A Call for Standardization and Validation of Text Style Transfer Evaluation

Phil Ostheimer, Mayank Nagda, Marius Kloft, Sophie Fellenz
surname@cs.uni-kl.de
RPTU Kaiserslautern-Landau

Abstract

Text Style Transfer (TST) evaluation is, in practice, inconsistent. Therefore, we conduct a meta-analysis on human and automated TST evaluation and experimentation that thoroughly examines existing literature in the field. The meta-analysis reveals a substantial standardization gap in human and automated evaluation. In addition, we also find a validation gap: only few automated metrics have been validated using human experiments. To this end, we thoroughly scrutinize both the standardization and validation gap and reveal the resulting pitfalls. This work also paves the way to close the standardization and validation gap in TST evaluation by calling out requirements to be met by future research.

1 Introduction

Text style transfer (TST) is the task of transferring text from one style to another. Examples include sentiment transfer (making a negative text more positive) (Shen et al., 2017), politeness transfer (Niu and Bansal, 2018), formality transfer (Rao and Tetreault, 2018), and many more. Anyone working and communicating with texts has most likely manually performed many of these transfer tasks countless times. Whether generalized (Reif et al., 2022) or task-specific models, evaluating their performance on TST is crucial to measure progress in task-specific text generation. In the last few years, there has been a surge in research on TST (see Figure 1, Appendix), the majority of which was in the last two years. When a field develops that quickly, the development of evaluation procedures often does not keep up.

TST research is now at a point where reflection on evaluation practices is becoming increasingly urgent, as progress in this field is becoming difficult if the performance of new methods cannot be reliably compared to existing methods. It can even become hard to determine the state of the art one should aim to improve. However, recent TST surveys (Toshevska and Gievska, 2022; Jin et al., 2022; Hu et al., 2022) focus on summarizing TST methods and their evaluation but do not point out evaluation and experimentation issues. Works on TST evaluation either focus on getting more reliable metrics for a particular type of TST (Mir et al., 2019), focus on a specific evaluation aspect for validation (Yamshchikov et al., 2021), or emphasize only shortcomings of automated evaluation in a multilingual setting of one particular type of TST (Briakou et al., 2021a). As a first step, it is, therefore, necessary to examine evaluation practices and experimental setups to point out the need for standardization of both, which is what we set out to do in this paper.

Early TST works (Shen et al., 2017; Li et al., 2018) focus on human evaluation, which is still considered the most reliable (Briakou et al., 2021b). Many later publications rely only on automated metrics, which can be validated by correlating them with human evaluations. So far, no comprehensive overview exists showing which metrics are validated and how leading to researchers using popular metrics rather than those with the highest correlations. In our meta-analysis, we counter-act this by reporting which metrics are validated and analyzing cases where the same metric received different validation results. To sum up, our contributions are:

1. Our examination of TST methods and TST evaluation practices highlights the need for standardization (Section 2).

2. Our overview of automated TST evaluation reveals an unmet need for validation (Section 3).

We conclude with a list of requirements to help close these two gaps.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10791–10815

2 A Meta-Analysis of Text Style Transfer

In this meta-analysis, we focus on TST publications in top-tier NLP and AI venues (see Appendix A.2 for selection details), resulting in 89 considered papers summarized in Table 7, 8, and 9. The dimensions of TST evaluation are fluency, content preservation, and style transfer strength, as well as other aspects. Therefore, we divide the automated and human evaluation subsections according to these aspects.

2.1 Automated Evaluation

For automated evaluation, we aggregate the findings of our meta-analysis in Table 1 (with a detailed description of conventions in Appendix A.4.1). Overall, 21/89 papers use only automated evaluation, and 33/89 use at least one metric that has not been validated. Statistically significant results are reported for only 3/89 papers.

Aspect	Count
No human evaluation	21/89
>1 non-validated metric	33/89
Statistical significance	3/89
Metrics	
Fluency	34
Content preservation	35
Style transfer strength	23
Other	24
Papers	
Fluency	45/89
Content preservation	66/89
Style transfer strength	77/89
Other	34/89

Table 1: Shown is the variety of automated TST evaluation metrics. All examined TST papers deploy at least one. However, using non-validated metrics is a common phenomenon, and statistical significance is reported only in a fraction of the investigated papers.

Fluency We find a total of 34 different setups to measure fluency automatically. 45/89 papers rely on automated metrics for measuring fluency. The by far most common metric (see Table 4) is the Perplexity (PPL) of a Language Model (LM). However, the pre-training procedure of the LM varies. It is either trained on all styles of a dataset (dos Santos et al., 2018; Dai et al., 2019; John et al., 2019; Cheng et al., 2020), trained on each style separately (Yang et al., 2018; Gong et al., 2019), pre-trained on external datasets (Logeswaran et al., 2018; Jain et al., 2019), or trained on the dataset and the TST output (backward/forward) (Zhao et al., 2018; Huang et al., 2020). 19/89 papers do not mention how the LM is trained. These numbers show the need for consistency.

Content Preservation We find 66 papers using a total of 35 different automated metrics to measure content preservation. Only 15 of these metrics have been validated.

The most common metrics (see Table 5) to measure content preservation are Source-BLEU and Ref-BLEU (applying BLEU (Papineni et al., 2002) to the input/output, reference/output respectively). However, as pointed out by Mir et al. (2019), Source-BLEU only measures n-gram overlaps and does not consider that a change of the sentence is necessary to change the style. This results in contradictory validation results (see Section 3). Post (2018) has shown that the reported BLEU scores heavily depend on preprocessing and several parameters (e.g., number of references, length penalty, maximum n-gram length, smoothing applied to 0-count n-grams). Works using BLEU for evaluation need this specification to make the results consistent and reproducible. In our meta-analysis, we could not find a single TST paper that specifies BLEU sufficiently (including above mentioned details). Ref-BLEU is the second most popular method to measure content preservation. In addition to the general problems with BLEU, we see two more shortcomings. On the one hand, the needed reference sentences are not available for all datasets. On the other hand, calculating BLEU scores between the output and multiple human references can improve its reliability and, thankfully, for the arguably most popular TST dataset—Yelp—Jin et al. (2019) and Luo et al. (2019) introduced additional reference sentences. These, however, are only used by 6/30 papers applying Ref-BLEU on Yelp.

