Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Prompt Tuning for Text Detoxification

Nikita Konodyuk
SberDevices
Moscow, Russia
nekonodyuk@sberbank.ru

Abstract

Text detoxification is a challenging style transfer task, that implies paraphrasing into a neutral form while preserving the meaning as closely as possible. In this paper, we present a lightweight approach based on a recently proposed prompt tuning technique. Using RuGPT3-XL (Generative Pretrained Transformer-3 for Russian) as a frozen backbone, we train only a sequence of continuous embeddings inserted before and after an input text. Even though the number of trainable parameters is less than 0.025% of their total number, our approach achieves competitive performance compared to the methods involving full model tuning and ranks 4th on the leaderboard of shared RUSSE Detox task.

Keywords: text detoxification, text generation, RuGPT3, prompt tuning
DOI: 10.28995/2075-7182-2022-21-1089-1096

Автоматический подбор затравок для детоксификации

Никита Конодюк
SberDevices
Москва, Россия
nekonodyuk@sberbank.ru

Аннотация

Детоксификация текстов является нетривиальной задачей переноса стиля, которая подразумевает парафразирование в нейтральную форму с как можно более точным сохранением смысла. В данной статье мы представляем эффективный метод, основанный на недавно предложенной технике автоматического подбора затравок. Используя RuGPT3-XL (Generative Pretrained Transformer-3 for Russian) как неизменяемую предобученную модель, мы обучаем только последовательность непрерывных векторных представлений, вставляемых до и после входного текста. Несмотря на то, что количество обучаемых параметров составляет менее чем 0.025% от их общего числа, наш метод достигает сравнимого качества относительно методов, включающих в себя дообучение всей модели, и занимает 4-е место в таблице результатов соревнования RUSSE Detox.

Ключевые слова: детоксификация текстов, генерация текстов, RuGPT3, подбор затравок

1 Introduction

Detection and elimination of toxicity in texts is an active area of research. The issue is particularly acute in the context of social networks that are trying to automate moderation and reduce the overall toxicity of the environment. Although this can already be achieved simply by blocking inappropriate messages, proactive correction, i.e. autosuggestions of detoxified messages, could provide a better user experience by preserving the possibility of constructive communication, at the same time decreasing the toxicity.

Text detoxification is the task of rewriting an offensive text into a neutral form preserving its meaning. It thus can be considered a style transfer task with toxic as a source style and neutral as a target and can be solved using traditional style transfer methods using parallel corpora.

Until recently, however, the parallel corpora of toxic and detoxified texts in Russian did not exist. That led to the domination of unsupervised approaches, such as (Dale et al., 2021).

RUSSE Detox shared task (Dementieva et al., 2022) provides the first parallel detoxification dataset for Russian, which allows exploring the capabilities of generic text-to-text methods in application to the task. In this paper, we present a solution based on prompt tuning. As a backbone, we use RuGPT3 of two scales: Large (760M parameters) and XL (1.3B). We show that training only a sequence of prompt embeddings is enough to adapt the backbone to the detoxification task and conduct experiments to find the optimal prompt tuning configuration.

We thus make the following contributions:

- We apply prompt tuning to adapt an LM backbone to text detoxification task using a parallel corpus of Russian data.
- We conduct experiments to determine the optimal length of trainable prompt for the task.
- We show that prompt tuning alone does not achieve satisfactory results and propose a decoding trick to handle the prompt tuning errors.

The remaning part of the paper proceeds as follows. Section 2 contextualizes the research by providing the background information on text detoxification methods and introducing prompt tuning. Section 3 describes the provided data and evaluation protocol. Section 4 specifies the approach. Section 5 presents the results of evaluation and additional experiments. Section 6 concludes the paper.

