Exploring Cross-lingual Textual Style Transfer with Large Multilingual Language Models

Daniil Moskovskiy[1] **Daryna Dementieva**[1,2] **Alexander Panchenko**[1]
[1]Skolkovo Institute of Science and Technology, Russia
[2]Technical University of Munich, Germany
{daniil.moskovskiy,daryna.dementieva,a.panchenko}@skoltech.ru

Abstract

Detoxification is a task of generating text in polite style while preserving meaning and fluency of the original toxic text. Existing detoxification methods are designed to work in one exact language. This work investigates multilingual and cross-lingual detoxification and the behavior of large multilingual models like in this setting. Unlike previous works we aim to make large language models able to perform detoxification without direct fine-tuning in given language. Experiments show that multilingual models are capable of performing multilingual style transfer. However, models are not able to perform cross-lingual detoxification and direct fine-tuning on exact language is inevitable.

1 Introduction

The task of Textual Style Transfer (Textual Style Transfer) can be viewed as a task where certain properties of text are being modified while rest retain the same[1]. In this work we focus on detoxification textual style transfer (dos Santos et al., 2018a; Dementieva et al., 2021a). It can be formulated as follows: given two text corpora $D^X = \{x_1, x_2, \dots x_n\}$ and $D^Y = \{y_1, y_2, \dots, y_n\}$, where X, Y - are two sets of all possible text in styles s^X, s^Y respectively, we want to build a model $f_\theta : X \to Y$, such that the probability $p(y_{gen}|x, s^X, s^Y)$ of transferring the style s^X of given text x (by generation y_{gen}) to the style s^Y is maximized (where s^X and s^Y are toxic and non-toxic styles respectively).

Some examples of detoxification presented in Table 1.

Textual style transfer gained a lot of attention with a rise of deep learning-based NLP methods. Given that, Textual Style Transfer has now a lot of specific subtasks ranging from formality style transfer (Rao and Tetreault, 2018; Yao and Yu, 2021) and simplification of domain-specific texts (Devaraj et al., 2021; Maddela et al., 2021) to emotion modification (Sharma et al., 2021) and detoxification (debiasing) (Li et al., 2021; Dementieva et al., 2021a).

There exist a variety of Textual Style Transfer methods: from totally **supervised** methods (Wang et al., 2019b; Zhang et al., 2020; Dementieva et al., 2021a) which require a parallel text corpus for training to **unsupervised** (Shen et al., 2017; Wang et al., 2019a; Xu et al., 2021) that are designed to work without any parallel data. The latter sub-field of research is more popular nowadays due to the scarcity of parallel text data for Textual Style Transfer. On the other hand, if we address Textual Style Transfer task as a Machine Translation task we get a significant performance boost (Prabhumoye et al., 2018).

The task of detoxification, in which we focus in this work, is relatively new. First work on detoxification was a sequence-to-sequence collaborative classifier, attention and the cycle consistency loss (dos Santos et al., 2018b). A recent work by (Laugier et al., 2021) introduces self-supervised model based on T5 model (Raffel et al., 2020) with a denoising and cyclic auto-encoder loss.

Both these methods are unsupervised which is an advantage but it comes from the major current problem of the textual style transfer. There is a lack of parallel data for Textual Style Transfer since there exist only few parallel datasets for English (Rao and Tetreault, 2018) and some other languages (Briakou et al., 2021). When it comes to detoxification there are only two parallel detoxification corpora available now and they both appeared only last year (Dementieva et al., 2021b). Most state-of-the-art methods rely on large amounts of text data which is often available for some well-researched languages like English but lacking for other languages almost entirely. Therefore, it is important to study whether cross-lingual (or at least multilingual) detoxifica-

[1]Hereinafter the data-driven definition of style is used. Therefore, we call style a characteristic of given dataset that differs from a general dataset (Jin et al., 2020).

Source text	Target text
What the f*ck is your problem?	What is your problem?
This whole article is bullshit.	This article is not good.
Yeah, this clowns gonna make alberta great again!	Yeah, this gonna make Alberta great again

Table 1: Examples of desired detoxification results.

tion is possible.

Multilingual language models such as mBART (Liu et al., 2020), mT5 (Xue et al., 2021) have recently become available. This work explores the possibility of multilingual and cross-lingual textual style transfer (Textual Style Transfer) using such large multilingual language models. We test the hypothesis that modern large text-to-text models are able to generalize ability of style transfer across languages.