Style Transfer Accuracy For automated evaluation of style transfer accuracy, we find 23 evaluation metrics, of which 14 have been validated. Overall, 77 papers use automated style transfer accuracy metrics. Table 6 provides an overview of the most common automated text style transfer strength metrics. The top two are TextCNN (Kim, 2014) and fastText (Joulin et al., 2017).

Other 34/89 papers measure a fourth aspect. For 29/89 it is an overall metric, and 8/89 measure another aspect of TST. We find eight validated metrics to measure overall performance. None of the metrics for other aspects have been validated.

2.2 Human Evaluation

Aspect	Count
Usage	68/89
Statistical analysis	2/68
Evaluations released	5/68
Underspecified	67/68
No. of evaluation schemes	24

Table 2: Shown are the aggregated insights of human TST evaluation. Despite being widespread, it is far from being standardized and, in most cases, lacks statistical analysis. It has many different setups that are often underspecified, and the annotations are not released.

For human evaluation, we aggregate the findings of our meta-analysis in Table 2 (with a detailed description of our conventions in Appendix A.3). Overall, 68/89 papers conduct a human evaluation. However, only a tiny fraction of 2/68 papers detect a statistically significant difference in model performance. Most papers fail to include a statistical test, be it a statistical power analysis before conducting the human evaluations or a statistical significance test afterward. This is a common problem in the NLP community (Card et al., 2020).

Releasing human evaluations is also relatively uncommon. Only 5/68 papers publish their human experimentation results. However, releasing them would facilitate reproducibility and the validation of new automated metrics. Reproducibility of human TST evaluation is a challenge, as is reproducibility of human evaluation in general (Belz et al., 2020). 67/68 papers conducting human evaluations have no adequate description. We consider the description of human evaluation to be adequate if the following is specified: annotators' background, the number of annotators, the number of annotators per sample, incentivization, data selection, questionnaire design, and rating scale (Briakou et al., 2021b). With published human evaluations, one could also easily estimate the parameters of a simulation-based power analysis as done by Card et al. (2020).

Overall, there is still room for improvement despite previous calls for standardization (Briakou et al., 2021b). For human evaluation, we find a total of 24 different evaluation schemes (viewed on a high level whether each aspect is present and evaluated relatively or absolutely, not considering different rating scales and questionnaire designs).

2.3 Experimentation

Aspect	Count
Multiple runs	5/89
Reproducibility	
Code provided	56/89
Evaluation code provided	42/89
Preprocessing specified	38/89

Table 3: Shown are the aggregated insights of TST experimentation: a lack of reporting multiple runs, hampered reproducibility by the missing provision of code, and the underspecification of the preprocessing pipeline.

We aggregate our meta-analysis' findings on experimentation details in Table 3 (with a detailed description of our conventions in Appendix A.5). In order to make statements about relative model performance, one usually runs the model multiple times with different seeds to be able to conduct a statistical significance test. A TST model's performance can significantly vary between runs (Tikhonov et al., 2019; Yu et al., 2021), indicating the need for reporting multiple runs. However, most (84/89) papers only report a single run.

Reproducing results is difficult since only 56/89 of the reviewed papers release their code. An even smaller fraction provides the complete evaluation code (42/89). Another aspect that can significantly influence the outcome is preprocessing. However, only 38/89 papers specify their preprocessing pipeline.

3 Automated Metrics and Their (Missing) Validation

In this section, we summarize existing automated metrics and their validity in terms of correlation with human evaluations. In Table 10, we give a detailed overview (the first of its kind) and describe our conventions in Appendix B. The most crucial convention is to assume that the validation of an automated metric for a particular TST task also generalizes to other TST tasks.

3.1 Fluency

Fluency is sometimes referred to as grammaticality, readability, or naturalness (Mir et al., 2019). It is commonly quantified by measuring the PPL of an LM on the TST output (40/45 reviewed papers). Mir et al. (2019) claim a limited correlation between sentence-level Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) LM PPL and human fluency evaluations and conclude that LM PPL is an inappropriate metric to measure fluency in the TST setting. On the other hand, Pang and Gimpel (2019) show a high correlation but also note that PPL is not equal to fluency. Briakou et al. (2021a) report a relatively low correlation for the PPL of a 5-gram KenLM (Heafield, 2011) with human evaluations and slightly higher correlations for Pseudo Log-Likelihoods (PLL) of BERT (Devlin et al., 2019), and XLM (Conneau et al., 2020). Overall, LM PPL for TST evaluation has only been validated on a fraction of the deployed TST datasets, namely Yelp[s] and GYAFC. Previous work is divided as to whether and to what extent this metric correlates with human evaluations. As reported in Section 2, many different architectures, training methods, and application methods of LMs for TST fluency evaluation exist. However, as a simplification, we assume the LM PPL to be a validated fluency metric.

Other approaches to evaluate fluency are based on BLEU. In contrast to Luo et al. (2019), Li et al. (2018) show no significant correlation for Ref-BLEU. Pryzant et al. (2020) report a low correlation of the Source-BLEU score with human evaluations. Wu et al. (2020) and Rao and Tetreault (2018) report moderate correlation for their metric (Heilman et al., 2014). Pryzant et al. (2020) show that their style classifier correlates more with human evaluations of fluency than the style transfer strength. None of these metrics is deployed by more than two papers.