2 Related Work

2.1 Text Detoxification

Text detoxification is a relatively new style transfer task, which is primarily solved using unsupervised style transfer methods because of the lack of sufficiently large parallel corpora. In (Santos et al., 2018), training an autoencoder with additional style classification and cycle-consistency losses is proposed. (Tran et al., 2020) apply pointwise corrections with subsequent postprocessing with a seq2seq model to improve fluency. (Laugier et al., 2021) fine-tune T5 as a denoising autoencoder, following the setting of (Santos et al., 2018). (Krause et al., 2020) propose training an auxiliary discriminator model to condition the LM during generation. (Dale et al., 2021) modify the approach of (Krause et al., 2020) by using a paraphrasing LM instead of a regular one. In (Dementieva et al., 2021), two models for the Russian language are proposed:

- condBERT masks and replaces toxic tokens following the approach of (Wu et al., 2019) with the difference that the toxicity level for masking is determined by a bag-of-words logistic regression model.
- detoxGPT is a RuGPT-3 model fine-tuned on a small parallel corpus of 200 samples.

Apart from (Dementieva et al., 2021), the research on toxicity elimination in Russian mostly focused on classification tasks ((Andrusyak et al., 2018), (Smetanin, 2020), (Zueva et al., 2020), etc.).

2.2 Prompt Tuning

With the wide success of large pre-trained language models, a range of techniques have arisen to adapt these foundation models to downstream tasks. Since BERT (Devlin et al., 2018), *fine-tuning* has been the prevalent adaptation technique. The emergence of GPT-3 (Brown et al., 2020), however, demonstrated that giant autoregressive LMs can be applied to multiple downstream tasks via *in-context learning*, without any task-specific tuning.

For example, solving a machine translation task in *few-shot* setting implies passing a task description and multiple training examples along with a test prompt:

```
Translate English to French:  ← task instructions
sea otter => loutre de mer ← examples
pepperming => menthe poivree ← examples
plush girafe => girafe peluche ← examples
cheese => ← test prompt
```

In *zero-shot* setting the model receives no training examples and should understand the task only from its description:

Toxic sentence	Detoxified sentence
из за таких пидоров мы и страдаем	Из-за таких людей мы и страдаем
хуй знает кто кум, но девушка красивая	неизвестно кто кум, но девушка красивая
порядок бы блять навёл !	Порядок бы навел

Table 1: Examples of detoxified sentences from the parallel corpus.

```
Translate English to French:  ← task instructions
cheese => ← test prompt
```

In both settings, the model is expected to generate the answer.

Although in-context learning is efficient on a wide range of tasks, the approach heavily relies on handcrafted task descriptions and prompt formats. However, manual prompt search is error-prone and requires human involvement. This led to the emergence of a range of automatic prompt search techniques.

In this paper, we adopt an approach that we will further refer to as *prompt tuning*. Consider an example task of toxicity detection. To classify the sentence `You're a duck.` as toxic or non-toxic in the zero-shot setting, we will have to handcraft a prompt like this:

```
Is this sentence toxic:  "You're a duck."?  Answer:
```

Note that the following ranges are manual instructions:

```
Is this sentence toxic:  "You're a duck."?  Answer:
```

Prompt tuning suggests the embeddings corresponding to task instructions to be learned automatically using the training data.

```
<instruction embeddings>You're a duck.<instruction embeddings>
```

In particular, the separate trainable embeddings (`<P[i]>` denotes the token, corresponding to the i-th trainable embedding) are optimized via gradient descent.

```
<P[0]><P[1]>...<P[i]>You're a duck.<P[i+1]>...<P[n]>
```

The method is loss-agnostic and thus can be used for multiple task types, such as text classification and text-to-text. In Section 4 we define our approach to using prompt tuning for style transfer in more detail.

Prefix-Tuning (Li and Liang, 2021) was first to propose the optimization of continuous prompts. The paper focused on text-to-text tasks, such as summarization and table description, and conducted experiments with GPT-2 and BART on E2E, WebNLG, DART, and XSUM datasets. The method outperformed fine-tuning baselines but required prefix embeddings of each transformer layer to be tuned separately.

