Reproducing the Metric-Based Evaluation of a Set of Controllable Text Generation Techniques

Michela Lorandi, Anya Belz
ADAPT Centre, Dublin City University, Ireland
{michela.lorandi, anya.belz}@adaptcentre.ie

Abstract

Rerunning a metric-based evaluation should be more straightforward, and results should be closer, than in a human-based evaluation, especially where code and model checkpoints are made available by the original authors. As this report of our efforts to rerun a metric-based evaluation of a set of single-attribute and multiple-attribute controllable text generation (CTG) techniques shows however, such reruns of evaluations do not always produce results that are the same as the original results, and can reveal errors in the reporting of the original work.

Keywords: Reproduction, metric evaluation, controllable text generation.

1. Introduction

Over the past few years, the fields of natural language processing (NLP) and machine learning (ML) have seen an increase in interest in reproducibility (Sinha et al., 2020; Branco et al., 2020; Belz et al., 2021; Belz and Thomson, 2023). Initially, efforts focussed on promoting and encouraging sharing of all resources needed to rerun experiments, but increasingly it became clear that exact reproduction of results is rarely the outcome even where metric evaluation is concerned. The question is what can be concluded in such situations beyond binary reproduced vs. not reproduced findings.

Belz et al. (2022; 2023) proposed QRA++, an approach to measuring how close results from two evaluations are, and how reproducible evaluation measures are, in order to facilitate comparison in terms of degree of reproducibility between different methods of evaluation. This approach enables comparable, quantified reproducibility results to be produced.

In this short report, we present our work rerunning the metric-based evaluation of a set of single and multiple-attribute controllable text generation techniques (Gu et al., 2022, 2023). In the case of all except one pair of scores from the original and reproduction evaluations, the two scores are not the same, and we apply QRA++ to quantify the differences.

We start with a summary of the QRA++ measures we use (Section 2), followed by a description of the specific original experiments we repeated in this reproduction study (Section 3). We then describe how we went about repeating the work (Section 4), before presenting the side-by-side results from the original work and our reproduction along with the QRA++ measures of their similarity (Section 5). We finish with some discussion and conclusions (Section 6).

2. QRA++ Measures

QRA++ distinguishes four types of results commonly reported in NLP and ML papers:

1. Type I results: single numerical scores, e.g. mean quality rating, error count, etc.
2. Type II results: sets of related numerical scores, e.g. set of Type I results .
3. Type III results: categorical labels attached to text spans of any length.
4. Type IV results: Qualitative findings stated explicitly or implied by quantitative results in the original paper.

The above are quantitatively assessed as follows:

1. Type I results: Small-sample coefficient of variation CV* (Belz, 2022).
2. Type II results: Pearson's r, Spearman's ρ.
3. Type III results: Multi-rater: Fleiss's κ; Multi-rater, multi-label: Krippendorff's α.
4. Type IV results: Proportion of findings that are / are not confirmed by the repeat experiment. To obtain comparable results we restrict ourselves to pairwise system ranks as findings.

In the work reported in this paper we have Type I, II and IV results, and therefore apply the corresponding quantitative measures above.

3. Original Work Being Repeated

In the present reproduction study, we carried out repeat evaluations of the main new systems presented by Gu et al. (2022) and Gu et al. (2023). The authors provide the code on GitHub[1] and the model checkpoints on Google Drive.[2]

[1] https://github.com/HappyGu0524/MultiControl

[2] https://drive.google.com/drive/folders/14XHSG4IAG1AL9t-SYoTUKnAs5ARqHd5f

More precisely, the experimental grid we reproduced looks as follows: {PriorCTG} x {Topic (World, Sports, Business, and Technology), Sentiment (Positive and Negative), Toxicity (Toxic and Non-Toxic)} x {no extension, extension} + {PriorCTG} x {Multi attribute (Topic, Sentiment and Non-Toxic)} x {no optim, optim} + {MultiCTG} x {Multi attribute (Topic, Sentiment and Non-Toxic)}. The individual systems (MultiCTG, PriorCTG +/- extend/optim) are described in the next section.

3.1. Systems included in reproduction

We included the results for the four main new systems from the original work (Gu et al., 2022, 2023) in our reproduction study; we abbreviate system names as follows: MultiCTG, PriorCTG, PriorCTG+extend, and PriorCTG+optim.

