Exploring Methods for Cross-lingual Text Style Transfer: The Case of Text Detoxification

Anonymous ACL submission

Abstract

Text detoxification is the task of transferring the style of text from toxic to neutral. While there are approaches yielding promising results in monolingual setup, e.g., (Dale et al., 2021; Hallinan et al., 2022), cross-lingual transfer for this task remains a challenging open problem (Moskovskiy et al., 2022). In this work, we present a large-scale study of strategies for cross-lingual text detoxification – given a parallel detoxification corpus for one language; the goal is to transfer detoxification ability to another language for which we do not have such a corpus. Moreover, we are the first to explore a new task where text translation and detoxification are performed simultaneously, providing several strong baselines for this task. Finally, we introduce new automatic detoxification evaluation metrics with higher correlations with human judgments than previous benchmarks. We assess the most promising approaches also with manual markup, determining the answer for the best strategy to transfer the knowledge of text detoxification between languages.

1 Introduction

The original monolingual task of text detoxification can be considered as text style transfer (TST), where the goal is to build a function that, given a source style s^{src}, a destination style s^{dst}, and an input text t^{src} to produce an output text t^{dst} such that: (i) the style is indeed changed (in case of detoxification from toxic into neutral); (ii) the content is saved as much as possible; (iii) the newly generated text is fluent.

The task of detoxification was already addressed with several approaches. Firstly, several unsupervised methods based on masked language modelling (Tran et al., 2020; Dale et al., 2021) and disentangled representations for style and content (John et al., 2019; dos Santos et al., 2018) were explored. More recently, Logacheva et al. (2022b) showed the superiority of supervised *seq2seq* models for detoxification trained on a parallel corpus of crowdsourced toxic ↔ neutral sentence pairs. Afterwards, there were experiments in multilingual detoxification. However, cross-lingual transfer between languages with multilingual *seq2seq* models was shown to be a challenging task (Moskovskiy et al., 2022). In this work, we aim to fill this gap and present an extensive overview of different approaches for cross-lingual text detoxification methods (tested in English and Russian), showing that promising results can be obtained in contrast to prior findings. Besides, we explore combining of two *seq2seq* tasks/models in a single one to achieve computational gains (i.e., avoid the need to store and perform inference with several models). Namely, we conduct simultaneous translation and style transfer experiments, comparing them to a step-by-step pipeline.

Monolingual Text Detoxification	
Data	En parallel corpus ✔
Original (En) Detox (En)	Its a crock of s**t, and you know it. It's quite unpleasant, and you know it.
Cross-lingual Detoxification Transfer (Ours #1)	
Data	En parallel corpus ✔, Ru parallel corpus ✘
Original (Ru) Detox (Ru)	Тварина е**ная, если это ее слова Она очень неправа, если это действительно еще слова
Simultaneous Detoxification&Translation (Ours #2)	
Data	En parallel corpus ✔, Ru parallel corpus ✔
Original (Ru) Detox (En)	Тварина е**ная, если это ее слова She's not a good person if its her words

Table 1: **Two new text detoxification setups** explored in this work compared to the monolingual setup.

The contributions of this work are as follows:

- We present a study on *cross-lingual detoxification transfer* and present new methods based on adapters and multi-task learning,

- We are the first to explore the task of *simultaneous detoxification and translation* and test several baseline approaches to solve it,

- We present a set of updated *metrics for automatic evaluation* of detoxification improving correlations with human judgements.[1]

2 Related Work

Text Detoxification Datasets Previously, several datasets for different languages were released for toxic and hate speech detection. For instance, there exist several versions of Jigsaw datasets – monolingual (Jigsaw, 2018) for English and multilingual (Jigsaw, 2020) covering 6 languages. In addition, there are corpora specifically for Russian (Semiletov, 2020), Korean (Moon et al., 2020), French (Vanetik and Mimoun, 2022) languages, *inter alia*. These are non-parallel classification datasets. In previous work on detoxification methods, such kind of datasets were used to develop and test unsupervised text style transfer approaches (Wu et al., 2019; Tran et al., 2020; Dale et al., 2021; Hallinan et al., 2022).

However, lately a parallel dataset *ParaDetox* for training supervised text detoxification models for English was released (Logacheva et al., 2022b) similar to previously parallel TST datasets for formality (Rao and Tetreault, 2018; Briakou et al., 2021). Pairs of toxic-neutral sentences were collected with a pipeline based on three crowdsourcing tasks. The first task is the main paraphrasing task. Then, the next two tasks – content preservation check and toxicity classification – are used to verify a paraphrase. Using this crowdsourcing methodology, a Russian parallel text detoxification dataset was also collected (Dementieva et al., 2022). We base our cross-lingual text detoxification experiments on these comparably collected and comparable in volume data (cf. Table 2).

	Train	Dev	Test	Total
English (Logacheva et al., 2022b)	18 777	988	671	20 436
Russian (Dementieva et al., 2022)	5 058	1 000	1 000	7 058

Table 2: **Parallel text detoxification datasets** used in our cross-lingual detoxification experiments.

Text Detoxification Models Addressing text detoxification task as *seq2seq* task based on a parallel corpus was shown to be more successful than the application of unsupervised methods by Logacheva et al. (2022b). For English methods, the fine-tuned BART model (Lewis et al., 2020) on English ParaDetox significantly outperformed all the baselines and other *seq2seq* models in both automatic and manual evaluations. For Russian in (Dementieva et al., 2022), there was released ruT5 model (Raffel et al., 2020) fined-tuned on Russian ParaDetox. These SOTA monolingual models for English[2] and Russian[3] are publicly available.

Multilingual Models Together with pre-trained monolingual language models (LM), there is a trend of releasing multilingual models covering more and more languages. For instance, the NLLB model (Costa-jussà et al., 2022) is pre-trained for 200 languages. However, large multilingual models can have many parameters (NLLB has 54.5B parameters), simultaneously requiring a vast amount of GPU memory to work with it.

As the SOTA detoxification models were fine-tuned versions of T5 and BART, we experiment in this work with multilingual versions of them – **mT5** (Xue et al., 2021) and **mBART** (Tang et al., 2020). The mT5 model covers 101 languages and has several versions. The mBART model has several implementations and several versions as well. We use mBART-50, which covers 50 languages. Also, we use in our experiments the **M2M100** model (Fan et al., 2021) that was trained for translation between 100 languages. All these models have less than 1B parameters (in *large* versions).

Cross-lingual Knowledge Transfer A common case is when data for a specific task is available for English but none for the target language. In this situation, techniques for knowledge transfer between languages are applied.

One of the approaches usually used to address the lack of training data is the translation approach. It was already tested for offensive language classification (El-Alami et al., 2022; Wadud et al., 2023). The idea is to translate the training data in the available language into the target language and train the corresponding model based on the new translated dataset.

The methods for zero-shot and few-shot text style transfer were already explored. In (Krishna et al., 2022), the operation between style and language embeddings is used to transfer style knowl-

[1] Code and datasets will be released openly.