Our contributions can be summarized as follows[2]:

1. We introduce a novel study of multilingual textual style transfer and conduct experiments with several multilingual language models and evaluate their performance.

2. We conduct cross-lingual Textual Style Transfer experiments to investigate whether multilingual language models are able to perform Textual Style Transfer without fine-tuning on a specific language.

2 Methodology

We formulate the task of **supervised** Textual Style Transfer as a sequence-to-sequence NMT task and fine-tune multilingual language models to translate from "toxic" to "polite" language.

2.1 Datasets

In this work we use two datasets for Russian and English languages. Aggregated information about datasets could be found in Table 2, examples from datasets can be found in A.1 and A.2.

Language	Train	Dev	Test
English	18777	988	671
Russian	5058	1000	1000

Table 2: Aggregated datasets statistics.

Russian data We use detoxification dataset[3] which consists of 5058 training sentences, 1000 validation sentences and 1000 test sentences.

English data We use ParaDetox (Dementieva et al., 2021b) dataset. It consists of 19766 *toxic* sentences and their *polite* paraphrases. This data is split into training and validation as 95% for training and 5% for validation. For testing we use a set of 671 toxic sentences.

2.2 Experimental Setup

We perform a series of experiments on detoxification using parallel data for English and Russian. We train models in two different setups: **multilingual** and **cross-lingual**.

Multilingual setup In this setup we train models on data containing both English and Russian texts and then compare their performance with baselines trained on these languages solely.

Cross-lingual setup In cross-lingual setup we test the hypothesis that models are able to perform detoxification without explicit fine-tuning on exact language. We fine-tune models on English and Russian separately and then test their performance.

2.3 Models

Scaling language models to many languages has become an emerging topic of interest recently (Devlin et al., 2019; Tan et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). We adopt several multilingual models to textual style transfer in our work.

Baselines We use two detoxification methods as baselines in this work - **Delete** method which simply deletes toxic words in the sentence according to the vocabulary of toxic words and **CondBERT**. The latter approach works in usual masked-LM setup by masking toxic words and replacing them with non-toxic ones. This approach was first proposed by (Wu et al., 2019) as a data augmentation

[2]All code is available online: `https://github.com/skoltech-nlp/multilingual_detox`

[3]`https://github.com/skoltech-nlp/russe_detox_2022`

method and then adopted to detoxification by (Dale et al., 2021).

mT5 mT5 (Xue et al., 2021) is a multilingual version of T5 (Raffel et al., 2020) - a text-to-text transformer model, which was trained on many downstream tasks. mT5 replicates T5 training but now it is trained on more than 100 languages.

mBART mBART (Liu et al., 2020) is a multilingual variation of BART (Lewis et al., 2020) - denoising autoencoder built with a sequence-to-sequence model. mBART is trained on monolingual corpora across many languages. We adopt mBART in sequence-to-sequence detoxification task via fine-tuning on parallel detoxification dataset.

2.4 Evaluation metrics

Unlike other NLP tasks, one metric is not enough to benchmark the quality of style transfer. The ideal Textual Style Transfer model output should *preserve the original content* of the text, *change the style* of the original text to target and the generated text also should *be grammatically correct*. We follow Dale et al. (2021) approach in Textual Style Transfer evaluation.

2.4.1 Content Preservation

Russian Content preservation score (**SIM**) is evaluated as a cosine similarity of LaBSE (Feng et al., 2020) sentence embeddings. The model is slightly different from the original one, only English and Russian embeddings are left.

English Similarity (**SIM**) between the embedding of the original sentence and the generated one is calculated using the model presented by Wieting et al. (2019). Being is trained on paraphrase pairs extracted from ParaNMT corpus (Wieting and Gimpel, 2018), this model's training objective is to select embeddings such that the similarity of embeddings of paraphrases is higher than the similarity between sentences that are not paraphrases.

2.4.2 Grammatic and language quality (fluency)

Russian We measure fluency (**FL**) 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 and insertion, word shuffling etc.) versions. Fluency is calculated as a difference between the probabilities of being corrupted for source and target sentences. The logic behind using difference is that we ensure that the generated sentence is not worse than the original one in terms of fluency.

English We measure fluency (**FL**) as a percentage of fluent sentences evaluated by the RoBERTa-based[4] (Liu et al., 2019) classifier of linguistic acceptability trained on CoLA (Warstadt et al., 2019) dataset.

