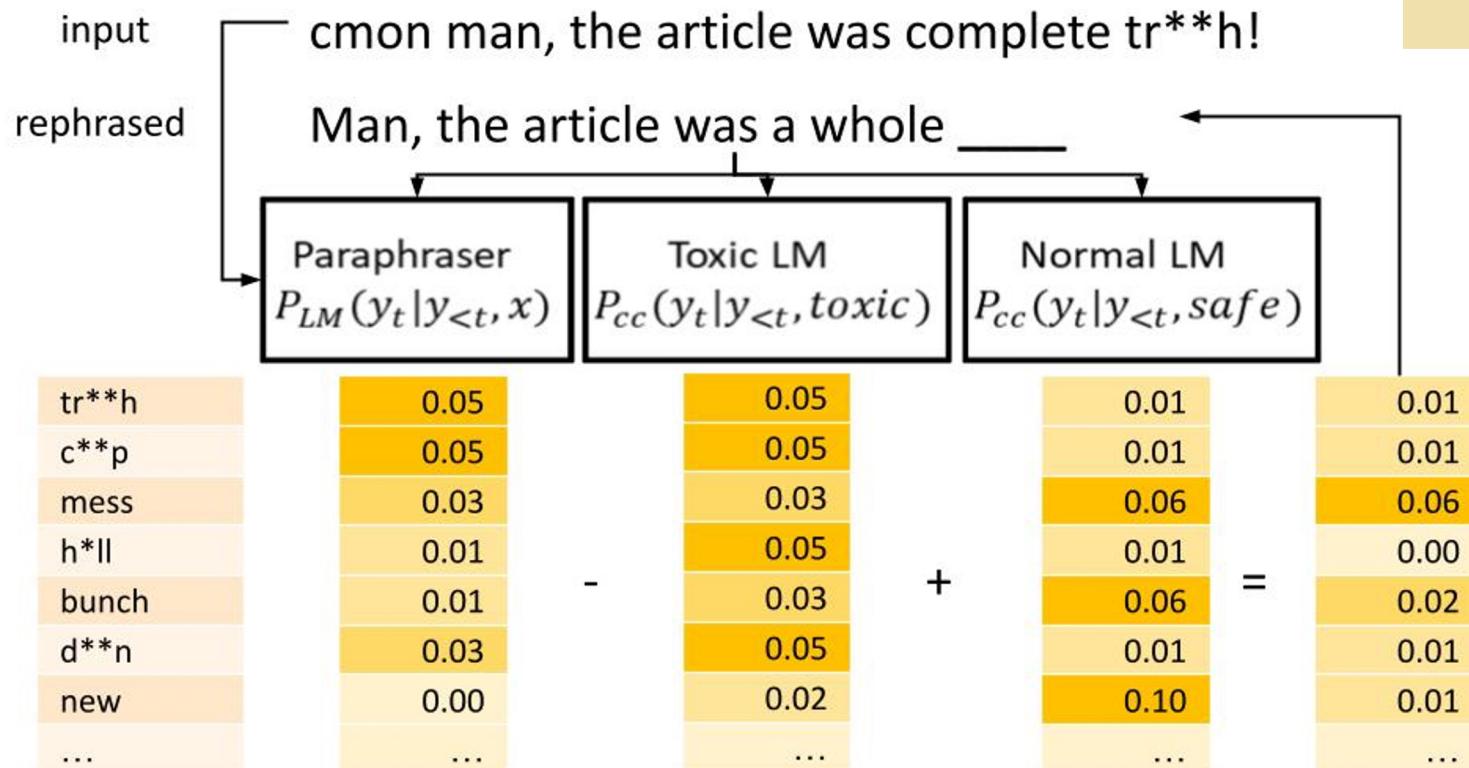
I hate reading this [MASK].



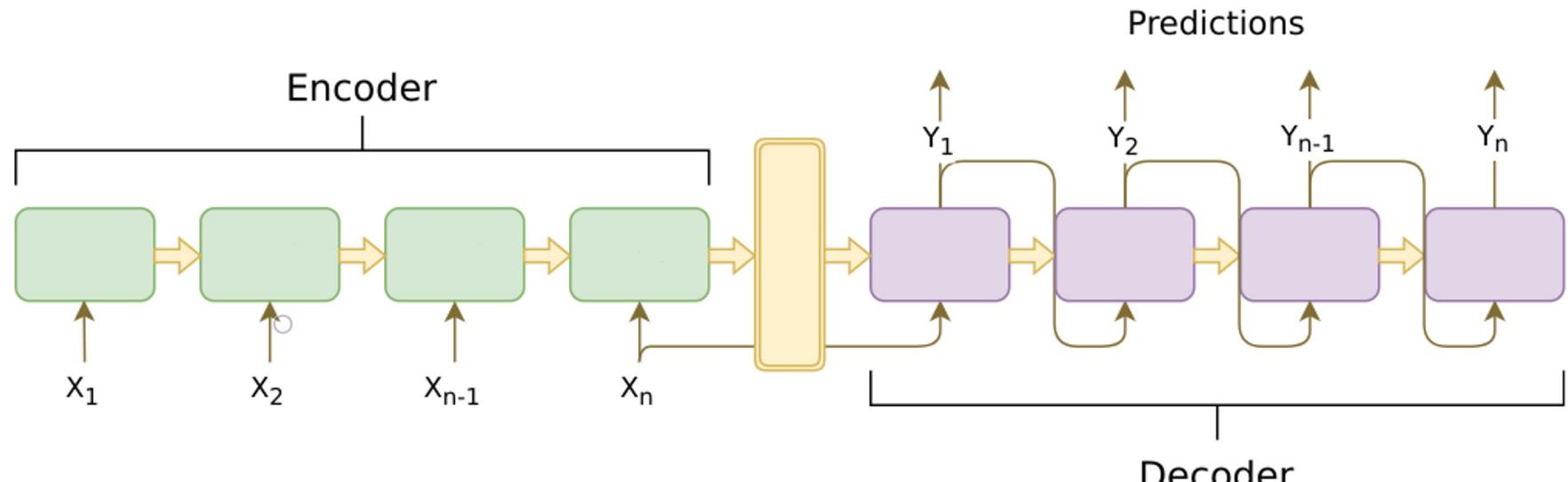
I hate reading this [MASK].



I hate reading this material.