[2] https://huggingface.co/s-nlp/bart-base-detox

[3] https://huggingface.co/s-nlp/ruT5-base-detox

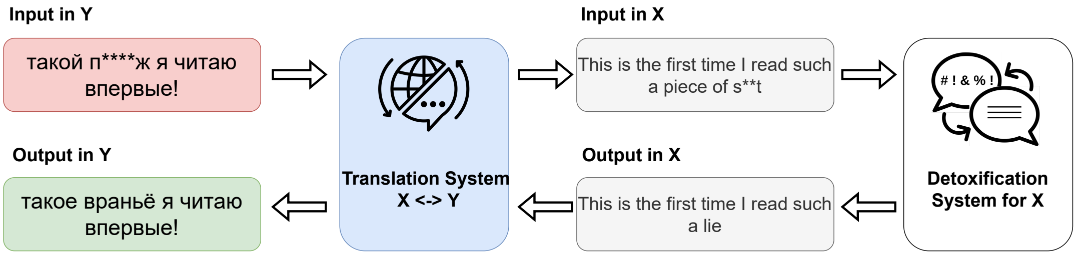

Figure 1: **Backtranslation approach**: (i) translate input text into resource-rich language; (ii) perform detoxification; (iii) translate back into target language.

edge to a new language. The authors in (Lai et al., 2022b) use adapter layers to incorporate the knowledge about the target language into a TST model.

For text detoxification, only in (Moskovskiy et al., 2022) cross-lingual setup was explored through the translation of inputs and outputs of a monolingual system. It has been shown that detoxification trained for English using a multilingual Transformer is not working for Russian (and vice versa). In this work, we present several approaches to cross-lingual detoxification, which, in contrast, yield promising results.

Simultaneous Text Generation&Translation The simultaneous translation and text generation was already introduced for text summarization. Several datasets with a wide variety of languages were created (Perez-Beltrachini and Lapata, 2021; Hasan et al., 2021). The main approaches to tackle this task – either to perform step-by-step text generation and translation or train a supervised model on a parallel corpus. To the best of our knowledge, there were no such experiments in the domain of text detoxification. This work provides the first experiments to address this gap.

3 Cross-lingual Detoxification Transfer

In this section, we consider the setup when a parallel detoxification corpus is available for a resource-rich language (e.g., English), but we need to perform detoxification for another language such corpus is unavailable. We test several approaches that differ by the amount of data and computational sources listed below.

3.1 Backtranslation

One of the baseline approaches is translating input sentences into the language for which a detoxification model is available. For instance, we first train a detoxification model on available English ParaDetox. Then, if we have an input sentence in another language, we translate it into English, perform detoxification, and translate it back into Russian (Figure 1). Thus, for this approach, we require two models (one model for translation and one for detoxification) and three inferences (one for translation from the target language into the available language, text detoxification, and translation back into the target language).

In previous work (Moskovskiy et al., 2022), **Google Translate API** and **FSMT** (Ng et al., 2019) models were used to make translations. In this work, we extend these experiments with two additional models for translation:

- **Helsinki OPUS-MT** (Tiedemann and Thottingal, 2020) – Transformer-based model trained specifically for English-Russian translation.[4]
- **Yandex** Translate API available from Yandex company and considered high/top quality for the Russian-English pair.[5]

We test the backtranslation approach with two types of models: (i) SOTA models for corresponding monolingual detoxification; (ii) multilingual LM.

3.2 Training Data Translation

Another way of how translation can be used is the translation of available training data. If we have available training data in one language, we can fully translate it into another and use it to train a separate detoxification model for this language (Figure 2). For translation, we use the same models described in the previous section.

As detoxification corpus is available for the target language in this setup, we can fine-tune either multilingual LM where this language is present or

[4] https://huggingface.co/Helsinki-NLP/opus-mt-ru-en
[5] https://tech.yandex.com/translate

monolingual LM if it is separately pre-trained for the required language. Compared to the previous approach, this method requires a fine-tuning step that implies additional computational resources.

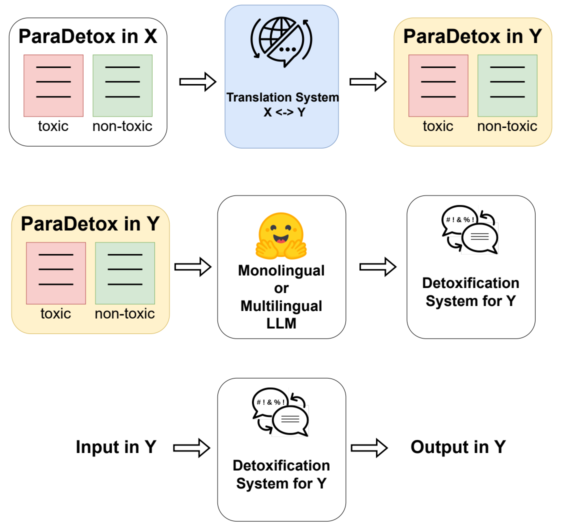

Figure 2: **Training Data Translation approach**: (i) translate available dataset into the target language; (ii) train detoxification model for the target language.

3.3 Multitask Learning

Extending the idea of using translated ParaDetox, we can add additional datasets that might help improve model performance.

We suggest multitasking training for cross-lingual detoxification transfer. We take a multilingual LM where resource-rich and target languages are available. Then, for the training, we perform multitask procedure which is based on the following tasks: (i) translation between the resource-rich language and target language; (ii) paraphrasing for the target language; (iii) detoxification for the resource-rich language for which original ParaDetox is available; (iv) detoxification for the target language based on translated data.

Even if the LM is already multilingual, we suggest that the translation task data help strengthen the bond between languages. As the detoxification task can be seen as a paraphrasing task as well, the paraphrasing data for the target language can add knowledge to the model of how paraphrasing works for this language. Then, the model is basically trained for the detoxification task on the available data.

For paraphrasing corpus, we use **Opusparcus** corpus (Creutz, 2018). For translation, we use corresponding `en-ru` parts of **Open Subtitles** (Lison and Tiedemann, 2016), **Tatoeba** (Tiedemann, 2020), and **news_commentary**[6] corpora.

3.4 Adapter Training

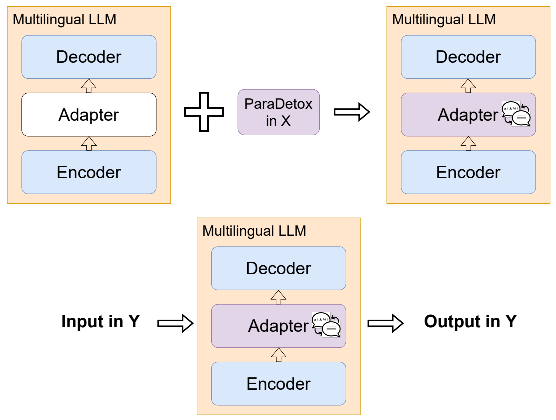

Figure 3: **Adapter approach**: (i) insert Adapter layer into Multilingual LM; (ii) train only Adapter for detoxification task on the available corpus.