2.4.3 Style transfer accuracy

Russian Style transfer accuracy (**STA**) is evaluated with a BERT-based (Devlin et al., 2019) toxicity classifier[5] fine-tuned from RuBERT Conversational. This classifier was additionally trained on Russian Language Toxic Comments dataset collected from `2ch.hk` and Toxic Russian Comments dataset collected from `ok.ru`.

English Style transfer accuracy (**STA**) is calculated with a style classifier - RoBERTa-based (Liu et al., 2019) model trained on the union of three Jigsaw datasets (Jigsaw, 2018). The sentence is considered toxic when the classifier confidence is above 0.8. The classifier reaches the AUC-ROC of 0.98 and F_1-score of 0.76.

2.4.4 Joint metric

Aforementioned metrics must be properly combined to get one *Joint* metric to evaluate Textual Style Transfer. We follow Krishna et al. (2020) and calculate **J** as an average of products of sentence-level *fluency*, *style transfer accuracy*, and *content preservation*:

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

2.5 Training

There is a variety of versions of large multilingual models available. In this work we use small and base versions of mT5[6,7] (Xue et al., 2021) and large version of mBART[8] (Liu et al., 2020).

[4] https://huggingface.co/roberta-large

[5] https://huggingface.co/SkolkovoInstitute/russian_toxicity_classifier

[6] https://huggingface.co/google/mt5-base

[7] https://huggingface.co/google/mt5-large

[8] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

	STA↑	SIM↑	FL↑	J↑	STA↑	SIM↑	FL↑	J↑
	Russian				**English**			
	Baselines							
Delete	0.532	0.875	0.834	0.364	0.810	0.930	0.640	0.460
condBERT (Dale et al., 2021)	0.819	0.778	0.744	0.422	**0.980**	0.770	0.820	0.620
	Multilingual Setup							
mT5 base	0.772	0.676	0.795	0.430	0.833	0.826	0.830	0.556
mT5 small	0.745	0.705	0.794	0.428	0.826	0.841	0.763	0.513
mT5 base*	0.773	0.676	0.795	0.430	0.893	0.787	**0.942**	**0.657**
mBART 5000	0.685	**0.778**	**0.841**	**0.449**	0.887	**0.889**	0.866	0.640
	Cross-lingual Setup							
mT5 base ENG	0.838	0.276	0.506	0.115	0.860	0.834	0.833	0.587
mT5 base RUS	0.676	0.794	0.846	0.454	0.906	0.365	0.696	0.171
mT5 small ENG	0.805	0.225	0.430	0.077	0.844	0.858	0.826	0.591
mT5 small RUS	0.559	0.822	0.817	0.363	0.776	0.521	0.535	0.169
mBART 3000 ENG	0.923	0.395	0.552	0.202	0.842	0.856	0.876	0.617
mBART 3000 RUS	0.699	0.778	0.858	0.475	0.547	0.778	0.888	0.299
mBART 5000 ENG	0.900	0.299	0.591	0.160	0.857	0.840	0.873	0.616
mBART 5000 RUS	0.724	0.746	0.827	0.457	0.806	0.484	0.864	0.242
	Backtranslation Setup							
mBART 5000 (Google)	0.675	0.669	0.634	0.284	0.678	0.762	0.568	0.284
mBART 5000 (FSMT)	0.737	0.633	0.731	0.348	0.744	0.746	0.893	0.415

Table 3: Evaluation of TST models. Numbers in **bold** indicate the best results. ↑ describes the higher the better metric. Results of unsuccessful TST depicted as gray. ENG and RUS depicts the data model have been trained on. mT5 base* was trained on all English and Russian data available (datasets were not equalized). Last row depicts backtranslation workaround for cross-lingual detoxification. We include only the best result for brevity.

Multilingual training In multilingual training setup we fine-tune models using both English and Russian data. We use Adam (Kingma and Ba, 2015) optimizer for fine-tuning with different learning rates ranging from $1 \cdot 10^{-3}$ to $5 \cdot 10^{-5}$ with linear learning rate scheduling. We also test different number of warmup steps from 0 to 1000. We equalize Russian and English data for training and use 10000 toxic sentences and their polite paraphrases for multilingual training in total. We train mT5 models for 40 thousand iterations[9] with a batch size of 8. We fine-tune mBART (Liu et al., 2020) for 1000, 3000, 5000 and 10000 iterations with batch size of 8.

Cross-lingual training In cross-lingual training setup we fine-tune models using only one dataset, e.g.: we fine-tune model on English data and check performance on both English and Russian data. Fine-tuning procedure was left the same: 40000 iterations for mT5 models and 1000, 3000, 5000 and 10000 iterations for the mBART.