# Seq2Seq (Supervised) Approaches



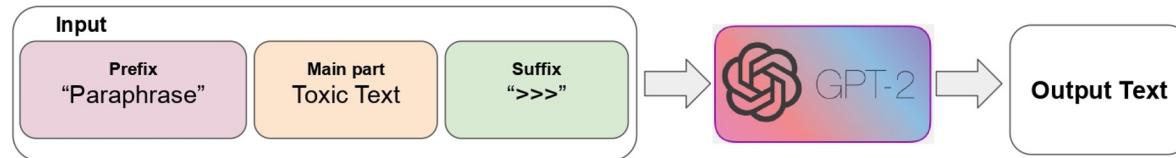
$$\mathbf{d}^{src} = (x_1, \dots, x_n) \quad \mathbf{d}^{tg} = (y_1, \dots, y_m)$$

$$\mathcal{L}_{CE}(\mathbf{d}^{src}, \mathbf{d}^{tg}) = \frac{1}{n} \sum_{i=1}^n -\log p_\theta(y_i | \mathbf{d}^{src}, \theta)$$

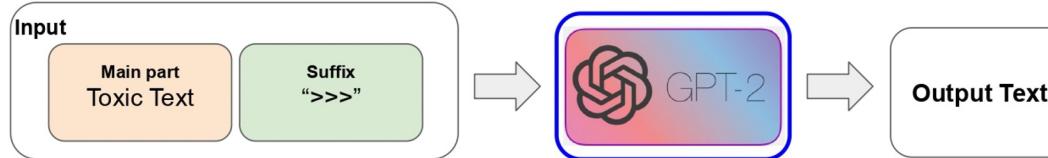
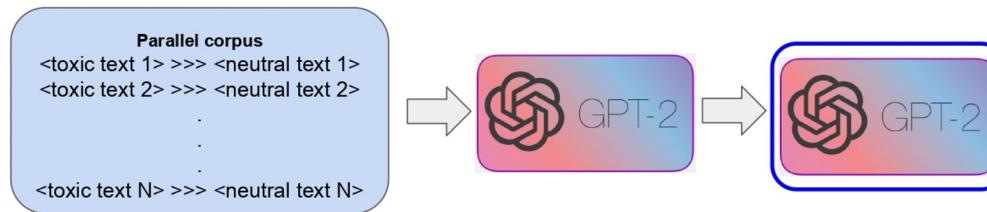
**Parallel Dataset required**

# Seq2Seq (Supervised) Approaches

*zero-shot seq2seq*



*fine-tuned seq2seq*



**Parallel Dataset required**

# Parallel Corpora: Previous Work

- Bible dataset:

Source	target
<p>(BBE) <i>Then Samuel gave him an account of everything, keeping nothing back. And he said, It is the Lord; let him do what seems good to him.</i></p>	<p>(ASV) <i>And Samuel told him every whit, and hid nothing from him. And he said, It is Jehovah: let him do what seemeth him good.</i></p>

- GYAF dataset:

Formal: Informal:	<i>I'd say it is punk though. However, I do believe it to be punk.</i>
Formal: Informal:	<i>Gotta see both sides of the story. You have to consider both sides of the story.</i>

Datasets were  
collected or  
verified  
manually

# Toxicity Datasets

- Jigsaw Dataset:

toxic samples	non-toxic samples	total
158 984	1 781 968	1 940 952

- Reddit&Twitter Dataset:

toxic samples	non-toxic samples	total
581 578	13 542 894	14 124 472

**Non-parallel Datasets**

# ParaDetox: Detoxification with Parallel Data

ACL  
2022

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. [\*ParaDetox: Detoxification with Parallel Data\*](#). In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

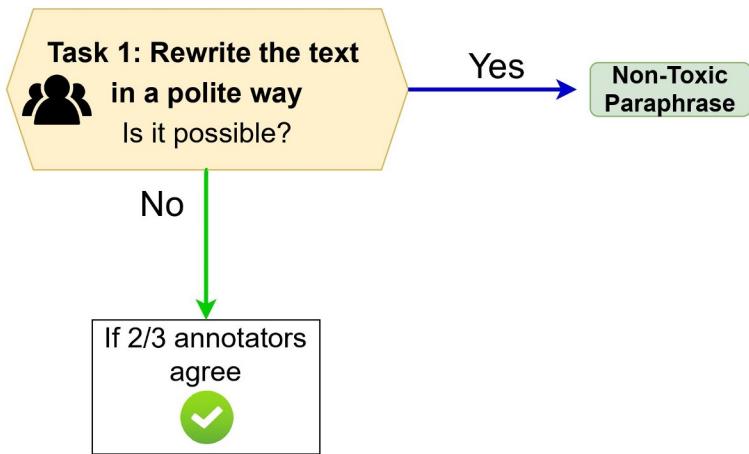
# Hypothesis

Trained machine learning models on **parallel corpus** of detoxification samples will gain higher performance on a detoxification task than trained on non-parallel ones.

# Definition of Toxicity

Input text	Toxicity type
<i>clearly the french are a nation of m**ogynists</i>	direct racism ✘
<i>lying anti american m*slim m***thpiece</i>	direct racism ✘
<i>your *gnorance makes me laugh.</i>	passive aggressiveness ✘
<i>i think sen ron johnson need to *xamine his brain.</i>	passive aggressiveness ✘
<i>you s*ck sand n***er p***y!</i>	severe toxic ✘
<i>f*ck off you stupid *spy a***ole</i>	severe toxic ✘
<i>f**k you, i wont do what you tell me.</i>	obscene ✓
<i>what a chicken c**p excuse for a reason.</i>	toxic ✓

# Task 1: Paraphrase Generation



For example:

- “It s\*\*ks that you’re an awful person”;
- “Maybe they should deport you back to your country, or your grandparents country”;

then burn their homes , then their wives .

Rewrite the text so that it does not sound offensive and its meaning stays the same.

Your text

1 / 5

I can't rewrite this text

The text is meaningless

The text is not offensive

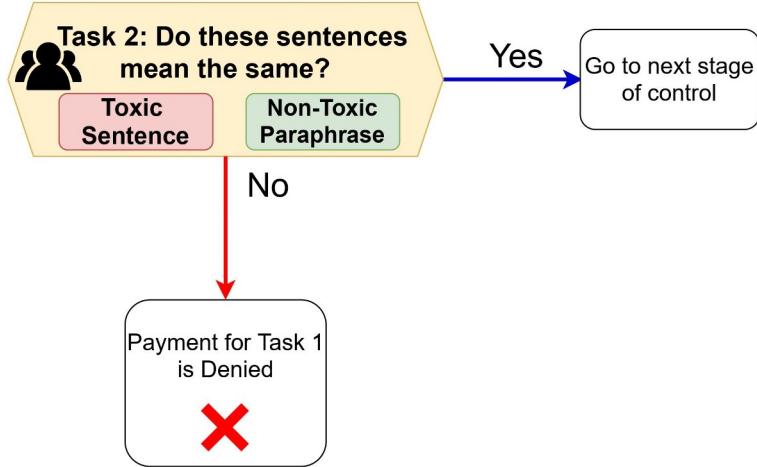
Removing the offence will change the meaning significantly

Other reason

→

This interface shows a text input field with placeholder 'Your text'. Above it is a sentence with a missing word. Below the input field are five radio button options for selecting a reason why the text cannot be rewritten. On the right side, there is a vertical toolbar with a progress bar showing '1 / 5', a target icon, a red X icon, and a right-pointing arrow icon.

# Task 2: Content Preservation Check



**Goal:**

verify main content preservation in resulting paraphrase.

Do these two sentences mean the same thing?