To eliminate the translation step, we present a new approach based on the Adapter Layer idea (Houlsby et al., 2019). The usual pipeline of *seq2seq* generation process is:

$$y = \text{Decoder}(\text{Encoder}(x)) \quad (1)$$

We add an additional Adapter layer in the model:

$$y = \text{Decoder}(\text{Adapter}(\text{Encoder}(x))), \quad (2)$$

where $Adapter = Linear(ReLU(Linear(x)))$ and gets as input the output embeddings from encoder.

Any multilingual pre-trained model can be taken for a base *seq2seq* model. Then, we integrate the Adapter layer between the encoder and decoder blocks. For the training procedure, we train the model on a monolingual ParaDetox corpus available. However, we do not update all the weights of all model blocks, only the Adapter. As a result, we force the Adapter layer to learn the information about detoxification while the rest of the blocks save the knowledge about multiple languages. We can now input the text in the target language during inference and obtain the corresponding detoxified output (Figure 3). Compared to previous approaches, the Adapter training requires only one model fine-tuning procedure and one inference step. While in (Lai et al., 2022b)

[6]https://huggingface.co/datasets/news_commentary

there were used several Adapter layers pre-trained specifically for the language, we propose to use only one layer between the encoder and decoder of multilingual LM that will incorporate the knowledge about the task.

For this approach, we experiment with the **M2M100** and **mBART-50** models. While the M2M100 model is already trained for the translation task, this version of mBART is pre-trained only on the denoising task. Thus, we additionally pre-train this model on paraphrasing and translation corpora used for the Multitask approach. During the training and inference with the mBART model, we explicitly identify which language the input and output are given or expected with special tokens.

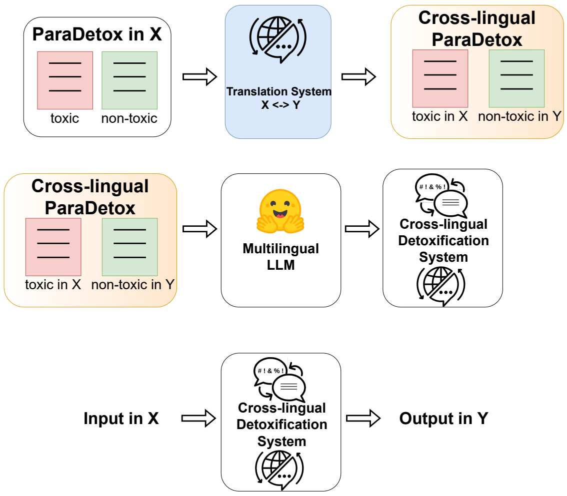

Figure 4: **Simultaneous Detox&Translate** approach is based on synthetic cross-lingual parallel corpus.

4 Detox&Translation

The setup of simultaneous detoxification and translation occurs when the toxic and non-toxic parts of the training parallel dataset are in different languages. For instance, a toxic sentence in a pair is in English, while its non-toxic paraphrase is in Russian.

The baseline approach to address text detoxification from one language to another can be to perform step-by-step detoxification and translation. However, that will be two inference procedures, each potentially with a computationally heavy seq2seq model. To save resources for one inference, in this section, we explore the models that can perform detoxification and translation in one step.

While for cross-lingual text summarization, parallel datasets were obtained, there are no such data for text detoxification. The proposed approach is creating a synthetic cross-lingual detoxification dataset (Figure 4). Then, we train simultaneously model for detoxification as well as for translation. The models described in the section above were also used for the translation step of parallel corpora.

5 Evaluation Setups

There are plenty of work developing systems for text detoxification. Yet, in each work, the comparison between models is made by automatic metrics that are not unified, and their choice may be arbitrary (Ostheimer et al., 2023). There are several recent works that studied the correlation between automatic and manual evaluation for text style transfer tasks – formality (Lai et al., 2022a) and toxicity (Logacheva et al., 2022a). Our work presents a new set of metrics for automatic evaluation for English and Russian languages, confirming our choice with correlations with manual metrics.

For all languages, the automatic evaluation consists of three main parameters:

- *Style transfer accuracy* (**STA**$_a$): percentage of non-toxic outputs identified by a style classifier. In our case, we train for each language corresponding toxicity classifier.
- *Content preservation* (**SIM**$_a$): measurement of the extent to which the content of the original text is preserved.
- *Fluency* (**FL**$_a$): percentage of fluent sentences in the output.

The aforementioned metrics must be properly combined to get one *Joint* metric to rank models. We calculate **J** as following:

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(x_i) \cdot \mathbf{SIM}(x_i) \cdot \mathbf{FL}(x_i), \quad (3)$$

where the scores $\mathbf{STA}(x_i)$, $\mathbf{SIM}(x_i)$, $\mathbf{FL}(x_i) \in \{0, 1\}$ meaning the belonging to the corresponding class.

5.1 Automatic Evaluation for English

Our setup is mostly based on metrics used by (Logacheva et al., 2022b): only the content similarity metric is updated as other metrics obtain high correlations with human judgments.

Style accuracy STA_a metric is calculated with a RoBERTa-based (Liu et al., 2019) style classifier trained on the union of three Jigsaw datasets (Jigsaw, 2018).

Content similarity Previous metric SIM_a^{old} is estimated as cosine similarity between the embeddings of the original text and the output computed with the model of (Wieting et al., 2019). This model is trained on paraphrase pairs extracted from ParaNMT (Wieting and Gimpel, 2018) corpus.

Our updated metric SIM_a is calculated as BLEURT score (Sellam et al., 2020). In (Babakov et al., 2022), a large investigation on similarity metrics for paraphrasing and style transfer tasks. The results showed that the BLEURT metric has the highest correlations with human assessments for text style transfer tasks for the English language.

Fluency FL_a is the percentage of fluent sentences identified by a RoBERTa-based classifier of linguistic acceptability trained on the CoLA dataset (Warstadt et al., 2019).

5.2 Automatic Evaluation for Russian

The set of old and new metrics is listed below (the old setup is based on the official shared task script (Dementieva et al., 2022)):

Style accuracy Previous metric STA_a^{old} is computed with a RuBERT Conversational classifier (Kuratov and Arkhipov, 2019) fine-tuned on Russian Language Toxic Comments dataset collected from `2ch.hk` and Toxic Russian Comments dataset collected from `ok.ru`.

In our updated metric STA_a, we change the toxicity classifier using the latest, the more robust version presented in (Gusev, 2022).

Content similarity Previous metric SIM_a^{old} is evaluated as a cosine similarity of LaBSE (Feng et al., 2022) sentence embeddings.

In our updated metric, SIM_a is still calculated as cosine similarity, but we use RuBERT Conversational fine-tuned on three additional datasets: Russian Paraphrase Corpus (Gudkov et al., 2020), RuPAWS (Martynov et al., 2022), and content evaluation part from Russian parallel corpus (Dementieva et al., 2022).