Back-translation approach to cross-lingual style transfer proved to work substantially better than the zero-shot setup discussed above. Nevertheless, both Google and FSMT did not yield scores comparable to monolingual setup. Besides, surprisingly Google yielded worse results than FSMT.

[9]According to (Xue et al., 2021) mT5 was not fine-tuned on downstream tasks as the original T5 model. Therefore, model requires more fine-tuning iterations for Textual Style Transfer.

3 Results & Discussion

Table 3 shows the best scores of both multilingual and cross-lingual experiments. In multilingual setup mBART performs better than baselines and mT5 for both English and Russian. Note that the table shows only the best results of the models. It is also notable that for mT5 increased training size for English data provides better metrics for English while keeping metrics for Russian almost the same. We also depict some of the generated detoxified sentences in the Table 3 in the part B of Appendix.

As for cross-lingual style transfer, results are negative. None of the models have coped with the task of cross-lingual Textual Style Transfer. That means that models produce the same or almost the same sentences for the language on which they were not fine-tuned so that toxicity is not eliminated. We provide only some scores here in the Table 6 for reference.

Despite the fact that our hypothesis about the possibility of cross-language detoxification was not confirmed, the presence of multilingual models pretrained in many languages gives every reason to believe that even with a small amount of parallel data, training models for detoxification is possible.

A recent work by (Lai et al., 2022) shows that cross-lingual formality Textual Style Transfer is possible. Lai et al. (2022) achieve this on XFORMAL dataset (Briakou et al., 2021) by adding language-specific adapters in the vanilla mBART architecture (Liu et al., 2020) - two feed-forward layers with residual connection and layer normalization (Bapna and Firat, 2019; Houlsby et al., 2019).

We follow the original training procedure described by Lai et al. (2022) by training adapters for English and Russian separately on 5 million sentences from News Crawl dataset[10]. We use batch size of 16 and 200 thousand training iterations. We also then train cross-attentions on our parallel detoxifcation data in the same way. However, models tend to duplicate input text without any detoxification. Thus, while the exact same original setup did not work for detoxification, more parameter search and optimization could lead to more acceptable results and we consider the approach by Lai et al. (2022) as a promising direction of a future work on multilingual and cross-lingual detoxification.

4 Conclusion

In this work we have tested the hypothesis that multilingual language models are capable of performing cross-lingual and multilingual detoxification. In the multilingual setup we experimentally show that reformulating detoxification (Textual Style Transfer) as a NMT task boosts performance of the models given enough parallel data for training. We beat simple (Delete method) and more strong (condBERT) baselines in a number of metrics. Based on our experiments, we can assume that it is possible to fine-tune multilingual models in any of the 100 languages in which they were originally trained. This opens up great opportunities for detoxification in unpopular languages.

However, our hypothesis that multilingual language models are capable of cross-lingual detoxification was proven to be false. We suggest that the reason for this is not a lack of data, but the model's inability to capture the pattern between toxic and non-toxic text and transfer it to another language by itself. This means that the problem of cross-lingual textual style transfer is still open and needs more investigation.

[10] https://data.statmt.org/news-crawl/

Acknowledgements

This work was supported by MTS-Skoltech laboratory on AI.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel R. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3199–3216. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7979–7996. Association for Computational Linguistics.

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

Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021b. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. In *Proceedings of the 2nd Crowd*

Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/)), pages 35–49, Copenhagen, Denmark. CEUR Workshop Proceedings.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018a. Fighting offensive language on social media with unsupervised text style transfer.

Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018b. 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

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.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2022. Multilingual pre-training with language and task adaptation for multilingual text style transfer. *CoRR*, abs/2203.08552.

Leo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1442–1461. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Mingzhe Li, Xiuying Chen, Min Yang, Shen Gao, Dongyan Zhao, and Rui Yan. 2021. The style-content duality of attractiveness: Learning to write eye-catching headlines via disentanglement. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13252–13260. AAAI Press.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

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.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. 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 3536–3553. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Alan W. Black, and Ruslan Salakhutdinov. 2018. Style transfer through multilingual and feedback-based back-translation. *CoRR*, abs/1809.06284.

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.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 194–205. ACM / IW3C2.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019a. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11034–11044.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019b. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3571–3576. Association for Computational Linguistics.

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 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, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing.

Haoran Xu, Sixing Lu, Zhongkai Sun, Chengyuan Ma, and Chenlei Guo. 2021. VAE based text style transfer with pivot words enhancement learning. *CoRR*, abs/2112.03154.