3.2 Content Preservation

There are two extensive studies for automated content preservation metrics by Mir et al. (2019) and Yamshchikov et al. (2021). However, both have limited scope: Mir et al. (2019) only report scores where style words have been masked or removed (not done by any other paper). Yamshchikov et al. (2021) report correlations only on the datasets themselves and not on actual model outputs. Both report Word Mover's Distance (WMD) (Kusner et al., 2015) having the highest correlation, outperforming Source-BLEU, other embedding-based metrics (such as also investigated by Fu et al. (2018) and Pang and Gimpel (2019)), and chrF (Popović, 2015).

Cao et al. (2020) find high correlation for Source-BLEU, whereas Pryzant et al. (2020) find low correlation. For Ref-BLEU, Li et al. (2018), Luo et al. (2019), and Cao et al. (2020) show a high, and Xu et al. (2012) a low correlation, whereas Briakou et al. (2021a) investigate Ref-BLEU and Source-BLEU among others, but show chrF having the highest correlation. The suitability of BLEU for TST evaluation remains questionable as Mir et al. (2019) point out that BLEU cannot capture whether words were changed on purpose.

3.3 Style Transfer Strength

Style transfer strength is usually evaluated by applying a sentence classifier trained to classify the output sentences by style. Early work (Xu et al., 2012) compares several metrics showing the highest correlation for Logistic Regression (LR). The two most popular methods nowadays are TextCNN (Kim, 2014) and fastText (Joulin et al., 2017). Luo et al. (2019) show high correlation for TextCNN, Pang and Gimpel (2019) for fastText, whereas Mir et al. (2019) validate both, showing slightly better correlations for TextCNN. Li et al. (2018) validate a Bi-directional LSTM for style classification and note that the correlation to human evaluations highly depends on the dataset. Rao and Tetreault (2018) report a moderate correlation to their style transfer strength metric. Also, Mir et al. (2019) show that the Earth Mover's Distance (EMD) in combination with TextCNN or fastText has a higher correlation with human evaluations than only TextCNN or fastText, but no other paper uses it.

3.4 Other

Niu and Bansal (2018) report a high correlation of Source-BLEU with overall dialogue quality. Rao and Tetreault (2018) report a relatively high correlation for Ref-BLEU compared to TERp and PINC with the overall human evaluations. This is in agreement with Wang et al. (2020). Wu et al. (2020) perform dataset-dependent studies and found no significant correlation between automated metrics and human scores.

3.5 Metrics Validated for Multiple Aspects

Some metrics, such as Ref-BLEU, Source-BLEU, PINC, and embedding-based metrics, are validated for multiple aspects. However, Ref-BLEU shows the highest correlation as an overall metric only for Rao and Tetreault (2018) (also outperforming

PINC), otherwise (Xu et al., 2012; Li et al., 2018; Luo et al., 2019; Cao et al., 2020; Wang et al., 2020; Briakou et al., 2021a), there is no clear picture. For Source-BLEU (Niu and Bansal, 2018; Mir et al., 2019; Pryzant et al., 2020; Cao et al., 2020; Yamshchikov et al., 2021; Briakou et al., 2021a) and also for embedding-based metrics (Xu et al., 2012; Fu et al., 2018; Mir et al., 2019; Pang and Gimpel, 2019; Wu et al., 2020; Yamshchikov et al., 2021; Briakou et al., 2021a), we find mixed results across the different aspects.

4 Conclusion & Future Work

Our research emphasizes the pressing need for standardization and validation in TST. While human evaluation is still considered to be the most reliable, it is expensive and time-consuming, hindering reproducibility. Many publications use automated metrics as surrogates, but only a fraction of them is validated. Furthermore, human and automated evaluations and experimental setups lack standardization, making it difficult to compare the performance of different models. Summarizing our results, we pose the following six requirements to be met by future TST research:

1. It needs to use validated metrics (see Table 10), focusing on those showing the highest correlations with human evaluations

2. In experiments, multiple runs need to be performed on different random seeds, reporting mean and standard deviation.

3. A statistical significance test needs to be performed on the results of automated metrics.

4. If a human evaluation is done, a statistical power analysis is necessary in advance, and all human evaluation details need to be published as suggested by Briakou et al. (2021b).

5. To improve reproducibility, researchers should always specify the preprocessing pipeline and publish their code (including evaluation code).

6. A comparison with state-of-the-art methods on the validated metrics is called for.

To help with these requirements, we plan a large-scale experimental comparison to rank existing methods according to validated metrics in the future and examine how automated metrics generalize from one TST task to the other. We also plan to (re-)validate existing automated metrics to help meet the first requirement.

Limitations

The present work only points out problems of existing research and presents no final solutions. We also simplify the assumption that an automated metric validated for a specific TST task generalizes to other tasks. However, this is problematic since there is, to our knowledge, no investigation of whether a validation on one task generalizes. This concern is motivated by the fundamental differences in how different TST tasks are defined. There are several different definitions, such as data-driven TST (e.g., sentiment transfer) and linguistically motivated TST (e.g., formality transfer) (Jin et al., 2022). Also, we consider only TST papers (no text simplification) and focus on top-tier NLP and AI venues (non-workshop).

Acknowledgements

The authors acknowledge support by the Carl-Zeiss Foundation, the BMBF awards 01|S20048, 03|B0770E, and 01|S21010C, and the DFG awards BU 4042/2-1, BU 4042/1-1, KL 2698/2-1, KL 2698/5-1, KL 2698/6-1, and KL 2698/7-1.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. Reprogen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 232–236. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Joel R. Tetreault, and Marine Carpuat. 2021a. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. 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 1321–1336. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel R. Tetreault, and Marine Carpuat. 2021b. A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 58–67.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel R. Tetreault. 2021c. 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.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1061–1071. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9263–9274. Association for Computational Linguistics.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6455–6469. Association for Computational Linguistics.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 190–200. The Association for Computer Linguistics.

Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. 2018. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4671–4682.

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2915–2924. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111. 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*, pages 8440–8451, Online. Association for Computational Linguistics.

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. European Language Resources Association (ELRA).

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. 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 5997–6007.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In

Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, pages 512–515. AAAI Press.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3286–3292. ACM.

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.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 193–202. ACM.

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.

Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li. 2020. Pre-train and plug-in: Flexible conditional text generation with variational autoencoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 253–262. Association for Computational Linguistics.

Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. 2019. Rethinking text attribute transfer: A lexical analysis. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 24–33. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style Transfer in Text: Exploration and Evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 955–964. IEEE Computer Society.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 340–348. Tsinghua University Press.

Hongyu Gong, Suma Bhat, Lingfei Wu, Jinjun Xiong, and Wen-Mei W. Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. 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 3168–3180. Association for Computational Linguistics.

Navita Goyal, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Abhilasha Sancheti. 2021. Multi-style transfer with discriminative feedback on disjoint corpus. 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 3500–3510. Association for Computational Linguistics.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Trans. Assoc. Comput. Linguistics*, 6:437–450.

Mengqiao Han, Ou Wu, and Zhendong Niu. 2017. Unsupervised automatic text style transfer using LSTM. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 281–292. Springer.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1576–1586. The Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A Probabilistic Formulation of Unsupervised Text Style Transfer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Ruining He and Julian J. McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on*

World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pages 507–517. ACM.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, Edinburgh, Scotland, UK, July 30-31, 2011*, pages 187–197. Association for Computational Linguistics.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor.*, 24(1):14–45.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward Controlled Generation of Text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1587–1596.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2213–2223. International Committee on Computational Linguistics.

Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D. Kelleher. 2020. Style versus content: A distinction without a (learnable) difference? In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2169–2180. International Committee on Computational Linguistics.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6554–6561. AAAI Press.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Comput. Linguistics*, 48(1):155–205.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5082–5093. Association for Computational Linguistics.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised Text Attribute Transfer via Iterative Matching and 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3095–3107. Association for Computational Linguistics.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. 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 6047–6052. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann, and Soujanya Poria. 2022. So different yet so alike! constrained unsupervised text style transfer. 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 416–431. Association for Computational Linguistics.

Heejin Kim and Kyung-Ah Sohn. 2020. How positive are you: Text style transfer using adaptive style embedding. In *Proceedings of the 28th International*

Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pages 2115–2125. International Committee on Computational Linguistics.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pages 181–184. IEEE Computer Society.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

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 2020, Online, November 16-20, 2020*, pages 737–762. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Shibamouli Lahiri. 2014. Complexity of word collocation networks: A preliminary structural analysis. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 96–105. The Association for Computer Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. 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 4241–4254. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y.-Lan Boureau. 2018. Multiple-Attribute Text Rewriting. In *International Conference on Learning Representations*.

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.

Joosung Lee. 2020. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 195–204. Association for Computational Linguistics.

Wouter Leeftink and Gerasimos Spanakis. 2019. Towards controlled transformation of sentiment in sentences. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence, ICAART 2019, Volume 2, Prague, Czech Republic, February 19-21, 2019*, pages 809–816. SciTePress.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text 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 3302–3311. Association for Computational Linguistics.

Jingjing Li, Zichao Li, Lili Mou, Xin Jiang, Michael R. Lyu, and Irwin King. 2020a. Unsupervised text generation by learning from search. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: a Simple Approach to

Sentiment and 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 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li. 2020b. DGST: a dual-generator network for text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7131–7136. Association for Computational Linguistics.

Yuan Li, Chunyuan Li, Yizhe Zhang, Xiujun Li, Guoqing Zheng, Lawrence Carin, and Jianfeng Gao. 2020c. Complementary auxiliary classifiers for label-conditional text generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8303–8310. AAAI Press.

Yi Liao, Lidong Bing, Piji Li, Shuming Shi, Wai Lam, and Tong Zhang. 2018. Quase: Sequence editing under quantifiable guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3855–3864. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ao Liu, An Wang, and Naoaki Okazaki. 2022. Semi-supervised formality style transfer with consistency training. 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 4689–4701. Association for Computational Linguistics.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8376–8383. AAAI Press.

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.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. 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 4262–4273. Association for Computational Linguistics.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5108–5118.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5116–5122. ijcai.org.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard H. Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Styleptb: A compositional benchmark for fine-grained controllable text 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 2116–2138. Association for Computational Linguistics.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7426–7441. Association for Computational Linguistics.

Yun Ma, Yangbin Chen, Xudong Mao, and Qing Li. 2021. Collaborative learning of bidirectional decoders for unsupervised text style transfer. 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 9250–9266. Association for Computational Linguistics.

Yun Ma and Qing Li. 2021. Exploring non-autoregressive text style transfer. 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 9267–9278. 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 *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June,* *2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1869–1881. Association for Computational Linguistics.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8671–8680. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics.

Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pages 1020–1025. IEEE Computer Society.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating Style Transfer for Text. 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 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick. 2021. Style pooling: Automatic text style obfuscation for improved classification fairness. 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 2009–2022. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Jonas Mueller, David K. Gifford, and Tommi S. Jaakkola. 2017. Sequence to Better Sequence: Continuous Revision of Combinatorial Structures. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2536–2544. PMLR.