Fluency Previous metric FL_a^{old} is measured with a BERT-based classifier (Devlin et al., 2019) trained to distinguish real texts from corrupted ones. The model was trained on Russian texts and their corrupted (random word replacement, word deletion, insertion, word shuffling, etc.) versions.

In our updated metric FL_a to make it symmetric with the English setup, fluency for the Russian language is also evaluated as a RoBERTa-based classifier fine-tuned on the language acceptability dataset for the Russian language RuCoLA (Mikhailov et al., 2022).

	Old metrics	New metrics
STA	0.472	**0.598**
SIM	0.124	**0.244**
FL	-0.011	**0.354**
J	0.106	**0.482**

Table 3: **New vs old evaluation**. Spearman's correlation between automatic vs manual setups for each old and new evaluation parameter based on systems scores for *Russian* language. All numbers denote the statistically significant correlation (p-value ≤ 0.05).

We use the manual assessments available from (Dementieva et al., 2022) to calculate correlations with manual assessments. We have 850 toxic samples in the test set evaluated manually via crowdsourcing by three parameters – toxicity, content, and fluency. We can see in Table 3 the correlations between human assessments and new metrics are higher than for the previous evaluation setup. This confirms the hypothesis for the usage of new automatic metrics. The more detailed results are presented in Appendix D.

To calculate **SIM** metric for **Detox&Translation** task we use the monolingual version of SIM for the target language, comparing the output with the input translated into the target language.

5.3 Manual Evaluation

As the correlation between automatic and manual scores still has room for improvement, we also evaluate selected models manually. We invited three annotators fluent in both languages to markup the corresponding three parameters of evaluation (instructions in Appendix F). A subset of 50 samples from the corresponding test sets were randomly chosen for this evaluation. The interannotator agreement (Krippendorff's α) reaches 0.74 (STA), 0.60 (SIM), and 0.71 (FL).

	STA	SIM	FL	J	STA	SIM	FL	J
	Russian				English			
Baselines: Monolingual Setup (on a language with a parallel corpus)								
Human references	0.788	0.733	0.820	0.470	0.950	0.561	0.836	0.450
Duplicate input	0.072	0.785	0.783	0.045	0.023	0.726	0.871	0.015
Monolingual models trained on monolingual parallel corpus								
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	**0.973**	0.468	0.788	0.362
ruT5-detox	**0.738**	**0.763**	**0.807**	**0.453**			—	
BART-detox			—		0.892	**0.624**	**0.833**	**0.458**
Multilingual models trained on parallel monolingual corpora								
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	**0.758**	**0.784**	0.392	**0.884**	0.599	**0.835**	**0.435**
M2M100+Adapter	**0.709**	0.747	0.754	**0.397**	0.876	0.601	0.785	0.413
mBART*+Adapter	0.650	**0.758**	0.778	0.383	0.863	**0.617**	0.829	**0.435**
Cross-lingual Text Detoxification Transfer (from a language with to a language without a parallel corpus)								
Backtranslation: monolingual model wrapped by two translations								
ruT5-detox (FSMT)			—		**0.680**	0.458	0.902	**0.324**
BART-detox (Yandex)	**0.601**	0.709	0.832	**0.347**			—	
mBART (Yandex)	0.595	0.710	**0.835**	0.345	0.661	**0.561**	**0.913**	0.322
Translation of parallel corpus and training model on it								
mBART RU-Tr (Helsinki)	0.429	0.773	0.780	0.257			—	
mBART EN-Tr (FSMT)			—		**0.762**	0.553	**0.871**	**0.354**
Multitask learning: translation of parallel corpus and adding relevant datasets								
mBART EN+RU-Tr	**0.552**	0.749	**0.783**	**0.320**			—	
mBART EN-Tr+RU			—		0.539	0.749	0.783	0.312
Adapter training: training multilingual models on monolingual corpus w/o translation								
M2M100+Adapter RU			—		0.422	**0.630**	0.779	0.186
M2M100+Adapter EN	0.340	**0.722**	0.675	0.160			—	
mBART*+Adapter RU			—		**0.697**	0.570	**0.847**	**0.315**
mBART*+Adapter EN	**0.569**	0.705	**0.776**	**0.303**			—	
Detox&Translation: Simultaneous Text Detoxification and Translation								
Step-by-step approach: monolingual detoxifier as a pivot + translation from/to the pivot								
ruT5-detox (FSMT)			—		**0.930**	0.396	**0.794**	**0.300**
BART-detox (Yandex)	0.775	**0.694**	**0.876**	**0.467**			—	
End-to-end models trained on cross-lingual parallel detoxification corpus								
mBART (Yandex)	**0.788**	0.562	0.744	0.333	**0.922**	**0.446**	**0.728**	**0.305**
mT5 (Yandex)	0.782	**0.592**	**0.790**	**0.361**	0.897	0.393	0.558	0.204

Table 4: **Automatic evaluation results**. Numbers in **bold** indicate the best results in the sub-sections. Rows in green indicate the best models per tasks. In *(brackets)*, the method of translation used for the approach is indicated. EN or RU denotes training corpus language – original monolingual ParaDetox, while EN-Tr or RU-Tr denotes translated versions of ParaDetox. mBART* states that the version of mBART fine-tuned on paraphrasing and translation data.

6 Results

The **automatic evaluation** results are presented in Table 4. We take test sets provided for both English and Russian datasets for evaluation. Firstly, we report scores of humans and trivial duplication of the input text. Then, we present strong baselines based on local edits – Delete and condBERT (Dale et al., 2021; Dementieva et al., 2021) – and, finally, SOTA *seq2seq* detoxification monolingual models based on T5/BART. Moreover, we report the performance of multilingual models (mBART/M2M100) trained on monolingual parallel corpus separately (RU/EN) or on the joint corpus (RU+EN) to check the credibility of training multilingual models for such a task. The results of the **manual evaluation** are reported in Table 5 comparing only the best models identified with automatic evaluation.

Additional results are available in appendices: Appendix A contains examples of models' outputs; Appendix B contains examples of toxic text translations; Appendix C compares approaches based on the linguistic and computational resources; Appendix E presents a comparison of different translation methods for each approach.

6.1 Cross-lingual Detoxification Transfer

From Table 4, we see that backtranslation approach performed with SOTA monolingual detoxification models yields the best TST scores. This is the only approach that does not require additional model fine-tuning.

	STA	SIM	FL	J
	English			
BART-detox (monolingual)	0.94	0.96	1.00	0.90
Backtr. ruT5-detox (FSMT)	**0.78**	**0.78**	**1.00**	**0.58**
mBART+Adapter RU	0.74	0.70	0.96	0.42
	Russian			
ruT5-detox (monolingual)	0.84	0.96	1.00	0.82
Backtr. BART-detox (Yandex)	0.78	0.56	**1.00**	0.40
mBART+Adapter EN	**0.80**	**0.92**	0.96	**0.72**

Table 5: **Manual evaluation results**. We report the SOTA monolingual models for each language for reference and the best multilingual models (based on Backtranslation and Adapter approaches).