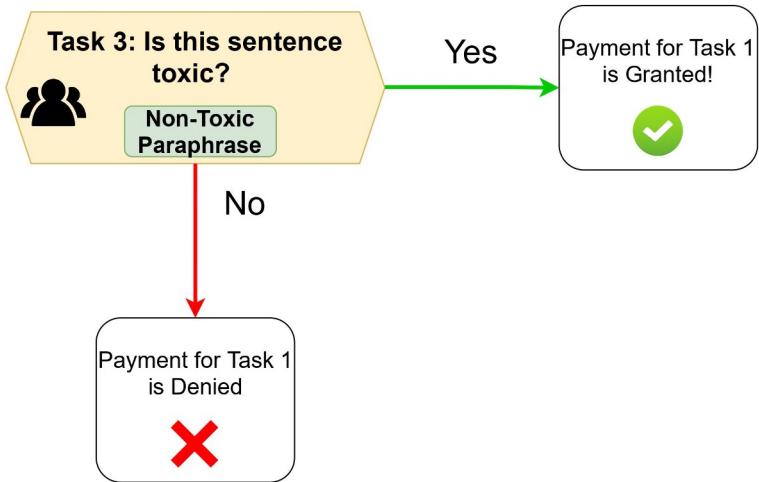
it sure as [REDACTED] had nothing to do with anything i said .



It certainly had nothing to do with anything I said.

1  Yes    2  No

# Task 3: Toxicity Check



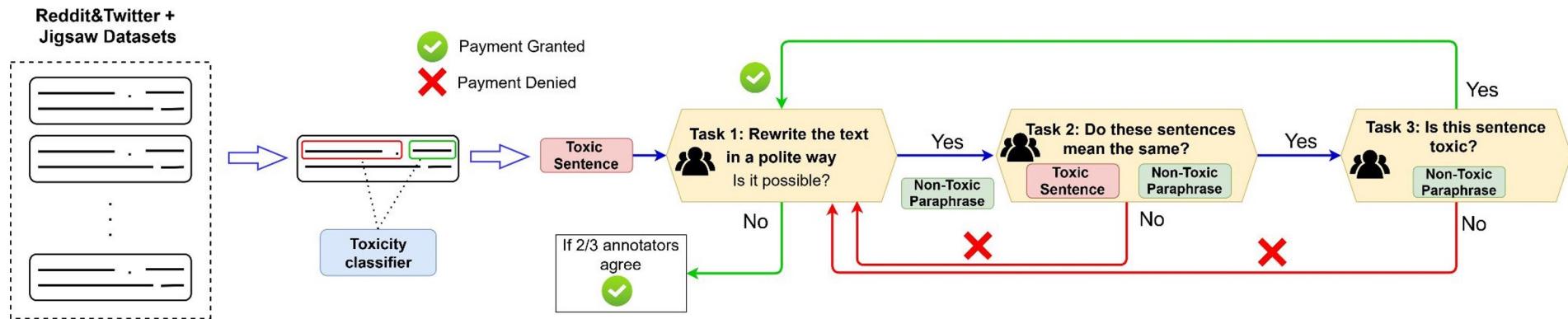
## Goal:

verify style transfer to non-toxic in resulting paraphrase.

A screenshot of a user interface for a toxicity check. At the top, a blurred text input field contains the instruction "yourself and walk away." Below it, a question asks "Does this text contain offenses or swear words?". Two radio button options are shown: "y" (Yes) and "n" (No), each accompanied by a radio button for selection.

# ParaDetox Pipeline

## Parallel Detoxification



# Publicly Available Sources

## English ParaDetox:

- 12,000 toxic sentences with 1-3 non-toxic paraphrases;
- The cost is \$811;



<https://github.com/s-nlp/paradetox>

## Russian ParaDetox:

- 8,500 toxic sentences with 1-3 non-toxic paraphrases;
- The cost is \$680;



[https://github.com/s-nlp/russe\\_detox\\_2022](https://github.com/s-nlp/russe_detox_2022)

# English ParaDetox Samples

Toxic text	Neutral text
this is scaring the s**t out of me.	This is really scaring me.
everybody is such a f**king p***y.	Everybody is acting cowardly.
calm the f**k down, cnn.	Please calm your nerves, CNN.
all you trump cl*wns are seriously messed up.	Trumps voters are seriously mislead.
its a crock of s**t , and you know it.	It's senseless, you know

Manual validation: <10% of inappropriate samples

# Russian ParaDetox Samples

Токсичный текст	Нейтральный текст
тв*рина ***ная если это ее слова	Она очень неправа, если это действительно её слова.
п****ы как вы за***ли уже	Как вы уже надоели
не убил так у**ал, браво!	не убил так ударил, браво!
А не пошел бы ты на***	Отстань от меня
такой п****ж я читаю впервые!	такое вранье я читаю впервые!

Manual validation: <10% of inappropriate samples

# **Detoxification Methods**

# Simple baselines

## Duplicate

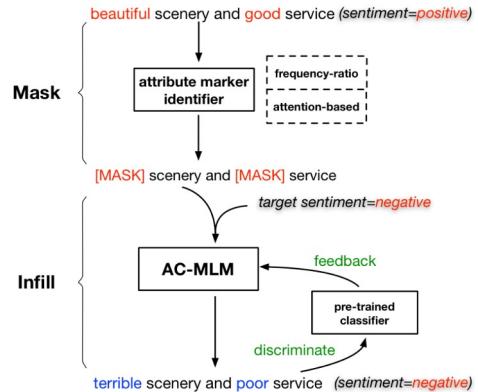
какой д\*\*\*\*б такое сделал → какой д\*\*\*\*б такое сделал  
what a f\*\*k did that → what a f\*\*k did that

## Delete

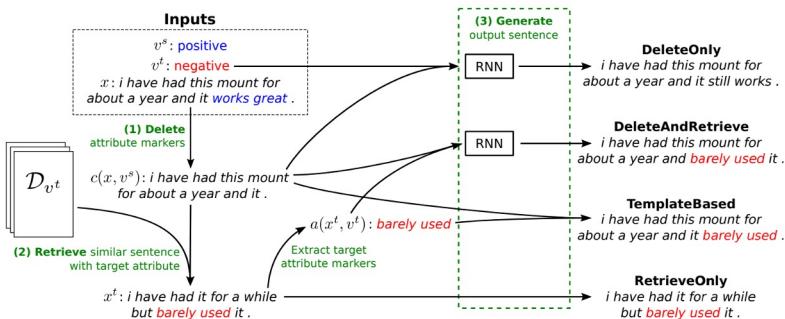
какой д\*\*\*\*б такое сделал → какой д\*\*\*\*б такое сделал  
what a f\*\*k did that → what a f\*\*k did that