Courtney Napoles, Matthew R Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Trans. Assoc. Comput. Linguistics*, 6:373–389.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised Evaluation Metrics and Learning Criteria for Non-Parallel Textual Transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 138–147, Hong Kong. Association for Computational Linguistics.

Kartikey Pant, Yash Verma, and Radhika Mamidi. 2020. Sentiinc: Incorporating sentiment information into sentiment transfer without parallel data. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 312–319. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8*

September, 2016, volume 1609 of *CEUR Workshop Proceedings*, pages 750–784. CEUR-WS.org.

Ellie Pavlick and Joel R. Tetreault. 2016. An empirical analysis of formality in online communication. *Trans. Assoc. Comput. Linguistics*, 4:61–74.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 866–876.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 480–489. AAAI Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sudha Rao and Joel R. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science, NLP+CSS@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 17–26. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A Recipe For Arbitrary Text Style Transfer with Large Language Models. *arXiv:2109.03910 [cs]*. ArXiv: 2109.03910.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. 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 815–825. Association for Computational Linguistics.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer Science*, pages 545–560. Springer.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205. AAAI.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. 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 4936–4945. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 1633–1650. USENIX Association.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12516–12526. Computer Vision Foundation / IEEE.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 91–101. ACL.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "Transforming" Delete, Retrieve, Generate Approach for Controlled Text 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 3267–3277. Association for Computational Linguistics.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9008–9015. AAAI Press.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. 2019. Style Transfer for Texts: Retrain, Report Errors, Compare with Rewrites. 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 3934–3943. Association for Computational Linguistics.

Martina Toshevska and Sonja Gievska. 2022. A review of text style transfer using deep learning. *IEEE Trans. Artif. Intell.*, 3(5):669–684.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2107–2114. International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. 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).

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 Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2236–2249. International Committee on Computational Linguistics.

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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019a. A hierarchical reinforced sequence operation method for unsupervised 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 4873–4883. Association for Computational Linguistics.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Yu Wu, Yunli Wang, and Shujie Liu. 2020. A dataset for low-resource stylized sequence-to-sequence generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9290–9297. AAAI Press.

Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive learning for unsupervised text style transfer. 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 2510–2521. Association for Computational Linguistics.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 979–988. Association for Computational Linguistics.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10534–10543. PMLR.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Trans. Assoc. Comput. Linguistics*, 4:401–415.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2899–2914. Indian Institute of Technology Bombay.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric. 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 14213–14220. AAAI Press.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised Text Style Transfer using Language Models as Discriminators. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3801–3807. ijcai.org.

Ping Yu, Yang Zhao, Chunyuan Li, and Changyou Chen. 2021. Rethinking sentiment style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1569–1582. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Ye Zhang, Nan Ding, and Radu Soricut. 2018a. SHAPED: shared-private encoder-decoder for text style adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1528–1538. Association for Computational Linguistics.

Yi Zhang, Tao Ge, and Xu Sun. 2020b. 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.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018b. Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1103–1108. Association for Computational Linguistics.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially Regularized Autoencoders. In *International Conference on Machine Learning*, pages 5902–5911. PMLR. ISSN: 2640-3498.

Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7135–7144. Association for Computational Linguistics.

A Meta-Analysis

This section describes the details of the conducted meta-analysis on TST evaluation.

A.1 Popularity of TST

TST has become a popular topic in research. When Searching for "text style transfer" on `dblp.uni-trier.de` (accessed Jan 18, 2023), we can find a steep increase in the number of publications per year (see Figure 1).

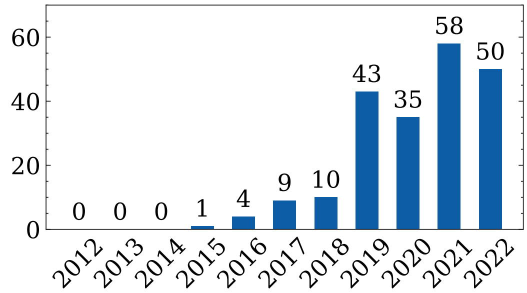

Figure 1: The number of TST publications has steeply increased in the last ten years.

A.2 Paper Selection

We consider the papers listed in a recent TST survey (Jin et al., 2022) [1] with the addition of more recent publications. We consider only TST papers (no text simplification) and focus on top-tier NLP and AI venues (non-workshop) with the following statistics: ACL:25, EMNLP:17, NAACL:9, COLING:8, AAAI:7, NeurIPS:5, ICML:4, ICLR:3, IJCAI:3, ECIR:2, EACL:1, CVPR:1, NLPCC:1, USENIX:1, ICAART:1, Royal Society Open Science (journal):1 resulting in 89 considered papers where 60 were already surveyed by Jin et al. (2022).

A.3 Human Evaluation

For human evaluation, we approach the meta-analysis with the following conventions used in Tables 7, 8, and 9.

- In the Statistical Analysis, Evaluations Released, and Adequately Specified columns, n/a refers to publications that did not conduct a human evaluation
- Statistical Analysis: whether either a statistical significance test on the obtained results or a power analysis in advance to estimate a sufficient number of evaluations to detect the hypothesized effect size is done
- Evaluations Released: whether the human evaluations have been released and are publicly available
- Adequately Specified: If the authors specify annotators' background, number of annotators, number of annotators per sample, incentivization, data selection, questionnaire design, rating scale

[1] `https://github.com/fuzhenxin/Style-Transfer-in-Text`

- Abbreviations (categorization according to Briakou et al. (2021b)) for fluency, content, style, and other: A=Absolute rating of the model output (independent from the output from other models, also includes ratings where the output sentence was rated relative to the input sentence), R=Relative to other TST models' output (e.g., ranking)
- Other: separate question; we do not consider combinations (e.g., geometric means) of other metrics as other

A.4 Automated Evaluation

A.4.1 Conventions

For automated evaluation, we approach the meta-analysis with the following conventions used in Tables 7, 8, 9, and 10.