MultiCTG: This is the core new CTG approach proposed by Gu et al. (2022) which directly searches for the intersection areas of multiple attribute distributions to achieve control over multiple control attributes. The attribute space is first estimated with an autoencoder structure, then the intersections are iteratively approached via joint minimisation of distances to points representing the controlled attributes.

PriorCTG: This is the core new CTG approach proposed by Gu et al. (2023), which utilises a form of latent-space control, more specifically an invertible transformation function, the Normalizing Flow, that maps the complex distributions in latent space to simple Gaussian distributions in **prior** space.

PriorCTG+extend: The **extend** control strategy additionally achieves opposite control, as in contrastive learning, by using negative weights when interpolating.

PriorCTG+optim: The **optim** control strategy additionally optimises the intersection of the single attribute representations in prior space to achieve multiple-attribute control.

All systems are trained on the IMDb movie reviews dataset (Maas et al., 2011), the AGNews dataset (Zhang et al., 2015), and the Jigsaw Toxic Comment Classification Challenge Dataset (cjadams, 2017), respectively, for control of sentiment, topic and detoxification attributes. Note that we did not include any of the baseline systems in the reproduction.

3.2. Evaluation metrics

The metrics in this section are all described in detail in Gu et al. (2022). The main set of metrics assesses single-attribute control performance (called 'attribute relevance' in the original papers), computed as the percentage of outputs that are classified as having the given intended control attribute value by a specific classifier.

For *Sentiment* control performance, the classifier is DeBERTa (He et al., 2020) finetuned on the Yelp dataset (Zhang et al., 2015).

For *Topic* control performance, the classifier is DeBERTa finetuned on the AGNews dataset (Zhang et al., 2015) utilizing the portion of dataset not used during the model's training.

For *Toxicity* control performance, there is a discrepancy between what the paper says and what is in the evaluation script shared on GitHub. According to the former, toxicity is measured with the Google Perspective API.[3] However, the script uses a toxicity classifier obtained by finetuning DeBERTa on the Jigsaw Toxic Comment Classification Challenge Dataset,[4] analogous to control performance assessment for the other control attributes. We ran the evaluation both with Perspective and with the DeBERTa classifier, and found that scores obtained with the latter were closer to the original scores, so those are what we used.

Multiple-attribute control performance is computed as the average of the single-attribute control performance scores for the three attributes being controlled.

Perplexity is calculated by GPT2-large following the Contrastive Prefix method (Qian et al., 2022). Note that we used our own implementation as no code was shared for this.

Distinctness (Li et al., 2016) is computed as the percentage of distinct n-grams in the continuations generated from a given set of prefixes. System-level 1-gram, 2-gram, and 3-gram distinctness scores are obtained by averaging over prefix-level distinctness scores. In multi-control setting, the average of system-level Distinct-1, 2 and 3 is computed. Here too we used our own implementation based on Yu et al. (2021) implementation, because the code was not shared either by Li et al. or by Gu et al.

This gives us six main types of metrics (the three classifier-based metrics, their average (for multiple-attribute control), perplexity, and distinctness). In Table 1 we additionally give the average over the individual control performance scores (**Avg.** columns) for sentiment, topic and toxicity.

4. Reproduction Work

Our first step was to download the code and model checkpoints from the authors' Github and Drive repositories, and recreate the environments on our machine with a GPU RTXA6000 with 48GB RAM.

We then re-executed the inference phase of the experiments involving PriorCTG from Gu et al. (2023), first those with single-attribute control, i.e.

[3] https://www.perspectiveapi.com/

[4] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/

Methods	Sentiment↑ (%)			Topic↑ (%)					Detox.↑	PPL.↓	Dist.-1/2/3↑ (%)
	Avg.	Pos.	Neg.	Avg.	W.	S.	B.	T.	(%)		
PriorCTG	97.1	99.9	94.3	95.9	95.5	99.3	90.2	98.7	90.7	61	42.0 / 79.7 / 88.4
PriorCTG Repro	98.2	99.9	96.6	94.8	93.4	97.8	88.5	99.5	96.9	59.7	41.9 / 79.5 / 88.4
PriorCTG+extend	99.7	99.9	99.5	97.8	97.9	99.4	94.0	99.8	95.7	61.6	42.4 / 79.4 / 88.1
PriorCTG+extend Repro	99.3	99.9	98.7	98.2	98.2	99.5	95.5	99.8	99.9	60.8	42.3 / 79.2 / 88.1

Table 1: Side-by-side metric results from original work (Gu et al., 2023) and reproduction study for **single-attribute control** (last two rows in Table 1 in the original paper). The results of the last two columns are obtained using our own implementation. For PriorCTG and PriorCTG+extend systems (see Section 3). Repro=Reproduction results.