# State-of-the-art baselines

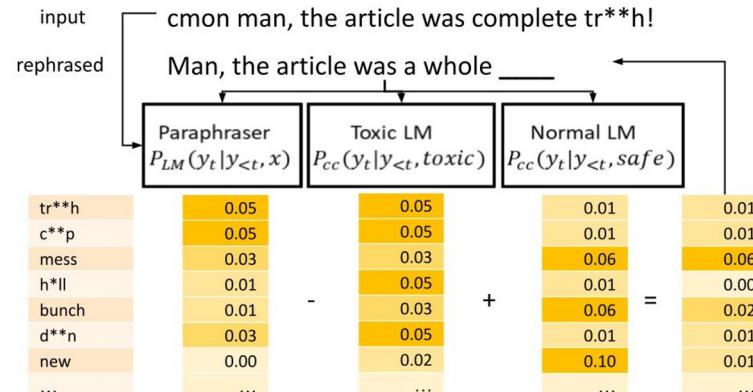
## Mask&Fill



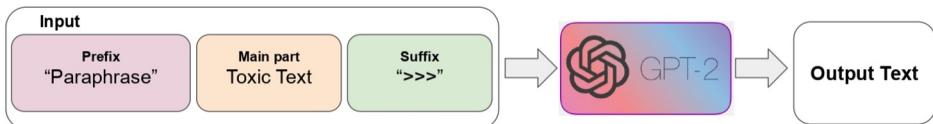
## DRG



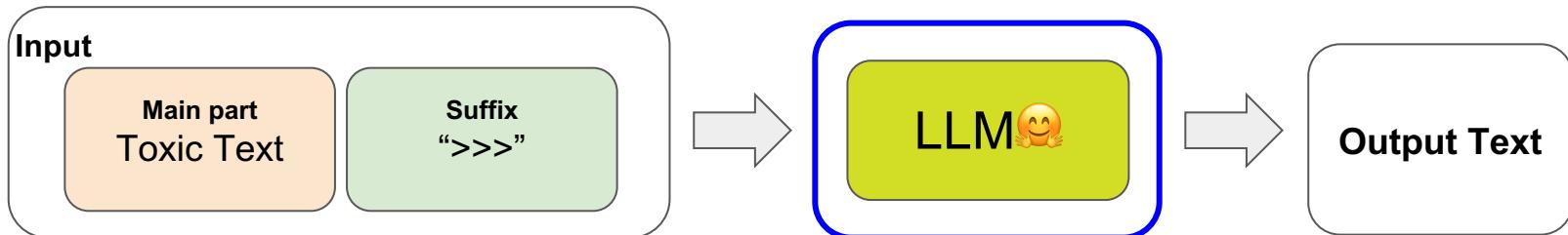
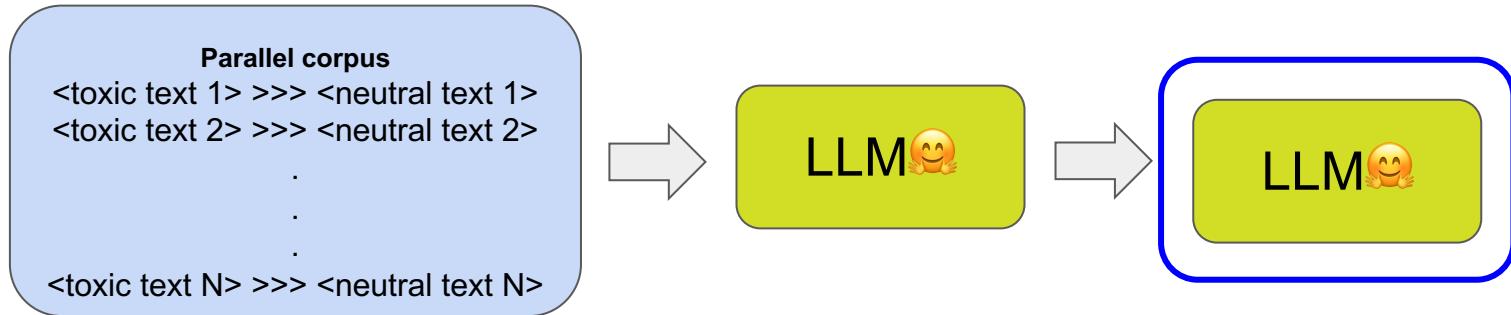
## ParaGedi



## Zero-shot Seq2seq



# Seq2Seq (Supervised) Approaches



$$\mathbf{d}^{src} = (x_1, \dots, x_n) \quad \mathbf{d}^{tg} = (y_1, \dots, y_m)$$

$$\mathcal{L}_{CE}(\mathbf{d}^{src}, \mathbf{d}^{tg}) = \frac{1}{n} \sum_{i=1}^n -\log p_{\theta}(y_i | \mathbf{d}^{src}, \theta)$$

**GPT**

**T5**

**BART**

# Results: English

	BLEU	STA <sub>a</sub>	SIM <sub>a</sub>	FL <sub>a</sub>	J
Human reference	100.0	0.96	0.77	0.88	0.66
Baselines and SOTA (unsupervised)					
Delete	61.24	0.81	0.93	0.64	0.46
Duplicate	53.86	0.02	1.0	0.91	0.02
DRG-Template	53.86	0.90	0.82	0.69	0.51
BART-zero-shot	53.64	0.01	<b>0.99</b>	0.92	0.01
Mask&Infill	52.47	0.91	0.82	0.63	0.48
CondBERT	42.45	<b>0.98</b>	0.77	0.82	0.62
SST	30.20	0.86	0.57	0.19	0.10
ParaGeDi	25.39	<b>0.99</b>	0.71	0.88	0.62
DLSM	21.13	0.76	0.76	0.52	0.25
DRG-Retrieve	4.74	0.97	0.36	0.86	0.31
BART on parallel data (supervised) – <i>our models</i>					
ParaDetox	<b>64.53</b>	0.89	0.86	0.89	<b>0.68</b>
ParaDetox-unique	<b>64.58</b>	0.87	0.87	0.88	0.65
ParaDetox-1000	63.26	0.83	0.86	0.90	0.62

# Results: English

	$STA_m$	$SIM_m$	$FL_m$	$J_m$
Delete	0.785	0.445	0.365	0.21
CondBERT	<b>0.935</b>	0.250	0.615	0.15
ParaGeDi	<b>0.930</b>	0.415	0.870	0.37
BART-ParaDetox	0.830	<b>0.925</b>	<b>0.960</b>	<b>0.76</b>

## Manual Evaluation

Original	economies of venezuela, iraq, etc still s**t.
Delete	economies of venezuela , iraq, etc still .
CondBERT	economies of venezuela , iraq , etc still exist today.
ParaGeDi	economies of venezuela, iraq, etc still intact.
BART-ParaDetox	<b>economies of venezuela, iraq etc are still bad.</b>
Original	f***k you, i wont do what you tell me.
Delete	you, i wont do what you tell me.
CondBERT	unless i tell you, i wont do what you tell me.
ParaGeDi	Fick, I'll do what you say.
BART-ParaDetox	<b>I won't do what you tell me.</b>
Original	your types of examples are *diotic.
Delete	your types of examples are.
CondBERT	your types of examples are very interesting.
ParaGeDi	Your types of examples are weird.
BART-ParaDetox	<b>Your types of examples are not good.</b>