- We add the automated metric to the evaluation dimension that the authors claim it measures
- Statistical Significance: Yes, if authors conduct a statistical significance test for their claimed results
- Fluency
  - LM based on LSTM (Hochreiter and Schmidhuber, 1997) (LSTM-LM)
  - LM based on Gated Recurrent Units (Cho et al., 2014) (GRU-LM)
  - LM based on Recurrent Neural Networks (RNN-LM) (Zhao et al., 2018)
  - GPT-2(Radford et al., 2019)
  - KenLM (Heafield, 2011)
  - RoBERTa (Liu et al., 2019)
  - Statistical Grammaticality Predictor (SGP) (Heilman et al., 2014)
  - Kneser LM (Kneser and Ney, 1995)
  - Transformer (Vaswani et al., 2017)
- Content
  - Cosine similarity of sentence embeddings (Cosine Sim)
  - Cosine distance of sentence embeddings (Cosine Dist)
  - Logistic Regression (LR)
  - BLEU (Papineni et al., 2002)
    - Source-BLEU: Comparing output to source sentences
    - Ref-BLEU: Comparing output to reference sentences, number of references (if more than 1) in brackets
  - chrF (Popović, 2015)
  - BERT (Devlin et al., 2019)
  - METEOR (Banerjee and Lavie, 2005)
  - BLEURT (Sellam et al., 2020)
  - COMET (Rei et al., 2020)
  - ROUGE (Lin, 2004)
  - CIDEr (Vedantam et al., 2015)
  - BERTScore (Zhang et al., 2020a)
  - CNN Similarity Measure (CNN-SM) (He et al., 2015)
  - Hu Sentiment Classifier (HSC) (Hu et al., 2016)
  - Davidson Classifier (DC) (Davidson et al., 2017)
  - Manning Classifier (MC) (Manning et al., 2014)
- Style
  - All models are classifiers if not differently specified
  - TextCNN (Kim, 2014)
  - fastText (Joulin et al., 2017)
  - Word Mover's Distance (WMD) based on Earth Mover's Distance (EMD) (Kusner et al., 2015)
  - Pretrained Transformer PT (Wolf et al., 2020)
  - Formality Classifier (FC) (Pavlick and Tetreault, 2016)
- Other
  - We do not consider combinations (e.g., geometric means) of other metrics as other
  - GLEU (Napoles et al., 2015)
  - PINC (Chen and Dolan, 2011)
  - Sentence Sim (Wieting et al., 2019)
  - Flesch-Kincaid readability index (Kincaid et al., 1975)
  - Li Diversity (LD) (Li et al., 2016a)

A.4.2 Ranking

We provide an overview of the most common automated metric setups for fluency in Table 4, for content preservation in Table 5, and for style transfer strength in Table 6.

Fluency	Count
GPT-2 PPL	9
GRU-LM PPL on dataset	3
5-gram KenLM PPL	3
LSTM-LM PPL on dataset	2
RNN-LM PPL forward/backward	2
SGP	2
3-gram LM PPL on dataset	2
LSTM-LM PPL	2
5-gram LM PPL	2
LM PPL on dataset	2

Table 4: Shown are the most common setups for automated fluency evaluation with at least two papers utilizing them, showing a great diversity of LM architectures and training setups (whether trained on all styles of the dataset at hand, on each style separately, forward/backward, or not further specified).

Style	Count
TextCNN	24
fastText	13
Classifier	10
BERT	6
LSTM	5
Bi-LSTM	3
CNN	3
GRU	3
LR	2
HSC	2
RoBERTa	2

Table 6: Shown are the most common setups for automated style transfer strength evaluation, with at least two papers utilizing them, showing the dominance of TextCNN and fastText. Often the classifier is not further specified.

Content	Count
Source-BLEU	33
Ref-BLEU	27
Cosine Sim	6
METEOR	4
BERTScore	4
Cosine Dist	3
Word Overlap	2
BERT fine-tuned with STS	2
WMD	2

Table 5: Shown are the most common setups for automated content evaluation, with at least two papers utilizing them, showing the dominance of BLEU-based metrics.

A.5 Experimentation

For experimentation, we approach the meta-analysis with the following conventions used in Tables 7, 8, 9, and 10.

- Datasets
  - Amazon (He and McAuley, 2016)
  - Beer reviews (McAuley et al., 2012)
  - Bible (Carlson et al., 2018)
  - Blogs (Schler et al., 2006)
  - Caption (Li et al., 2018)
  - DIAL (Blodgett et al., 2016)
  - Europarl (Koehn, 2005)
  - FlickrStyle10K (Gan et al., 2017)
  - Gender (Reddy and Knight, 2016)
  - Gigaword (Napoles et al., 2012)
  - Gutenberg (Lahiri, 2014)
  - GYAFC (Rao and Tetreault, 2018)
  - IBC (Sim et al., 2013)
  - IMDb (Diao et al., 2014)
  - IMDb2 (Maas et al., 2011)
  - MSD (Cao et al., 2020)
  - MTFC, TCFC (Wu et al., 2020)
  - MovieTriples (Serban et al., 2016)
  - Opinosis (Ganesan et al., 2010)
  - Paper-News (Fu et al., 2018)
  - Paraphrase corpus (Creutz, 2018)
  - Personality captioning (Shuster et al., 2019)
  - Political slant (Voigt et al., 2018)
  - Reddit (dos Santos et al., 2018)
  - Reddit2 (Baumgartner et al., 2020)
  - ROC (Mostafazadeh et al., 2016)
  - Rotten Tomatoes reviews (Pang and Lee, 2005)
  - Shakespeare (Xu et al., 2012)
  - SimpWiki (den Bercken et al., 2019)
  - SST (Socher et al., 2013)
  - Toxicity [2]
  - Trump speeches [3]
  - TV Series Transcripts (Li et al., 2016b)
  - Twitter (dos Santos et al., 2018)
  - Twitter Persona (Li et al., 2016b)