GPT Understands Too (Liu et al., 2021) focused on NLU tasks and proposed BiLSTM reparameterization of trainable prompt. The method outperformed fine-tuned GPT-2 on multiple SuperGLUE (Wang et al., 2019) tasks but required the prompts to be adapted jointly with model weights.

The Power of Scale for Parameter-Efficient Prompt Tuning (Lester et al., 2021) conducted experiments on SuperGLUE with T5 as a backbone and achieved performance competitive with fine-tuning by using longer prompts without reparameterization. It was also demonstrated that prompt tuning becomes more competitive with scale and that prompt initialization from vocabulary embeddings leads to more stable training.

3 Dialog Evaluation 2022: Detoxification Shared Task

3.1 Data

The organizers of the RUSSE-2022 Detoxification shared task introduced a parallel text detoxification dataset in Russian collected via Yandex.Toloka crowdsourcing platform. We show examples of the parallel data in Table 1. The dataset contains 8622 examples overall and is split into training (6947 examples), validation (800), and test (875) partitions.

3.2 Evaluation

The shared task is evaluated with the three metrics of style transfer quality, following the setup of (Krishna et al., 2020):

- **Style Transfer Accuracy (STA)** is automatically evaluated using a BERT-based toxicity classifier.
- **Content Preservation (SIM)** is automatically evaluated as the cosine similarity of embeddings of the source and detoxified sentences using a LaBSE model (Feng et al., 2020).
- **Fluency (FL)** is automatically evaluated using an acceptability classifier trained on a synthetically generated dataset of normal and corrupted sentences.

These metrics are aggregated into **Joint (J)** score by multiplication.

ChrF is computed as an additional reference-based metric following the machine translation evaluation setup.

On the stage of manual evaluation, STA, SIM, and FL are computed via crowdsourcing.

3.3 Evaluation Issues

Although the metrics collected during the human and automatic evaluation were the same, the automatic approximation was not accurate enough to yield a reliable correlation with human scores. As a result, automatic evaluation of the reference answers from the validation partition gave a 0.44 joint score, making model-based evaluation not informative. At the same time, the ChrF has proven to be closer to human assessment than model-based metrics.

4 Approach

We handle text detoxification as a text-to-text task and use prompt tuning to adapt a pre-trained GPT model to a parallel corpus.

In particular, each pair from the parallel corpus is formatted as

```
<P*N>{toxic_text}<P*M>{normal_text}<EOS>
```

where `<P*N> := <P[1]>...<P[N]>` and `<EOS>` is an end-of-sequence token. The EOS token is appended to train the model to limit its generation to detoxified text. The N and M are prompt length constants, further, we refer to $N + M$ as prompt length. In our experiments, only N is varied, and M is a constant of value 20.

As described in Section 2.2, each `<P[i]>` token corresponds to an automatically inserted trainable embedding. At the training stage, for each sequence in a batch, we compute LM loss only for tokens corresponding to the `{normal_text}<EOS>` part of the input. The gradients are then propagated to the trainable embeddings, and they are the only parameters that are updated.

On inference stage, we pass the prompt

```
<P*N>{toxic_text}<P*M>
```

and the detoxification result is generated autoregressively.

4.1 Hyperparameters

In our experiments, we default to the hyperparameters listed in Table 2. Our final submission is created using the same parameters, but with RuGPT3 XL as a backbone.

4.2 Postprocessing

We encountered an unexpected issue at the inference stage, that emerged due to the autoregressive nature of our model. To obtain reproducible and deterministic outputs, we used beam search as a decoding method. However, in some cases, the model yielded the EOS token before the actual end of the text, which resulted in content loss and SIM metric decrease. To overcome this issue, while keeping the decoding process deterministic, we utilized a heuristic approach. In particular, for each sentence, we

Parameter	Value
`learning_rate`	$1e-1$
`batch_size`	2
`# steps`	100k
`backbone`	RuGPT3 Large
`prompt_length`	120
`postprocessing`	Sortmax

Table 2: Prompt tuning hyperparameters

Backbone	STA	SIM	FL	J	ChrF1
Large	0.7516	0.7726	0.8128	0.4774	0.5498
XL	0.7455	0.7794	0.8195	0.4756	0.5658

Table 3: Automatic evaluation with respect to model size. The hyperparameters except backbone are listed in Table 2.