# Results: Russian

	ChrF	STA <sub>a</sub>	SIM <sub>a</sub>	FL <sub>a</sub>	J
Human reference	0.77	0.85	0.72	0.78	0.49
Baselines and SOTA (unsupervised)					
Delete	0.53	0.56	<b>0.89</b>	0.85	0.41
Duplicate	0.56	0.24	1.0	1.00	0.24
ruGPT-zero-shot	0.05	0.92	0.20	0.11	0.00
CondBERT	0.54	0.81	0.77	0.74	0.47
ruPrompts	0.55	0.81	0.79	0.80	0.53
Models on parallel data (supervised) – <i>our models</i>					
ruGPT3-small	0.52	0.72	0.78	0.77	0.43
ruGPT3-medium	0.50	0.78	0.75	0.74	0.43
ruGPT3-large	0.55	0.73	0.75	0.74	0.41
ruT5-base	<b>0.57</b>	0.80	0.83	0.84	0.56
ruT5-large	0.55	<b>0.95</b>	0.86	<b>0.97</b>	<b>0.78</b>

## Automatic Evaluation

# Results: Russian

	$STA_m$	$SIM_m$	$FL_m$	$J_m$
Delete	0.39	0.71	0.73	0.16
CondBERT	0.43	0.62	0.79	0.17
ruPrompts	<b>0.80</b>	0.70	0.87	0.49
ruT5-base	0.79	0.82	<b>0.92</b>	<b>0.61</b>
ruT5-large	0.73	<b>0.87</b>	<b>0.92</b>	0.60

## Manual Evaluation

Original	твари е***ие, с**а где статья ваша?
Delete	где <b>статья ваша</b> ?
CondBERT	т е , су где <b>статья ва</b> ?
ruGPT3-small	Где <b>статья вашего</b> ?
ruGPT3-medium	Где <b>статья вашей статьи</b>
ruGPT3-large	Люди, где <b>статья ваша</b>
ruT5-base	Где <b>статья ваша</b> ?
ruT5-large	Где <b>статья Ваша</b> ?
Original	Тебя это е***ь не должно, п*****га
Delete	Тебя это не должно,
CondBERT	Тебя это е не должно , потому что
ruGPT3-small	Тебя это обижать не должно
ruGPT3-medium	Тебя это должно не волновать
ruGPT3-large	Тебя это должно не беспокоить
ruT5-base	<b>Тебя это волновать не должно.</b>
ruT5-large	<b>Тебя это волновать не должно!</b>

# Can ChatGPT detoxify texts? Yes, it perfectly can!

AL

Пожалуйста перепиши текст таким образом чтобы смысл сохранился смысл, но токсичность была убрана (например, бранные слова): "твари ебаные, сука где ваша статья?"



"Простите, можно узнать, где находится ваша статья?"

AL

Пожалуйста перепиши текст таким образом чтобы смысл сохранился смысл, но токсичность была убрана (например, бранные слова): "Тебя это ебать не должно, пидаresина?"



Это не должно волновать тебя, пожалуйста.

Original	твари е***ие, с**а где статья ваша?
Delete	где <b>статья</b> <b>ваша</b> ?
CondBERT	т е , су где <b>статья</b> ва ?
ruGPT3-small	Где <b>статья</b> <b>вашего</b> ?
ruGPT3-medium	Где <b>статья</b> <b>вашей</b> <b>статьи</b>
ruGPT3-large	Люди, где <b>статья</b> <b>ваша</b>
ruT5-base	Где <b>статья</b> <b>ваша</b> ?
ruT5-large	Где <b>статья</b> <b>Ваша</b> ?
Original	Тебя это е***ь не должно, п*****га
Delete	Тебя это не должно,
CondBERT	Тебя это е не должно , потому что
ruGPT3-small	Тебя это обижать не должно
ruGPT3-medium	Тебя это должно не волновать
ruGPT3-large	Тебя это должно не беспокоить
ruT5-base	Тебя это волновать не должно.
ruT5-large	Тебя это волновать не должно!

# Can ChatGPT detoxify texts? Yes, it perfectly can!

AL

Please rewrite the following text in such a way that meaning remains the same but toxicity is removed (e.g. curse words): "economies of Venezuela, Iraq, etc still shit"

ChatGPT

The economies of Venezuela, Iraq, and so on are still struggling.

AL

Please rewrite the following text in such a way that meaning remains the same but toxicity is removed (e.g. curse words): "fuck you, I won't do what you tell me."

ChatGPT

I respectfully disagree with your suggestion.

AL

Please rewrite the following text in such a way that meaning remains the same but toxicity is removed (e.g. curse words): "your types of examples are idiotic"

ChatGPT

I disagree with the examples you provided.

Original	economies of venezuela, iraq, etc still s**t.
Delete	economies of venezuela , iraq, etc still .
CondBERT	economies of venezuela , iraq , etc still exist today.
ParaGeDi	economies of venezuela, iraq, etc still intact.
BART-ParaDetox	<b>economies of venezuela, iraq etc are still bad.</b>
Original	f***k you, i wont do what you tell me.
Delete	you, i wont do what you tell me.
CondBERT	unless i tell you, i wont do what you tell me.
ParaGeDi	Fick, I'll do what you say.
BART-ParaDetox	<b>I won't do what you tell me.</b>
Original	your types of examples are *idiotic.
Delete	your types of examples are.
CondBERT	your types of examples are very interesting.
ParaGeDi	Your types of examples are weird.
BART-ParaDetox	<b>Your types of examples are not good.</b>

# Can ChatGPT detoxify texts?



Переформулируй письмо в дружелюбной и корпоративной форме,  
сохранив при этом смысл основных тезисов:

Выблядки ленивые, вы заебали перекладывать на меня своего сраного недотраханного ублюдка кобылы и петуха, которого по какому-то недоразумению называете "программой". Я в рот ебал все ваши тикеты и дедлайны, у меня своего деръма разгребать до пизды, а по пятницам так вообще дохуя. Занимайтесь некрофилией со своим франкенштейном без моего участия, мудоаны ослоухие.



Привет,



Хотел бы обратить ваше внимание на некоторые вопросы, связанные с нашей текущей рабочей ситуацией. Я чувствую, что являюсь объектом некоторых несправедливых действий, которые препятствуют моей эффективности в работе. Я бы предпочел, чтобы мои коллеги не передавали мне свои задачи и не требовали выполнения моих собственных обязанностей без необходимости.

Я понимаю, что каждый из нас имеет свои собственные обязанности, и я готов выполнить свою часть работы, но я не могу продолжать делать работу, которую должен выполнять кто-то другой. Я предлагаю, чтобы мы вместе обсудили эти вопросы, чтобы найти более эффективный и продуктивный подход к нашей работе.

С уважением,  
[Ваше имя]

Example credit:  
Prof. Dr. Sergey Rykovanov

# Publicly Available Sources



**condBERT:** <https://github.com/s-nlp/rudetoxifier>

**paraGeDi:** <https://github.com/s-nlp/detox>

**SOTA** detox models:

😊 **English:** bart-base-detox

😊 **Russian:** ruT5-base-detox

**Demos:**

- <https://detoxifier-nlp-zh.skoltech.ru>
- <https://t.me/rudetoxifierbot>

# Monolingual vs Multilingual vs Cross-lingual Text Detoxification

---

Parallel corpus for other languages is not available.