[2] https://www.tensorflow.org/datasets/catalog/civil_comments

[3] www.kaggle.com/binksbiz/mrtrump

- Twitter Sordoni (Li et al., 2016b)
- Twitter2 (Pardo et al., 2016)
- Wikipedia (Xu et al., 2016)
- Wikipedia2 (Radford et al., 2018)
- Yahoo (Zhang et al., 2015)
- Yelp[s] (Shen et al., 2017; Li et al., 2018), additional references by Jin et al. (2019) and Luo et al. (2019)
- Yelp[l] (Lample et al., 2018), (Xu et al., 2018), (Zhang et al., 2018b), (Guu et al., 2018)[4]
- Youtube2text (Chen and Dolan, 2011)

- >1 Run: No, if not mentioned or no standard deviation reported
- Code Provided: Yes only if code is still available
- Evaluation Code Provided: Yes, only if the code is still available
- Preprocessing Specified: No, if not explicitly mentioned

A.6 Summary Tables

Table 7, 8, and 9 summarize our meta-analysis.

[4] https://www.yelp.com/dataset/challenge

	Human Evaluation							Automated Evaluation					Experiments				
Source	Statistical Analysis	Evaluations Released	Adequately Specified	Fluency	Content	Style	Other	Statistical Significance	Fluency	Content	Style	Other	Datasets	>1 Run	Code Provided	Eval. Code Provided	Prepr. Specified
Xu et al. (2012)	No	No	No	A	A	A	Overall A	No	-	-	Cosine sim, LM, LR	Semantic Adequacy: Ref-BLEU; Lexical Dissimilarity: PINC	Shakespeare	No	No	No	No
Li et al. (2016b)	No	No	No	-	-	-	Consistency R	No	-	-	-	Overall: Ref-BLEU	Twitter Persona, Twitter Sordoni, TV Series Transcripts	No	Yes	No	No
Mueller et al. (2017)	n/a	n/a	n/a	-	-	-	-	Yes	-	-	-	Overall: LM Log Likelihood	Beer reviews, Shakespeare	Yes	No	No	No
Hu et al. (2017)	n/a	n/a	n/a	-	-	-	-	No	-	-	HSC	-	IMDb, SST	No	Yes	Yes	No
Gan et al. (2017)	No	No	No	-	-	R	-	No	-	-	-	Overall: Source-BLEU, METEOR, ROUGE, CIDEr	FlickrStyle10K, Youtube2text	No	No	No	Yes
Shen et al. (2017)	No	No	No	A	-	A	Overall R	No	-	-	TextCNN	-	Yelp[s]	No	Yes	Yes	No
Han et al. (2017)	n/a	n/a	n/a	-	-	-	-	No	-	-	-	Overall: Ref-BLEU	Shakespeare	No	No	No	No
Fu et al. (2018)	No	No	No	-	A	-	-	No	-	Cosine Dist	LSTM	-	Amazon, Paper-News	Yes	Yes	Yes	Yes
Li et al. (2018)	No	Yes	No	A	A	A	-	No	-	Ref-BLEU	Bi-LSTM	-	Yelp[s], Amazon, Captions	No	Yes	Yes	No
Xu et al. (2018)	No	Yes	No	-	A	A	-	No	-	Source-BLEU	TextCNN	-	Yelp[l], Amazon	No	Yes	Yes	No
Prabhumoye et al. (2018)	No	No	No	A	A	-	-	No	-	-	CNN	-	Gender, Political Slant, Yelp[s]	No	Yes	Yes	No
dos Santos et al. (2018)	n/a	n/a	n/a	-	-	-	-	No	LSTM-LM PPL on dataset	Cosine Dist	DC	-	Twitter, Reddit	No	No	No	Yes
Liao et al. (2018)	No	No	No	A	A	-	-	No	-	-	TextCNN, MC	-	Yelp[s]	No	Yes	No	No
Zhang et al. (2018b)	No	No	No	A	A	A	-	No	-	Source-BLEU	TextCNN	-	Yelp[l]	No	Yes	Yes	Yes
Zhao et al. (2018)	No	No	No	A	A	A	-	No	RNN-LM PPL forward/backward	Source-BLEU	fastText	-	Yelp[s], Yahoo	No	Yes	Partially	No
Rao and Tetreault (2018)	No	No	No	A	A	A	Overall R	No	SGP	CNN-SM	FC	Overall: Ref-BLEU, PINC, TERp	GYAFC	No	No	No	Yes
Zhang et al. (2018a)	n/a	n/a	n/a	-	-	-	-	Yes	-	-	Classifier	Overall: ROUGE	Gigaword	Yes	No	No	Yes
Logeswaran et al. (2018)	No	No	No	A	A	A	-	No	LM PPL on external dataset	Own based on BLEU, Ref-BLEU	HSC	-	Yelp[s], IMDb, Shakespeare	No	No	No	No
Chen et al. (2018)	No	No	No	A	A	A	-	No	-	-	TextCNN	Overall: Ref-BLEU; Diversity: Source-BLEU	Yelp[s]	No	Yes	Yes	Yes
Yang et al. (2018)	n/a	n/a	n/a	-	-	-	-	No	LM PPL on dataset per style	Source-BLEU	TextCNN	Overall: Ref-BLEU	Yelp[s]	No	No	No	Yes
Carlson et al. (2018)	n/a	n/a	n/a	-	-	-	-	No	-	-	-	Overall: Ref-BLEU; Diff from source: PINC	Bible	No	Yes	Yes	Yes
Niu and Bansal (2018)	No	No	No	-	-	A	Overall A	No	-	Source-BLEU	SVM, CNN, LSTM-CNN	-	Stanford Politeness Corpus, MovieTriples	No	Yes	Yes	Yes
Guu et al. (2018)	No	No	No	A	A	-	Plausibility A	No	-	-	-	-	Yelp[l], Billionword	No	Yes	No	No
Shetty et al. (2018)	No	No	No	-	A	-	-	No	-	METEOR	LSTM	-	Blogs, Own	No	No	No	No
Jain et al. (2019)	No	No	No	-	-	-	Readability R	No	4-gram KenLM PPL on external dataset	Cosine Sim	-	Overall: Ref-BLEU; Readbility: Flesch-Kincaid readability index	Own	No	Yes	Yes	No
Dai et al. (2019)	No	No	No	R	R	R	-	No	5-gram KenLM PPL on dataset	Source-BLEU	fastText	Overall: Ref-BLEU	Yelp[s], IMDb	No	Yes	Yes	No
John et al. (2019)	No	No	No	A	A	A	-	No	3-gram LM PPL on dataset	Cosine Sim, word overlap	TextCNN	-	Yelp[s], Amazon	No	Yes	Yes	No
Kajiwara (2019)	n/a	n/a	n/a	-	-	-	-	No	-	-	-	Overall: Ref-BLEU, F1 on added, kept and deleted words	GYAFC	No	No	No	Yes
Wu et al. (2019a)	No	No	No	A	A	A	-	No	-	Ref-BLEU	TextCNN	-	Yelp[s], Amazon	No	Yes	Yes	No
Jin et al. (2019)	No	No	No	A	A	A	-	No	LSTM-LM PPL	Ref-BLEU[5,4]	TextCNN	-	Yelp[s], GYAFC	No	No	No	No
Li et al. (2019)	No	No	No	R	R	R	Overall R	No	-	Ref-BLEU, Source-BLEU	TextCNN	-	IMDb, GYAFC, Amazon, Yelp[s], Yahoo	No	Yes	Yes	No