Backbone	STA	SIM	FL	J
Large	0.803	0.703	0.866	0.493
XL	0.778	0.809	0.903	0.568

Table 4: Human with respect to model size. The hyperparameters except backbone are listed in Table 2.

Postprocessing	STA	SIM	FL	J	ChrF1
-	0.8292	0.6243	0.6463	0.3547	0.4439
Beam-Longest	0.7622	0.7451	0.7739	0.4440	0.5261

Table 5: Automatic evaluation with respect to postprocessing procedure. The hyperparameters except postprocessing method and prompt length are listed in Table 2. The prompt length is 105.

generated multiple candidates and selected the longest detoxified sentence. This postprocessing method is further denoted as *Beam-Longest*. The empirical results are reported in Section 5.3.

5 Results

5.1 Automatic Evaluation

In Table 3 we show the scores of automatic evaluation of prompt tuning using RuGPT-3 Large and RuGPT-3 XL models as a backbone. Note that the automatically estimated Joint score of both model scales is approximately equal, while ChrF of XL is significantly higher.

5.2 Human Evaluation

In Table 4 we show the performance of the same model scales in terms of human evaluation. The XL model significantly outperforms Large in the Joint score, which is not reflected by the automatic Joint score but correlates with ChrF.

5.3 Postprocessing

In Table 6 we compare the detoxification quality with and without Beam-Longest postprocessing. Without the postprocessing, the SIM score is lower because of the early truncation and consequent con-

Toxic	ты ебнулся , дядя ?
Reference	ты упал .,дядя
Detoxified	Ты что?
Detoxified (+Beam-Longest)	Ты ненормальный
Comment	Short example. Slightly better without Beam-Longest.
Toxic	это твари а не люди
Reference	Это плохие люди.
Detoxified	Это плохие люди
Detoxified (+Beam-Longest)	это нехорошие люди
Comment	Short example. Correct detoxification in both cases.
Toxic	да и этот пиндос живёт нанаши деньги.....
Reference	Да и этот живет на наши деньги.....
Detoxified	Да и этот человек живёт нанаши деньги
Detoxified (+Beam-Longest)	да и этот человек живёт нанаши деньги.....
Comment	Mid-length example. Correct detoxification in both cases.
Toxic	что у этих людей в башке!?походу насрато!
Reference	что у этих людей в голове!?
Detoxified	Что у этих людей в голове?
Detoxified (+Beam-Longest)	что у этих людей в голове!?похоже нас обманули!
Comment	Mid-length example. Beam-Longest adds unnecessary details.
Toxic	вспоминается на ум одна фраза(сказочный долбоёб,зачем его только из больницы выписали)так и здесь,очевидное невероятное
Reference	вспоминается на ум одна фраза(сказочный человек,зачем его только из больницы выписали)так и здесь,очевидное невероятное
Detoxified	вспоминается на ум одна фраза(зачем его только из больницы выписали)так и здесь,очевидное невероятное
Detoxified (+Beam-Longest)	вспоминается на ум одна фраза(сказочный человек зачем его только из больницы выписали)так и здесь,очевидное невероятное
Comment	Long example. Correct detoxification in both cases. Slightly better with Beam-Longest.
Toxic	гавнюки храбрые когда крутят женщин и стариков выхватывая из толпы по пять рыл на одного, а перед чурками отделались........
Reference	они храбрые, когда крутят женщин и стариков, выхватывая из толпы впятером одного, а перед кавказцами стушевались.
Detoxified	Храбрые когда крутят женщин и стариков выхватывая из толпы по пять человек на одного
Detoxified (+Beam-Longest)	Храбрые когда крутят женщин и стариков выхватывая из толпы по пять человек на одного, а перед кавказцами отделались........
Comment	Long example. Generation without Beam-Longest cuts the end of the sentence.