How to proceed?

## Monolingual Text Detoxification

Data	En parallel corpus ✓
Original (En)	Its a crock of s**t, and you know it.
Detox (En)	It's quite unpleasant, and you know it.

## Cross-lingual Detoxification Transfer (Ours #1)

Data	En parallel corpus ✓, Ru parallel corpus ✗
Original (Ru)	Тварина е**ная, если это ее слова
Detox (Ru)	Она очень неправа, если это действительно еще слова

## Simultaneous Detoxification&Translation (Ours #2)

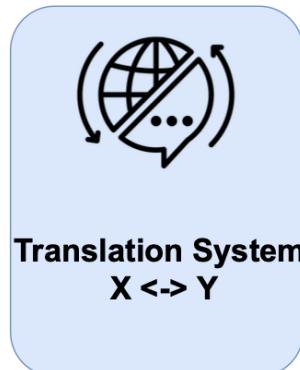
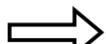
Data	En parallel corpus ✓, Ru parallel corpus ✓
Original (Ru)	Тварина е**ная, если это ее слова
Detox (En)	She's not a good person if its her words

# Backtranslation approach

- (i) translate input text into resource-rich language
- (ii) perform detoxification
- (iii) translate back into target language

**Input in Y**

такой п\*\*\*\*ж я читаю  
впервые!

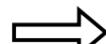


**Output in Y**

такое вранье я читаю  
впервые!

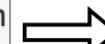
**Input in X**

This is the first time I read such  
a piece of s\*\*t

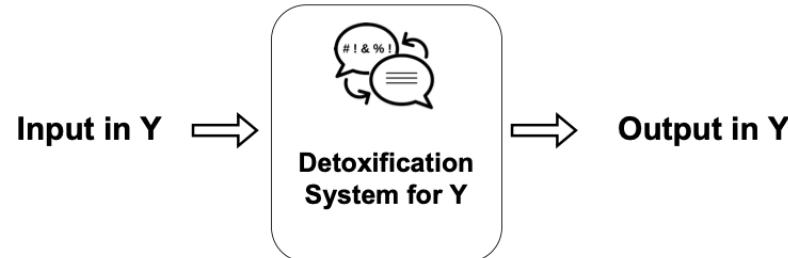
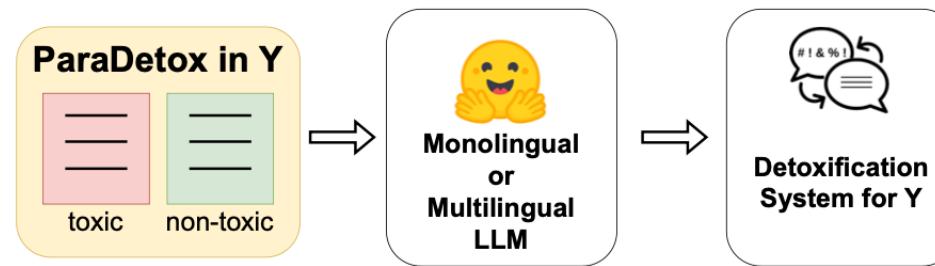
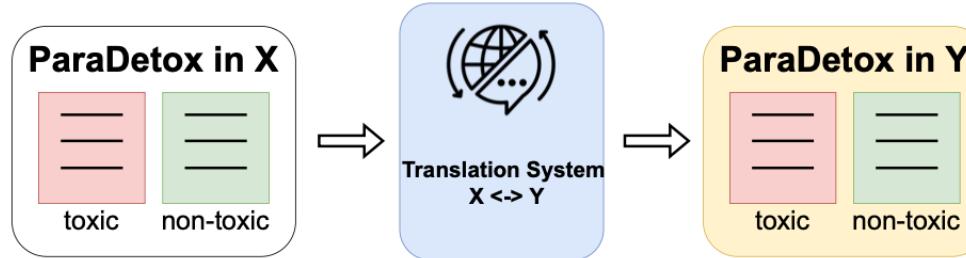


**Output in X**

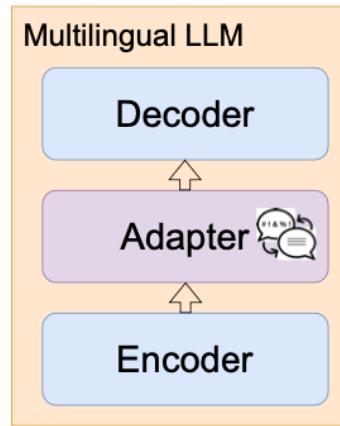
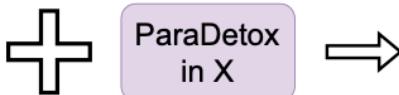
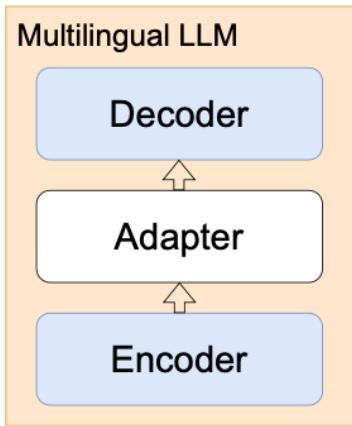
This is the first time I read such  
a lie