Training Data Translation approach for both languages shows the J score at the level of condBERT baseline. While SIM and FL scores are the same or even higher than monolingual SOTA, the STA scores drop significantly. Some toxic parts in translated sentences can be lost while translating the toxic part of the parallel corpus. It is an advantage for the Backtranslation approach as we want to reduce toxicity only in output, while for training parallel detox corpus, we lose some of the toxicity representation. However, this approach can be used as a baseline for monolingual detoxification (examples of translation outputs in Appendix B). Addition of other tasks training data to a translated ParaDetox yields improvement in the performance for the Russian language in Multitask setup. Paraphrasing samples can enrich toxicity examples that cause the increment in STA.

The adapter for the M2M100 model successfully compresses detoxification knowledge but fails to transfer it to another language. The results are completely different for additionally fine-tuned mBART. This configuration outperforms all unsupervised baselines and the Training Data Translation approach. Still, the weak point for this approach and the STA score, while not all toxicity types, can be easily transferred. However, Adapter Training is the most resource-conserving approach: it does not require additional data creation and has only one inference step. The adapter approach can be the optimal solution for cross-lingual detoxification transfer.

Finally, according to manual evaluations, Backtranslation is the best choice if we want to transfer knowledge to the English language. However, for another low-resource language, the Adapter approach seems to be more beneficial. In the Backtranlsation approach for the Russian language, we have observed a huge loss of content. That can be a case of more toxic expressions in Russian, which are hard to translate precisely into English before detoxification. As a result, we can claim that the Adapter approach is the most efficient and precise way to transfer detoxification knowledge transfer from English to other languages.

6.2 Detox&Translation

At the bottom of Table 4, we report experiments of baseline approaches: detoxification with monolingual detoxification SOTA, then translation into the target language.

We can observe that our proposed approaches for this task for English perform better than the baselines. While for Russian, the results are slightly worse; our models require fewer computational resources during inference. Thus, we can claim that simultaneous style transfer with translation is possible with multilingual LM.

7 Conclusion

We present the first of our knowledge extensive study of cross-lingual text detoxification approaches for English and Russian languages. We also update the automatic evaluation setups for both languages that achieve the highest correlations with human judgments.

The automatic evaluation shows that the Backtranslation approach achieves the highest performance. However, this approach is bounded to the translation system availability and requires three steps during inference. The Training Data Translation approach can be a good baseline for a separate monolingual detoxification system in the target language. On the other hand, the Adapter approach requires only one inference step and performs slightly worse than Backtranslation. The adapter method showed the best manual evaluation scores when transferring from English to Russian.

We present the first study of detoxification and translation in one step. We show that the generation of a synthetic parallel corpus where the toxic part is in one language, and the non-toxic is in another using NMT is effective for this task. Trained on such a corpus, multilingual LMs perform at the level of the backtranslation requiring fewer computations.

8 Limitations

One of the obvious limitations of this work is the usage of only two languages for our experiments – English and Russian. There is a great opportunity for improvement to experiment with more languages and their pairs to transfer knowledge in a cross-lingual style. At the same time, our work provides a result for transferring between languages from different language families.

The possibility of solving the detoxification task, in any case, requires the presence of a corpus of toxicity classification for the language. Firstly, creating a test set and building a classifier for STA evaluation is necessary. Also, having some embedding model for the language is important to calculate the SIM score for evaluation. For FL, in this work, we use classifiers. However, such classifiers can not be present in other languages. As a result, the FL score can be replaced with perplexity.

Also, we test only some subset of multilingual LM. There are new larger models that can potentially solve the task more precisely. However, in our experiments, we use the models that fit in one GPU.

Finally, we do not explore enough which types of toxicities can be successfully transferred between English-Russian pair and which are not. That can be a possibility for more data-driven research work.

9 Ethics Statement

Text detoxification technology has various practical applications, such as the automatic moderating output of generative neural networks to prevent reputation losses of companies trained and deployed such networks: think of a chatbot that could respond impolitely due to inherent biases in training data. On the other hand, completely automatic text detoxification of user-generated content should be done with extreme care. Instead, a viable use-case is to suggest that the user rewrite a toxic comment (e.g., to save her digital reputation as the 'internet remembers everything'). It is, however, crucial to leave the freedom to a person to express comment in the way she wants, given legal boundaries. Text detoxification (including the cross-lingual variant studied in this paper) is useful. Yet, each concrete integration should be done considering a balance between the need to protect users (e.g., children playing video games and being offended by other gamers in a chat) and the freedom of speech. Each social media platform and country has its own approach to this issue, besides these policies dynamically emerging in time. Therefore, making comprehensive guidelines for using this technology is beyond the scope of our study.

References

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 300–321, Dublin, Ireland. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv e-prints*, pages arXiv–2207.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daryna Dementieva, Varvara Logacheva, Irina Nikishina, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. RUSSE-2022: Findings of the first Russian detoxification task based on parallel corpora. In *Computational Linguistics and Intellectual Technologies*.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Methods for detoxification of texts for the russian language. *Multimodal Technol. Interact.*, 5(9):54.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 189–194. Association for Computational Linguistics.

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6048–6056.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Vadim Gudkov, Olga Mitrofanova, and Elizaveta Filippskikh. 2020. Automatically ranked Russian paraphrase corpus for text generation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 54–59, Online. Association for Computational Linguistics.

Ilya Gusev. 2022. Russian texts detoxification with levenshtein editing. *CoRR*, abs/2204.13638.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2022. Detoxifying text with marco: Controllable revision with experts and anti-experts. *CoRR*, abs/2212.10543.

Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *CoRR*, abs/2112.08804.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Jigsaw. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge. Accessed: 2021-03-01.

Jigsaw. 2020. Jigsaw multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification. Accessed: 2021-03-01.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 424–434. Association for Computational Linguistics.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7439–7468. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *CoRR*, abs/1905.07213.

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022a. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2022b. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 262–271. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina, and Alexander Panchenko. 2022a. 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.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022b. 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.

Nikita Martynov, Irina Krotova, Varvara Logacheva, Alexander Panchenko, Olga Kozlova, and Nikita Semenov. 2022. Rupaws: A russian adversarial dataset for paraphrase identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 5683–5691. European Language Resources Association.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. Rucola: Russian corpus of linguistic acceptability. *CoRR*, abs/2210.12814.

Jihyung Moon, Won-Ik Cho, and Junbum Lee. 2020. Beep! korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2020, Online, July 10, 2020*, pages 25–31. Association for Computational Linguistics.

Daniil Moskovskiy, Daryna Dementieva, and Alexander Panchenko. 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.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 314–319. Association for Computational Linguistics.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation. *CoRR*, abs/2306.00539.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9408–9423. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Aleksandr Semiletov. 2020. Toxic russian comments. https://www.kaggle.com/alexandersemiletov/toxic-russian-comments. Accessed: 2021-07-22.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1174–1182. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT - building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine*

Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020, pages 479–480. European Association for Machine Translation.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2107–2114, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Natalia Vanetik and Elisheva Mimoun. 2022. Detection of racist language in french tweets. *Inf.*, 13(7):318.