Methods	Average↑ (%)	Sentiment↑ (%)	Topic↑ (%)	Detoxification↑ (%)	PPL.↓	Dist.↑ (%)
MultiCTG	87.4 ± 10.9	86.7 ± 10.5	84.8 ± 14.2	90.7 ± 7.4	31.3	59.0
MultiCTG Repro	88.4 ± 8.3	84.9 ± 11.5	84.5 ± 14.4	95.9 ± 5.5	31.5	59.2
PriorCTG	89.9 ± 8.7	88.0 ± 10.6	87.4 ± 8.5	94.3 ± 3.2	38.9	65.3
PriorCTG Repro	91.1 ± 6.7	88.0 ± 10.2	87.1 ± 11.2	98.3 ± 1.6	38.3	65.2
PriorCTG+optim	92.2 ± 8.6	92.5 ± 8.5	89.3 ± 11.0	94.9 ± 3.4	33.0	61.7
PriorCTG+optim Repro	93.2 ± 7.2	91.8 ± 9.7	89.3 ± 12.4	98.6 ± 1.1	32.5	62

Table 2: Side-by-side metric results from original work (Gu et al., 2022, 2023) and reproduction study for **multiple-attribute control**. Results for MultiCTG are from the third to last row in Gu et al. (2022). Original results for the other two systems are from the last two rows in Table 3 in Gu et al. (2023). The results of the last two columns are obtained using our own implementation. For system and metrics descriptions see Section 3). Repro=Reproduction results.

where Topic, Sentiment or Toxicity are being controlled individually, and then those with multiple-attribute control, where Topic, Sentiment and Toxicity are being controlled at the same time.

For multiple-attribute control we also re-executed the inference phase of the experiments involving MultiCTG from Gu et al. (2022). This gave us sets of 35 × 5 = 175 outputs (35 inputs from the PPLM Prompts test set × 5 repetitions of prompting and collecting the outputs) for each system/attribute combination.

Note that as in the original work, outputs are generated for all values of all controlled attributes (single-attribute case) or for all combinations of controlled attribute values (multiple-attribute case), results for all of which except Toxicity=toxic ('Detox(ification)' in the tables) are reported in the results tables. In the multiple-attribute case, the average over different attribute value combinations, along with the corresponding standard deviation, is reported.

For the evaluation, we computed the metrics listed in Section 3. Recall from Section 3.2 that we used the script provided by the authors for Sentiment, Topic and Toxicity control performance assessment. However, we coded our own scripts to compute Perplexity and Distinct-n, as scripts are not provided for these. We also use our own code for the standard deviations in the multiple-attribute table. For all scripts we use parameters as provided by the authors.

Note that as a result of some of the evaluation scripts not being shared, we have two distint reproduction situations (which in QRA++ is reflected in the measurement conditions): (a) for the classifier-based control-performance measures, we use our outputs (regenerated by us using the original authors' code) and evaluate them with the original authors' scripts; and (b) for perplexity and distinctness, we use our outputs *and* our evaluation scripts. In the former case differences in scores can only be due to differences in *executing* the original authors' code, whereas in the latter case, differences can be due to both execution and differences in the evaluation code.

In order to avoid this dual possible source of difference for perplexity and distinctness scores, we decided to re-evaluate the original authors' outputs with our own script. This means that the scores in our tables are not the same as in the two original papers for these two metrics. But it means CV* scores and other reproducibility measures are comparable across all metrics.

5. Side-by-Side Results and QRA++ Assessment

Tables 1 and 2 present side-by-side evaluation results for the original and reproduction work, for each of the six metrics from Section 3, plus, in Table 1 only, averages over individual control performance scores (**Avg.** columns). Reall that we reevaluated the original authors' outputs in terms of Perplexity and Distinctness (see preceding section).