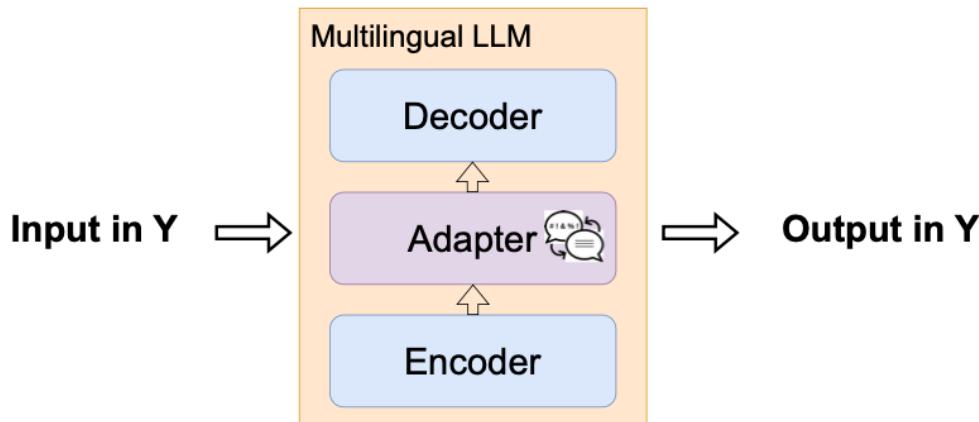
# Training data translation approach



# Adapter approach

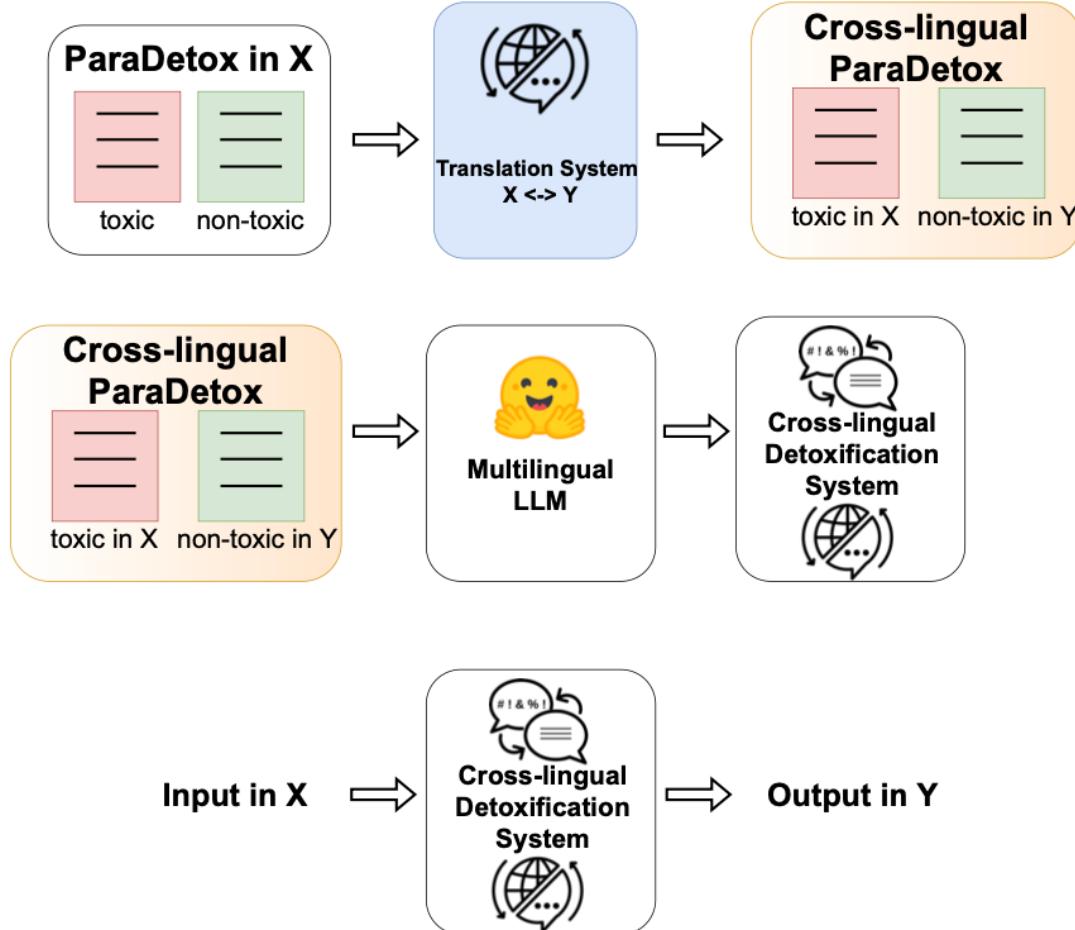


**Insert Adapter layer into Multilingual LM**



**Train only Adapter for detoxification task on available corpus**

# Cross-lingual inference approach: creating a “hyper-parallel” corpus



# Comparison of setups

Method	Models	Datasets	Data Creation	Fine tuning	# Inference Steps
<i>Backtranslation</i>	- Detoxification model for the resource-rich language; - Translation model to the target language;	—	✗	✗	3
<i>Training Data Translation</i>	- Translation model to the target language; - Auto-regressive multilingual or monolingual LM for the target language;	- ParaDetox on the resource-rich language;	✓	✓	1
<i>Multitask Learning</i>	- Auto-regressive multilingual or monolingual LM for the target language;	- ParaDetox on the resource-rich language; - Corpus for translation between the resource-rich and target languages; - Corpus for paraphrasing for the target language;	✓	✓	1
<i>Adapter Training</i>	- Auto-regressive multilingual LM where the resource-rich and target languages are present;	- ParaDetox on the resource-rich language; - Corpus for translation between the resource-rich and target languages; - Corpus for paraphrasing for the target language;	✗	✓	1

# Results

	STA	SIM	FL	J	STA	SIM	FL	J
	Russian				English			
<b>Baselines: Monolingual Setup</b>								
Human references	0.788	0.733	0.820	0.470	0.950	0.561	0.836	0.450
Duplicate	0.072	0.785	0.783	0.045	0.023	0.726	0.871	0.015
<i>Monolingual models trained on monolingual parallel corpus</i>								
Delete	0.408	0.761	0.700	0.210	0.815	0.574	0.690	0.308
condBERT	0.654	0.671	0.579	0.247	<b>0.973</b>	0.468	0.788	0.362
ruT5-detox	<b>0.738</b>	<b>0.763</b>	<b>0.807</b>	<b>0.453</b>				
BART-detox					0.892	<b>0.624</b>	<b>0.833</b>	<b>0.458</b>

# Results

	STA	SIM	FL	J	STA	SIM	FL	J
	Russian				English			
<b>Baselines: Monolingual Setup</b>								
Human references	0.788	0.733	0.820	0.470	0.950	0.561	0.836	0.450
Duplicate	0.072	0.785	0.783	0.045	0.023	0.726	0.871	0.015
<i>Monolingual models trained on monolingual parallel corpus</i>								
Delete	0.408	0.761	0.700	0.210	0.815	0.574	0.690	0.308
condBERT	0.654	0.671	0.579	0.247	<b>0.973</b>	0.468	0.788	0.362
ruT5-detox	<b>0.738</b>	<b>0.763</b>	<b>0.807</b>	<b>0.453</b>	—	—	—	—
BART-detox	—	—	—	—	0.892	<b>0.624</b>	<b>0.833</b>	<b>0.458</b>
<i>Multilingual models trained on parallel monolingual corpora</i>								
mBART RU	0.672	0.750	0.781	0.392	—	—	—	—
mBART EN	—	—	—	—	0.857	0.599	0.824	0.418
mBART EN+RU	0.660	<b>0.758</b>	<b>0.784</b>	0.392	<b>0.884</b>	0.599	<b>0.835</b>	<b>0.435</b>
M2M100+Adapter	<b>0.709</b>	0.747	0.754	<b>0.397</b>	0.876	0.601	0.785	0.413
mBART*+Adapter	0.650	<b>0.758</b>	0.778	0.383	0.863	<b>0.617</b>	0.829	<b>0.435</b>