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.

Zonghai Yao and Hong Yu. 2021. Improving formality style transfer with context-aware rule injection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1561–1570. Association for Computational Linguistics.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3221–3228. Association for Computational Linguistics.

A Data

A.1 English Dataset

Table 4 shows examples of sentence pairs from ParaDatex parallel detoxification corpora. There are several polite paraphrases for each toxic sentence in this dataset (Dementieva et al., 2021b), this is a consequence of the way these parallel data are collected. Leaving only one paraphrase for one source sentence we could get 6000 unique pairs of toxic sentences and their polite paraphrases. However, in this work we use data as is.

Original	my computer is broken and my phone too!! wtf is this devil sh*t???
Detoxed	My computer is broken and my phone too! So disappointed! My computer is broken and my phone too, what is this? Both my computer and phone are broken.
Original	sh*t is crazy around here
Detoxed	It is crazy around here. Stuff is crazy around here. Something is crazy around here.
Original	delete the page and shut up
Detoxed	Delete the page and stay silent. Please delete the page. Delete the page.
Original	massive and sustained public pressure is the only way to get these b*stards to act.
Detoxed	Massive and sustained public pressure is the only way to get them to act. Massive and sustained pressure is the only way to get these people to act.
Original	f*ck you taking credit for some sh*t i wanted to do
Detoxed	You are taking credit for something I wanted to do You're taking credit fro something i wanted to do.
Original	you gotta admit that was f*ckin hilarious though!
Detoxed	you got to admit that was very hilarious though! you gotta admit that was hilarious though!

Table 4: Example sentences from ParaDetox parallel detoxification corpora. Sentence in red is original (toxic) sentence, below are its polite paraphrases. Note that for the purpose of an overall correctness explicit words are masked with "*".

A.2 Russian Dataset

Table 5 shows examples from Russian parallel detoxification corpus.

Original	Х*рню всякую пишут,из-за этого лайка.v Долбо**изм.
Translation	*They write all sorts of bullshit, because of this like. Stupidity.*
Detoxed	Чушь всякую пишут, из- за этого лайка.
Translation	*They write all sorts of nonsense, because of this like.*
Original	А нахрена тогда ты здесь это писал?
Translation	*Why the f*ck did you post it here?*
Detoxed	Зачем ты это писал?
Translation	*Why did you post it?*
Original	Е*анутые. Отобрать оружие и лодку и штраф тысяч 500
Translation	*F*ckers. Take away weapons, boat and give a fine of 500 thousand.*
Detoxed	Отобрать оружие и лодку и штраф тысяч 500.
Translation	*Take away weapons, boat and give a fine of 500 thousand.*
Original	Не поверишь, сколько е**нутых на планете.
Translation	*You won't believe how many f*cked up people are on the planet.*
Detoxed	Не поверишь сколько таких на планете.
Translation	*You won't believe how many people like that are there on the planet.*

Table 5: Example sentences from Russian parallel detoxification corpora. Sentence in red is original (toxic) sentence, below are its polite paraphrases.

B Generation Examples

Table 6 contains detoxification examples for different models. It is notable that in some cases models generate almost the same results. This can be explained by the similarity of the training procedure and the fact that the reference answer was the same.

Original	After all its hard to get a job if you're stupid.
mBART 5000 ENG	After all its hard to get a job if you don't understand.
mBART 3000 ENG	After all its hard to get a job if you are not smart.
mT5 base (multilingual)	It's hard to get a job if you don't know.
Original	Fuck minimum security, put him in real prison.
mBART 5000 ENG	Put him in a real prison with minimum security.
mBART 3000 ENG	I don't care about minimum security, put him in real prison.
mT5 base	I don't care about minimum security, put him in real prison.
Original	Sure he should get punished ,but three f*cking years?
mBART 5000 ENG	Sure he should get punished, but three years?
mBART 3000 ENG	He should get punished, but three years?
mT5 base	He'd get punished, but three years?
Paraphrase	Что у этих людей в башке!? походу насрато!
Translation	*What's going on in these people's heads!? It looks like shit!*
mBART 5000 ENG	Что у этих людей в голове!? Походу бесчеловечно.
mBART 3000 ENG	Что у этих людей в голове? Походу ненормально!
mT5 base	походу этих людей!? походу!

Table 6: Some detoxified sentences produced by our fine-tuned models. Gray text refers to the original sentence, below are its paraphrases.