System	*CV* between original and reproduction scores for each evaluation measure*												
	Sent avg	Sent pos	Sent neg	Topic avg	Topic W	Topic S	Topic B	Topic T	Detox	PPL	Dist-1	Dist-2	Dist-3
Prior-CTG	1.12	0	2.4	1.15	2.22	1.52	1.9	0.8	6.59	2.15	0.24	0.25	0
Prior-CTG+ext	0.4	0	0.8	0.41	0.31	0.1	1.58	0	4.28	1.3	0.24	0.25	0
Average	0.76	0	1.6	0.78	1.27	0.81	1.74	0.4	5.44	1.725	0.24	0.25	0

Table 3: CV* for each pair of original and reproduction metric scores, for the Prior-CTG and Prior-CTG+extend systems, and the average over both systems.

System	*CV* between original and reproduction scores for each evaluation measure*					
	Avg	Sentiment	Topic	Detox	PPL	Distinct-n
Multi-CTG	1.13	2.09	0.35	5.56	0.64	0.34
Prior-CTG	1.32	0.0	0.34	4.14	1.52	0.15
Prior-CTG+optim	1.08	0.76	0.0	3.81	1.52	0.48
Average	1.18	0.95	0.23	4.5	1.23	0.32

Table 4: CV* for each pair of original and reproduction metric scores, for the Multi-CTG, Prior-CTG and Prior-CTG+optim systems, and the average over all three.

5.1. Type IV results

Regarding Type IV results (findings), here we are assessing relative performance between systems, such that each pairwise ranking counts as one finding. Note that statistical significance was not computed in the original work.

For single-attribute control (Table 1), in the original work, Prior CTG+extend has higher scores than PriorCTG according to all metrics except for Perplexity and 2-gram and 3-gram Distinctness where PriorCTG scores are very slightly higher. For Sentiment/Pos, scores are identical. In our reproduction evaluations, these two systems are ranked the same way in all cases, giving us a perfect proportion of 13/13 findings upheld for this table.

For multiple-attribute control (scores in Table 2), the same type of analysis gives us a proportion of 18/18 findings upheld (pairwise ranks confirmed).

5.2. Type I results

For Type I results, we computed CV* values for all individual system/metric level original and reproduction scores. We report the individual scores, as well as the mean per metric.

For single-attribute control (scores in Table 1), Table 3 shows CV* scores for each pair of original and reproduction metric scores, for the Prior-CTG and Prior-CTG+extend systems, and the average over both systems (last row).

One clear tendency is that the Prior-CTG system has better reproducibility scores across the board than Prior-CTG+extend (except for distinctness metrics where the two systems are tied).

Looking at metric-level differences ('Average' row), we can see that Perplexity and (by a smaller margin) Detoxification Control have lower reproducibility than the other metrics.

For multiple-attribute control (scores in Table 2), Table 4 shows CV* scores for each pair of original and reproduction metric scores, for the Multi-CTG, Prior-CTG and Prior-CTG+optim systems, and the average over all three (last row). We can see that here too, the Perplexity and Detoxification Control metrics have the poorest reproducibility.

We can also see a slight tendency for the classifier scores for the Prior-CTG+optim system to have better reproducibility than the other two systems (but not for PPL and Distinct-n), but the picture is more mixed than for the single-attribute control systems.

5.3. Type II results

For Type II results we compute Pearson's correlation coefficients between sets of metric scores in two ways, (i) for each metric (i.e. how do all the scores for each metric correlate between original and reproduction), and (ii) for each system (i.e. how do all the scores for each system correlate).

For single-attribute control (scores in Table 1), system-level Pearson's between all metric results in the original and reproduction runs is above 0.99 for both Prior-CTG and Prior-CTG+extend. Mean metric-level Pearson's is perfect (but note that we have only two score pairs all of which are ranked identically).

For multiple-attribute control (scores in Table 2), system-level Pearson's between all metric results in the original and reproduction runs is above 0.99 for all three systems. Metric-level Pearson's is above 0.99 for all metrics except the sentiment-classifier metric which at $r =$ 0.969 is slightly lower than the other metrics. Mean metric-level r is 0.994.