Table 6: Example of detoxification with and without Beam-Longest postprocessing.

Prompt Length	STA	SIM	FL	J	ChrF1
40	0.6819	**0.7929**	**0.8151**	0.4312	0.5464
80	0.7622	0.7451	0.7739	0.4440	0.5261
120	0.7516	0.7726	0.8128	0.4774	0.5498
255	**0.7823**	0.7595	0.7915	**0.4836**	**0.5512**

Table 7: Automatic evaluation with respect to prompt length. The hyperparameters except prompt length are listed in Table 2.

#	Team Name	STA	SIM	FL	J
	Human References	0.888	0.824	0.894	0.653
1	SomethingAwful	0.794	**0.872**	0.903	**0.633**
	T5 (baseline)	0.791	0.822	**0.925**	0.606
2	FRC CSC RAS	0.734	0.865	0.918	0.598
3	Mindful Squirrel	**0.824**	0.791	0.846	0.582
4	**team_ruprompts – Ours**	0.778	0.809	0.903	0.568
	Ruprompts (baseline) – Ours	0.803	0.703	0.866	0.493

Table 8: Final standings: top 4 teams and other relevant submissions.

tent loss. At the same time, the FL score also improves with Beam-Longest, which is not expected and may be attributed to the training details of the scoring model, e.g. using text truncation as one of the corruption types. The examples of detoxified sentences with and without Beam-Longest are listed in Table 1. The positive effect of postprocessing is observed mostly for longer sentences. On the other hand, for short sentences, the choice of the longest candidate may sometimes be suboptimal.

5.4 Prompt Length

In Table 7 we compare different lengths of trainable prompt. The longer prompts perform better in terms of both J and ChrF metrics. An interesting result is that the highest SIM and FL scores are obtained using a shorter prompt. This effect can be explained as follows. During training, the prompt does not directly adapt to the text detoxification task. Instead, it first learns to simply copy the input sentence without modifications, and only after that adapts to the required transformations. The prompt length of 40 may probably have an insufficient capacity to fully adapt to the detoxification task after learning to copy the input, which leads to higher SIM and FL scores since they are maximized by minimizing the number of corrections of input text.

5.5 Final Standings

In Table 8, we show the final standings by human evaluation. Our Large submission was provided as a baseline and ranked 9th (including baselines), and the XL submission ranked 4th (excluding baselines).

5.6 Parameter Efficiency

Given that the median length of prompt in our experiments is 120, and the embedding size of RuGPT3 Large is 1536, the median number of trainable parameters is approximately 184K. Considering that the total number of parameters of RuGPT3 Large itself is 760M, we are adjusting the number of parameters comparable with only 0.024% of all model parameters. In the case of RuGPT3 XL, this figure decreases to 0.019%.

6 Conclusion

In this paper, we present our submission to the RUSSE Detox shared task. We show that prompt tuning can be successfully applied to detoxification tasks and that as little as 0.024% trainable parameters are sufficient to achieve competitive results.

References

Bohdan Andrusyak, Mykhailo Rimel, and Roman Kern. 2018. Detection of abusive speech for mixed sociolects of russian and ukrainian languages. // *RASLAN*, P 77–84.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. *arXiv preprint arXiv:2109.08914*.

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 Technologies and Interaction*, 5(9):54.

Daryna Dementieva, Irina Nikishina, Varvara Logacheva, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. *arXiv preprint arXiv:2102.05456*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. *arXiv preprint arXiv:1805.07685*.

Sergey Smetanin. 2020. Toxic comments detection in russian. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue*.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. *arXiv preprint arXiv:2011.00403*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. // *International Conference on Computational Science*, P 84–95. Springer.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in russian hate speech detection. *arXiv preprint arXiv:2010.11666*.