Md. Anwar Hussen Wadud, Muhammad F. Mridha, Jungpil Shin, Kamruddin Nur, and Aloke Kumar Saha. 2023. Deep-bert: Transfer learning for classifying multilingual offensive texts on social media. *Comput. Syst. Sci. Eng.*, 44(2):1775–1791.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. "mask and infill" : Applying masked language model to sentiment transfer. *CoRR*, abs/1908.08039.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

A Examples of Detoxification Models Outputs

Model	Source	Generated
	Monolingual Models	
ruT5-detox	Судя по -всему, дебилы, которые готовы отдать за этот таз 1млн. руб закончились, а бабки как то осваивать надо А учиться чему что уехала за бугор бл**ь что она в беларуси не осталось и дальше не боролась оппозиция называется П**деж. Бабка бы удавилась за свои кровные. Вброс расчитан на тупое быдло.	Судя по всему те люди, которые готовы отдать за этот таз 1 млн руб закончились, а деньги как то осваивать надо А учиться чему? Что уехала заграницу, что она в Беларуси не осталась и дальше не боролась. Оппозиция называется вброс. бабка бы удавилась за свои кровные. вброс расчитан на глупых людей
BART-detox	He is a f**king parasite. This is a joke , are you all f**king retards? They are two faced f**ks.	He is not good. This is a joke, are you all naive? They are two faced people.
	Backtranslation	
ruT5-detox (FSMT)	He is a f**king parasite. This is scaring the s**t out of me. They are two faced f**ks.	He is not a good person, he is a very bad person. This scares the hell out of me. They are bad people
BART-detox (Yandex)	Лучше б руководство выкинули бы на хрен!!!! Ты бы уже п**ло язык за зубами держал Вот таким е**нашкам ещё права дают!!!	Лучше бы руководство уволили!!!! Ты бы держал рот на замке Дают права таким людям! Они не милые.
	Multitask Learning	
mBART RU-Tr (Helsinki)	Г**но,оно и в Африке г**но. П**оры живущие за счёт народа, в биологии называются паразитами, организм живущий за счёт другова А потом подобные идиоты Сапсаны под откос пускают.	Это и в Африке нехорошо. Люди живущие за счёт народа, в биологии называются паразитами, организм живущий за счет другого. А потом такие люди под откос пускают, как Сапсан.
mBART EN-Tr (FSMT)	Вот х**и вам бабам еще надо? такой прискурант озвучил! Политика это вообще один большой фейспалм, стадо п**оров, на**ывающих друг друга. Как можно было такую уродку выдвигать в депутаты?	вот что вам еще надо? такой прискурант озвучил! Политика - это вообще один большой фейспалм, где люди разговаривают друг с другом. Как можно было её выдвигать в депутаты?
	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 ldiot 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**нутая 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)	тьфу 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 anymore There's nothing else to show.

Table 6: Examples of models' output: baselines and our proposed approaches.

B Examples of Translation Outputs

Translator	Source	Translated	Quality
FSMT (Ng et al., 2019)	what the f**k is this comment ?	Что это за х**ня?	✘
FSMT	the fact we are doing this now and it started when i was number number is fucking insane !	тот факт, что мы делаем это сейчас, и это началось, когда я был номером, это гребано!	✘
FSMT	the germans today are the same blind brainless followtheleader sheep they were in and throughout their history.	сегодняшние немцы такие же слепые безмозглые овчарки, какими они были в своей истории.	✔
FSMT	the guy was f**king shot !	парень гребаный стрелял!	✘
FSMT	maybe now no one will be rammed in the *ss .	Может быть, теперь никто не будет ругаться в з**ницу.	✘
OPUS-MT (Tiedemann and Thottingal, 2020)	what the f**k is this comment ?	Что это за х**ня?	✘
OPUS-MT c	the fact we are doing this now and it started when i was number number is f**king insane !	То, что мы делаем сейчас и это началось когда я был номером - сумасшествие!	✘
OPUS-MT	the germans today are the same blind brainless followtheleader sheep they were inand throughout their history .	Сегодня немцы - те же слепые безмозглые овцы, что и во всей их истории.	✔
OPUS-MT	the guy was f**king shot !	Парня застрелили!	✘
OPUS-MT	maybe now no one will be rammed in the a** .	Может быть, теперь никто не будет запихнут в ж**у.	✘
OPUS-MT	it s actually a great idea if you want to keep nuts out and get s**t done .	Это действительно отличная идея, если ты хочешь держаться подальше и заниматься д*рьмом.	✘
Yandex.Translate	what the f**k is this comment ?	что, черт возьми, это за комментарий?	✔
Yandex.Translate	the fact we are doing this now and it started when i was number number is f**king insane !	тот факт, что мы делаем это сейчас, и это началось, когда я был номером номер, чертовски безумен!	✔
Yandex.Translate	the germans today are the same blind brainless followtheleader sheep they were inand throughout their history .	немцы сегодня - такие же слепые безмозглые овцы, следующие за лидером, какими они были на протяжении всей своей истории.	✔
Yandex.Translate	the guy was f**king shot !	этого парня, б**дь, застрелили!	✔
Yandex.Translate	maybe now no one will be rammed in the a** .	может быть, теперь никого не будут таранить в з*дницу.	✘
Yandex.Translate	it s actually a great idea if you want to keep nuts out and get s**t done .	на самом деле это отличная идея, если вы хотите не сходить с ума и довести дело до конца.	✔
Google.Translate	what the f**k is this comment ?	что за бред этот комментарий?	✔
Google.Translate	the fact we are doing this now and it started when i was number number is f**king insane !	тот факт, что мы делаем это сейчас, и это началось, когда я был номером номер, чертовски безумен!	✔
Google.Translate	the germans today are the same blind brainless followtheleader sheep they were inand throughout their history .	нынешние немцы - такие же слепые безмозглые овцы, следующие за вожаками, которыми они были на протяжении всей своей истории.	✔
Google.Translate	the guy was f**king shot !	парень был чертовски застрелен!	✔
Google.Translate	maybe now no one will be rammed in the a** .	может теперь никто не будет таранить под з*д.	✘
Google.Translate	it s actually a great idea if you want to keep nuts out and get s**t done .	на самом деле это отличная идея, если вы хотите держаться подальше от орехов и делать д*рьмо.	✔

Table 7: Examples of translations from English to Russian.