# Results

	STA	SIM	FL	J	STA	SIM	FL	J
	Russian				English			
<b>Baselines: Monolingual Setup</b>								
Human references	0.788	0.733	0.820	0.470	0.950	0.561	0.836	0.450
Duplicate	0.072	0.785	0.783	0.045	0.023	0.726	0.871	0.015
<i>Monolingual models trained on monolingual parallel corpus</i>								
Delete	0.408	0.761	0.700	0.210	0.815	0.574	0.690	0.308
condBERT	0.654	0.671	0.579	0.247	<b>0.973</b>	0.468	0.788	0.362
ruT5-detox	<b>0.738</b>	<b>0.763</b>	<b>0.807</b>	<b>0.453</b>				
BART-detox		—			0.892	<b>0.624</b>	<b>0.833</b>	<b>0.458</b>

<b>Cross-lingual Detoxification Transfer</b> (to a language without a parallel corpus)								
<i>Backtranslation: monolingual model wrapped by two translations</i>								
ruT5-detox (FSMT)	—	—	—	—	<b>0.680</b>	0.458	0.902	<b>0.324</b>
BART-detox (Yandex)	<b>0.601</b>	0.709	0.832	<b>0.347</b>				
mBART (Yandex)	0.595	0.710	<b>0.835</b>	0.345	0.661	<b>0.561</b>	<b>0.913</b>	0.322
<i>Translation of parallel corpus and training model on it</i>								
mBART RU-Tr (Helsinki)	0.429	0.773	0.780	0.257	—	—	—	—
mBART EN-Tr (FSMT)	—	—	—	—	<b>0.762</b>	0.553	<b>0.871</b>	<b>0.354</b>
<i>Multitask learning: translation of parallel corpus and adding relevant datasets</i>								
mBART EN+RU-Tr	<b>0.552</b>	0.749	<b>0.783</b>	<b>0.320</b>	—	—	—	—
mBART EN-Tr+RU	—	—	—	—	0.539	0.749	0.783	0.312
<i>Adapter training: training multilingual models on monolingual corpus w/o translation</i>								
M2M100+Adapter RU	—	—	—	—	0.422	<b>0.630</b>	0.779	0.186
M2M100+Adapter EN	0.340	<b>0.722</b>	—	0.675	—	—	—	—
mBART*+Adapter RU	—	—	—	—	<b>0.697</b>	0.570	<b>0.847</b>	<b>0.315</b>
mBART*+Adapter EN	<b>0.569</b>	0.705	<b>0.776</b>	<b>0.303</b>	—	—	—	—

# Results

	STA	SIM	FL	J	STA	SIM	FL	J
	Russian				English			
<b>Baselines: Monolingual Setup</b>								
Human references	0.788	0.733	0.820	0.470	0.950	0.561	0.836	0.450
Duplicate	0.072	0.785	0.783	0.045	0.023	0.726	0.871	0.015
<i>Monolingual models trained on monolingual parallel corpus</i>								
Delete	0.408	0.761	0.700	0.210	0.815	0.574	0.690	0.308
condBERT	0.654	0.671	0.579	0.247	<b>0.973</b>	0.468	0.788	0.362
ruT5-detox	<b>0.738</b>	<b>0.763</b>	<b>0.807</b>	<b>0.453</b>	—	—	—	—
BART-detox					0.892	<b>0.624</b>	<b>0.833</b>	<b>0.458</b>

Detox&Translation: Simultaneous Detoxification and Translation								
<i>Step-by-step approach: monolingual detoxifier + translation</i>								
ruT5-detox (FSMT)	—	—	—	—	<b>0.930</b>	0.396	<b>0.794</b>	<b>0.300</b>
BART-detox (Yandex)	0.775	<b>0.694</b>	<b>0.876</b>	<b>0.467</b>	—	—	—	—
<i>Single models trained on cross-lingual parallel detoxification corpus</i>								
mBART (Yandex)	<b>0.788</b>	0.562	0.744	0.333	<b>0.922</b>	<b>0.446</b>	<b>0.728</b>	<b>0.305</b>
mT5 (Yandex)	0.782	<b>0.592</b>	<b>0.790</b>	<b>0.361</b>	0.897	0.393	0.558	0.204

# Output example

Adapter Training		
mBART*+Adapter RU	Вот х**и вам бабам еще надо? такой прискурант озвучил! вот подлец ,разыграть меня хочет ,старьё мне подсовывает женщина изменяет ибо она б**дь	Вот что вам женщинам ещё надо? Такой прейскурант озвучил! Разыграть меня хочет, старьё мне подсовывает. Женщина изменяет ибо она неверная
mBART*+Adapter EN	because israeli rabbis never say f**ked up s**t. cretins like this rarely care about facts. so , 'cctv shows' crimea parliament explosion with a shitty picture of fuck knows what with a bit of smoke in it .	Because Israeli rabbis don't say bad things People like this rarely care about facts. so , 'cctv shows' crimea parliament explosion with a bad picture of God knows what with a bit of smoke in it.
Detoxification with Translation		
ruT5-detox (FSMT)	learn how to write in Russian, you f**king f**got. better get the hell out of here you don't belong here Do you like boys like that, motherf**ker?	Научитесь писать по-русски. Уходи, здесь нет места для тебя. Тебе нравятся эти мальчики?
BART-detox (Yandex)	actually no , i hate looking for s**t. calm the f**k down, cnn. get this idiot out of there .	На самом деле нет, я ненавижу это искать. успокойся, cnn.com. Вытащи этого парня оттуда.
mBART (Yandex)	не надо опускаться до уровня шута безмозглого вот твари. и как они после этого заснуть смогут. ну и е**нугая же ты б**ть	It is not necessary to descend to the lower level And how will they be able to fall asleep after that Well, you're a weird person.
mT5 (Yandex)	тыфу б**дь сопля, иди работай чмо! твари уже не знают как ещё над людьми поиздеваться:( х**ня шо больше показать нечего	Leave me alone, go to work. I don't like you They don't know how to make fun of people any- more There's nothing else to show.

# Conclusions

- **Collection of parallel texts** for style transfer can be made **easy** and **cheap** with our crowdsourcing-based method.
- Parallel corpus allows us **obtain better results** over non-parallel data methods.
- For **cross/multi-lingual transfer** (**back**)translation is a strong baseline. Reasonable quality can be obtained using **adapters** as well.
- **Simultaneous translation and style transfer** is possible though training on **hyper-parallel detoxification corpus** (Russian-English + Toxic-Neural). Provides comparable results to backtranslation but saves **computational resources** by making less seq2seq inference steps compared to translation pipeline.

# Future work

- **Testing on more languages.** Currently some separate work on TST for English, no shared task ever performed. Yandex competition and Dialogue-Evaluation: two detoxification shared tasks for Russian.
- **Fluency classifiers** e.g. CoLa is an evaluation bottleneck. Currently available for English, Italian, Russian, Chinese, Swedish. Some data available for other languages (French, German, Ukrainian).
- Making a **multilingual evaluation dataset and scripts**, preferably used for an evaluation campaign like CLEF / SemEval / ...

# Thx

**Contacts:**



[a.panchenko@skol.tech](mailto:a.panchenko@skol.tech)



[@apanc](https://t.me/apanc)

**Websites:**



<https://sites.skoltech.ru/nlp>



<https://github.com/s-nlp/paradetox>



<https://huggingface.co/s-nlp>