6. Discussion and Conclusion

The main challenges in carrying out our reproduction study were (i) lack of clarity in the paper with respect to what the averages and standard deviations in results tables were computed over, and (ii) discrepancies between the shared code and what the paper said, e.g. the paper says toxicity was assessed with Perspective, whereas the shared evaluation script has a toxicity classifier.

Our quantified reproducibility assessments revealed a high degree of reproducibility at the study level for Type II and Type IV results. For Type I results, study-level CV* (computed as the mean of metric-level means) was 1.154 for single-attribute control, and 1.402 for multiple-attribute control. While this compares well to reproducibility results in human evaluations which very rarely achieve study-level CV* below 5 in pairwise comparisons of original study and one reproduction, it does confirm once again that even with identical code, we cannot necessarily expect to get the same results.

In terms of metric-level CV*, the Detoxification control metric had notably worse reproducibility than the others which may be partly but not entirely explainable by the fact that only Toxicity=nontoxic was taken into account here.

In terms of the results that tend to be considered as most important, Type IV results or findings upheld, reproducibility was perfect with all pairwise rankings being identical in the original and reproduction experiments.

Bibliographical References

Anya Belz. 2022. A metrological perspective on reproducibility in nlp. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393.

Anya Belz and Craig Thomson. 2023. The 2023 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter Van Uytvanck, Joao Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with reprolang2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545. European Language Resources Association (ELRA).

Julia Elliott Lucas Dixon Mark McDonald nithum Will Cukierski cjadams, Jeffrey Sorensen. 2017. Toxic comment classification challenge.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Koustuv Sinha, Joelle Pineau, Jessica Forde, Rosemary Nan Ke, and Hugo Larochelle. 2020. Neurips 2019 reproducibility challenge. 6.

Kirstie Whitaker. 2017. The MT Reproducibility Checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2251–2268, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A. Perplexity and Distinct-n implementation

No code was shared to compute perplexity and Distinct-n, hence we used our own implementation. Perplexity is calculated using the evaluate library of HuggingFace[5] using GPT-2 Large.

System-level Distinct-n (n=1, 2, 3) is the average Distinct-n at prefix-level, which is computed as the number of unique n-grams in the set of generated outputs with the same prefix over the total amount of tokens. GPT-2 is used to tokenise the texts.

Table 5 and 6 show Perplexity and Distinct-n results reported in the original work, the results of the original study computed using our implementation and the reproduction using our implementation.

[5] https://huggingface.co/docs/evaluate/en/index

Methods	**PPL.** (%) ↓	**Dist.-1/2/3**↑ (%)
PriorCTG	54.3	29.1 / 70.1 / 86.9
PriorCTG using our implementation	61	42.0 / 79.7 / 88.4
PriorCTG Repro	59.7	41.9 / 79.5 / 88.4
PriorCTG+extend	54.6	29.8 / 70.5 / 86.8
PriorCTG+extend using our implementation	61.6	42.4 / 79.4 / 88.1
PriorCTG+extend Repro	60.8	42.3 / 79.2 / 88.1

Table 5: Side-by-side metric results from original work (Gu et al., 2023), original work (Gu et al., 2023) computed using our own implementation and reproduction study using our own implementation for **single-attribute control** (last two rows in Table 1 in the original paper). For PriorCTG and PriorCTG+extend systems (see Section 3). Repro=Reproduction results.

Methods	**PPL.**↓	**Dist.**↑ (%)
MultiCTG	28.4	49.5
MultiCTG using our implementation	31.3	59.0
MultiCTG Repro	31.5	59.2
PriorCTG	34.7	55.5
PriorCTG using our implementation	38.9	65.3
PriorCTG Repro	38.3	65.2
PriorCTG+optim	29.6	51.6
PriorCTG+optim using our implementation	33.0	61.7
PriorCTG+optim Repro	32.5	62

Table 6: Side-by-side metric results from original work (Gu et al., 2022, 2023), original work (Gu et al., 2022, 2023) computed using our own implementation and reproduction study using our own implementation for **multiple-attribute control**. Results for MultiCTG are from the third to last row in Gu et al. (2022). Original results for the other two systems are from the last two rows in Table 3 in Gu et al. (2023). For system and metrics descriptions see Section 3). Repro=Reproduction results.