Translator	Source	Translated	Quality
FSMT (Ng et al., 2019)	бл**ь, ты хоть себя слышишь?) ты говоришь что я экстрасенс, а потом говоришь, что нет	Do you even hear yourself?) You say I'm a psychic, and then you say no.	✘
FSMT (Ng et al., 2019)	лично я хочу чтоб мр*зи сели на пожизненое	Personally, I want them to sit down for life.	✘
FSMT (Ng et al., 2019)	тварь,трус! ничего человеческого не осталось	Creature, c*ward! There is nothing human left.	✘
FSMT (Ng et al., 2019)	От этого пострадают только всякие усть-переп**дюйск-телекомы с 3.5 сотрудниками	Only those with 3.5 employees will be affected.	✔
FSMT (Ng et al., 2019)	иди н**ер, верните иваныча, черти!	Go n**her, bring back Ivanich, devils!	✔
OPUS-MT (Tiedemann and Thottingal, 2020)	бл**ь, ты хоть себя слышишь?) ты говоришь что я экстрасенс, а потом говоришь, что нет	Can you f**king hear yourself?) You say I'm a psychic, and then you tell me no.	✔
OPUS-MT (Tiedemann and Thottingal, 2020)	лично я хочу чтоб мр*зи сели на пожизненое	Personally, I want the b*stards to sit down for life.	✔
OPUS-MT (Tiedemann and Thottingal, 2020)	тварь,трус! ничего человеческого не осталось	You son of a b**ch! There's nothing human left.	✔
OPUS-MT (Tiedemann and Thottingal, 2020)	От этого пострадают только всякие усть-переп**дюйск-телекомы с 3.5 сотрудниками	This will only cause damage to any of the three-way telecoms with 3.5 employees.	✘
OPUS-MT (Tiedemann and Thottingal, 2020)	эти бл**и совсем о**ели тв*ри конченые	These f**king things are so f**ked up.	✘
OPUS-MT (Tiedemann and Thottingal, 2020)	иди н**ер, верните иваныча, черти!	Go f**k yourself, get the Ivanich back!	✘
Yandex.Translate	бл**ь, ты хоть себя слышишь?) ты говоришь что я экстрасенс, а потом говоришь, что нет	Can you f**king hear yourself?) You say I'm a psychic, and then you tell me no.	✔
Yandex.Translate	лично я хочу чтоб мр*зи сели на пожизненое	Personally, I want the sc*m to go to prison for life.	✔
Yandex.Translate	тварь,трус! ничего человеческого не осталось	You coward! There's nothing human left.	✔
Yandex.Translate	От этого пострадают только всякие усть-переп**дюйск-телекомы с 3.5 сотрудниками	Only Ust-perep**dyuisk telecoms with 3.5 employees will suffer from this	✔
Yandex.Translate	эти бляди совсем о**ели твари конченые	these whores are completely f**ked up creatures are finished	✘
Yandex.Translate	иди н**ер, верните иваныча, черти!	go to hell, bring Ivanovich back, damn it!	✔
Google.Translate	бл**ь, ты хоть себя слышишь?) ты говоришь что я экстрасенс, а потом говоришь, что нет	f**k, can you even hear yourself?) you say that I'm a psychic, and then you say that I'm not	✔
Google.Translate	лично я хочу чтоб мр*зи сели на пожизненое	I personally want the sc*m to sit on a life sentence	✔
Google.Translate	тварь,трус! ничего человеческого не осталось	creature, c*ward! nothing human left	✔
Google.Translate	От этого пострадают только всякие усть-переп**дюйск-телекомы с 3.5 сотрудниками	Only all sorts of Ust-Perep**duysk-Telecoms with 3.5 employees will suffer from this	✔
Google.Translate	эти бл**и совсем охуели тв*ри конченые	these whores are completely f**ked up by the finished creatures	✘
Google.Translate	иди н**ер, верните иваныча, черти!	go to hell, bring Ivanovich back, d*mn it!	✔

Table 8: Examples of translations from Russian to English.

C Comparison of Proposed Approaches based on Required Resources

Method	**Models**	**Datasets**	**Data Creation**	**Fine tuning**	**# Inference Steps**
Backtranslation	- Detoxification model for the resource-rich language; - Translation model to the target language;	—	✘	✘	3
Training Data Translation	- Translation model to the target language; - Auto-regressive multilingual or monolingual LM for the target language;	- ParaDetox on the resource-rich language;	✔	✔	1
Multitask Learning	- 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
Adapter Training	- 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

Table 9: Comparison of the proposed approaches for cross-lingual detoxification transfer based on required computational and data resources. As one may observe, backtranslation approach requires 3 runs of seq2seq models, while other approaches are based on a single (end2end) model and require only one run.

D Human vs Automatic Evaluation Correlations for Old and New Setups

The detailed correlation results of new and old automatic metrics for the Russian language: (i) based on system score (Table 10); (ii) based on system ranking (Table 11).

In the first approach, we concatenate all the scores of all systems for corresponding metrics in one vector and calculate Spearman's correlation between such vectors for human and automatic evaluation. For the second approach, we rank the systems based on the corresponding metric, get the vector of the systems' places in the leaderboard, and calculate Spearman's correlation between such vectors for human and automatic evaluation. We can observe improvements in correlations for both setups with newly presented metrics.

Metric	STA_a^{old}	SIM_a^{old}	FL_a^{old}	J_a^{old}
STA_m	0.472	**-0.324**	-0.121	0.120
SIM_m	**-0.062**	0.124	0.084	-0.026
FL_m	0.018	**-0.087**	-0.011	-0.132
J_m	0.271	**-0.138**	-0.031	0.106
Metric	STA_a	SIM_a	FL_a	J_a
STA_m	**0.598**	-0.071	**0.130**	**0.516**
SIM_m	-0.012	**0.244**	**0.217**	**0.176**
FL_m	**0.107**	0.054	**0.354**	**0.229**
J_m	**0.370**	0.096	**0.259**	**0.482**

Table 10: Spearman's correlation coefficient between automatic VS manual metrics based on systems scores for **Russian** language. All numbers denote the statistically significant correlation (p-value ≤ 0.05).

Metric	STA_a^{old}	SIM_a^{old}	FL_a^{old}	J_a^{old}
STA_m	0.235	**-0.657**	-0.200	0.138
SIM_m	0.130	0.015	0.240	0.248
FL_m	-0.024	-0.284	0.024	0.002
J_m	0.169	-0.116	0.204	0.231
Metric	STA_a	SIM_a	FL_a	J_a
STA_m	**0.811**	-0.231	**0.600**	**0.692**
SIM_m	**0.240**	**0.732**	**0.349**	**0.648**
FL_m	**0.292**	**0.305**	**0.868**	**0.613**
J_m	**0.433**	**0.565**	**0.534**	**0.802**

Table 11: Spearman's correlation coefficient between automatic VS manual metrics based on system ranking for **Russian** language. All numbers denote the statistically significant correlation (p-value ≤ 0.05)

E Comparison of Translation Methods

Here, we provide a thorough comparison of all mentioned translation methods for presented approaches: (i) Cross-lingual Detoxification Transfer (Table 12); (ii) Detox&Translation (Table 13). Additionally, we provide the experiments for *multilingual* setup (where the detoxification models are trained on datasets in both languages simultaneously) for *Training Data Translation* approach in Table 14.

	STA	**SIM**	**FL**	**J**	**STA**	**SIM**	**FL**	**J**
	Russian				**English**			
	Cross-lingual Detoxification Transfer							
	Backtranslation							
ruT5-detox (FSMT)	—				**0.680**	0.458	**0.902**	**0.324**
ruT5-detox (Google)	—				0.643	0.565	0.884	0.311
ruT5-detox (Yandex)	—				0.627	**0.579**	0.896	0.316
ruT5-detox (Helsinki)	—				0.631	0.544	0.892	0.297
BART-detox (FSMT)	0.547	0.628	0.772	0.258	—			
BART-detox (Google)	0.578	**0.721**	0.815	0.333	—			
BART-detox (Yandex)	0.601	0.709	**0.832**	**0.347**	—			
BART-detox (Helsinki)	**0.607**	0.591	0.776	0.277	—			
mBART (FSMT)	0.545	0.629	0.781	0.263	**0.706**	0.460	0.844	0.269
mBART (Helsinki)	**0.599**	0.598	0.774	0.276	0.671	0.503	0.859	0.285
mBART (Yandex)	0.595	0.710	**0.835**	**0.345**	0.661	**0.561**	**0.913**	**0.322**
mBART (Google)	0.566	**0.722**	0.808	0.325	0.668	0.547	0.887	0.312
	Training Data Translation							
mBART RU-Tr (FSMT)	**0.432**	0.758	**0.781**	0.253	—			
mBART RU-Tr (Yandex)	0.384	**0.773**	0.780	0.228	—			
mBART RU-Tr (Helsinki)	0.429	**0.773**	0.780	**0.257**	—			
mBART EN-Tr (FSMT)	—				**0.762**	0.553	**0.871**	**0.354**
mBART EN-Tr (Yandex)	—				0.648	**0.623**	0.838	0.320
mBART EN-Tr (Helsinki)	—				0.646	0.618	0.858	0.319

Table 12: Evaluation of TST models. Numbers in **bold** indicate the best results by each parameter inside of the subsections. Rows in green indicate the best models chosen for the main results comparison. EN-Tr or RU-Tr denote translated versions of ParaDetox.

	STA	**SIM**	**FL**	**J**	**STA**	**SIM**	**FL**	**J**
	Russian				**English**			
	Detox&Translation							
	Detoxification with Translation							
ruT5-detox (Yandex)	—				0.834	**0.494**	0.705	0.297
ruT5-detox (Google)	—				0.829	0.490	0.686	0.284
ruT5-detox (FSMT)	—				**0.930**	0.396	**0.794**	**0.300**
ruT5-detox (Helsinki)	—				0.811	0.442	0.770	0.279
BART-detox (Yandex)	**0.774**	**0.699**	**0.876**	**0.470**	—			
BART-detox (Google)	0.773	0.680	0.845	0.440	—			
BART-detox (FSMT)	0.674	0.490	0.802	0.266	—			
BART-detox (Helsinki)	0.674	0.614	0.802	0.325	—			
	Cross-lingual Training Data							
mBART (Yandex)	**0.788**	**0.562**	**0.744**	**0.333**	**0.922**	**0.446**	**0.728**	**0.305**
mBART (Google)	0.749	0.516	0.727	0.277	0.894	0.365	0.703	0.230
mT5-base (Yandex)	**0.773**	**0.569**	**0.721**	**0.315**	**0.880**	**0.414**	**0.655**	**0.250**
mT5-base (Google)	0.765	0.473	0.602	0.218	0.861	0.343	0.573	0.173
mT5-large (Yandex)	**0.782**	**0.592**	**0.790**	**0.361**	**0.897**	0.393	0.558	0.204
mT5-large (Google)	0.745	0.536	0.708	0.280	0.846	**0.410**	**0.713**	**0.250**

Table 13: Evaluation of TST models. Numbers in **bold** indicate the best results by each parameter inside the subsections. Rows in green indicate the best models chosen for the comparison of the main results.

	STA	SIM	FL	J	STA	SIM	FL	J
	Russian				**English**			
	Multilingual Detoxification							
	Training Data Translation							
mBART EN+RU-Tr (FSMT)	**0.490**	0.734	**0.788**	**0.278**	0.863	0.633	0.838	0.450
mBART EN+RU-Tr (Yandex)	0.410	**0.771**	0.786	0.249	0.852	0.636	0.826	0.440
mBART EN+RU-Tr (Helsinki)	0.458	**0.771**	0.784	0.276	0.881	0.550	0.739	0.360
mBART EN-Tr+RU (FSMT)	0.613	0.775	0.781	0.370	0.692	0.583	0.861	0.327
mBART EN-Tr+RU (Yandex)	0.453	0.769	0.784	0.272	0.768	**0.593**	0.857	0.376
mBART EN-Tr+RU (Helsinki)	0.584	0.780	0.782	0.356	**0.792**	0.583	**0.870**	**0.386**

Table 14: Evaluation of TST models. Numbers in **bold** indicate the best results by each parameter inside the subsections. EN-Tr or RU-Tr denote translated versions of ParaDetox.

F Manual Evaluation Instructions

Here, we present the explanation of labels that annotators had to assign for each of the three evaluation parameters. We adapt the manual annotation process described in (Logacheva et al., 2022a):

Toxicity ($\mathbf{STA}_m$) *Is this text offensive?*

- **non-toxic** (1) — the sentence does not contain any aggression or offence. However, we allow covert aggression and sarcasm.
- **toxic** (0) — the sentence contains open aggression and/or swear words (this also applies to meaningless sentences).

Content ($\mathbf{SIM}_m$) *Does these sentences mean the same?*

- **matching** (1) — the output sentence fully preserves the content of the input sentence. Here, we allow some change of sense which is inevitable during detoxification (e.g., replacement with overly general synonyms: *idiot* becomes *person* or *individual*). It should also be noted that content and toxicity dimensions are independent, so if the output sentence is toxic, it can still be good in terms of content.
- **different** (0) — the sense of the transferred sentence differs from the input. Here, the sense should not be confused with the word overlap. The sentence is different from its original version if its main intent has changed (cf. *I want to go out* and *I want to sleep*). The partial loss or change of sense is also considered a mismatch (cf. *I want to eat and sleep* and *I want to eat*). Finally, when the transferred sentence is senseless, it should also be considered *different*.

Fluency ($\mathbf{FL}_m$) *Is this text correct?*

- **fluent** (1) — sentences with no mistakes, except punctuation and capitalization errors.
- **partially fluent** (0.5) — sentences which have orthographic and grammatical mistakes, non-standard spellings. However, the sentence should be fully intelligible.
- **non-fluent** (0) — sentences which are difficult or impossible to understand.

However, since all the input sentences are user-generated, they are not guaranteed to be fluent in terms of this scale. People often make mistakes and typos and use non-standard spelling variants. We cannot require that a detoxification model fixes them. Therefore, we consider the output of a model fluent if the model did not make it less fluent than the original sentence. Thus, we evaluate both the input and the output sentences and define the final fluency score as **fluent** (1) if the fluency score of the output is greater or equal to that of the input, and **non-fluent** (0) otherwise.