Table 7: Shown are the papers considered for the meta-analysis.

	Human Evaluation							Automated Evaluation					Experiments				
Source	Statistical Analysis	Evaluations Released	Adequately Specified	Fluency	Content	Style	Other	Statistical Significance	Fluency	Content	Style	Other	Datasets	>1 Run	Code Provided	Eval. Code Provided	Prepr. Specified
Shang et al. (2019)	No	No	No	A	A	A	-	No	-	Ref-BLEU	TextCNN	Overall: GLEU	GYAFC, Own	No	No	No	Yes
Sudhakar et al. (2019)	No	No	No	R	R	-	-	No	GPT-2 PPL on dataset per style	Source-BLEU	fastText	Overall: GLEU	Yelp[s], Amazon, Captions, Political Slant, Gender	No	Yes	Yes	No
Wang et al. (2019b)	n/a	n/a	n/a	-	-	-	-	No	-	BERT fine-tuned with STS	LSTM	Overall: Ref-BLEU, PINC	GYAFC	No	Yes	Partially	Yes
Tikhonov et al. (2019)	n/a	n/a	n/a	-	-	-	-	No	-	Source-BLEU	Classifier	Overall: Ref-BLEU	Yelp[s]	Yes	Yes	Yes	No
Leeftink and Spanakis (2019)	No	No	No	-	-	A	Overall A	No	-	-	Attention-RNN	-	IMDb2, Rotten Tomatoes reviews	No	No	No	Yes
Lample et al. (2018)	No	No	No	A	A	A	Overall R	No	5-gram Kneser LM PPL	Ref-BLEU, Source-BLEU	fastText	-	Yelp[s], Yelp[l], Amazon	No	Yes	Yes	Yes
Luo et al. (2019)	No	No	No	A	A	A	-	No	-	Ref-BLEU[4,4]	TextCNN	-	Yelp[s], GYAFC	No	Yes	Yes	No
Wu et al. (2019b)	No	No	No	A	A	A	-	No	-	Ref-BLEU	Bi-LSTM	-	Yelp[s], Amazon	No	No	No	No
Fu et al. (2019)	n/a	n/a	n/a	-	-	-	-	No	-	-	LR, TextCNN, pivot classifier	-	Yelp[s], Amazon, Captions, Political Slant, Gender, Paper-News	No	Yes	Yes	No
(Mir et al., 2019)	No	Yes	No	A	A	A	-	No	LSTM-LM PPL, unigram and neural logistic regression classifiers as adversarials	Style removal, style masking with Source-BLEU, METEOR, embedding average, greedy matching, vector extrema, WMD	TextCNN, fastText	-	Yelp[s]	No	Yes	Yes	No
Romanov et al. (2019)	n/a	n/a	n/a	-	-	-	-	No	LM PPL on dataset	Cosine Sim	GRU	-	Paper-News, Shakespeare	No	Yes	No	No
Gong et al. (2019)	No	No	No	A	A	A	-	No	GRU-LM PPL on dataset per style	Cosine Dist	LSTM	-	Yelp[s], GYAFC	No	No	No	No
Wang et al. (2019a)	No	No	No	A	A	A	-	No	LM PPL on dataset	Ref-BLEU	fastText	-	Yelp[s], Amazon, Captions	No	Yes	Partially	No
Li et al. (2020c)	n/a	n/a	n/a	-	-	-	-	No	-	Source-BLEU	Classifier	-	Personality captioning, FlickrStyle10K, Yahoo, Yelp[s]	No	No	No	No
Liu et al. (2020)	No	No	No	A	A	A	-	No	3-gram Kneser LM PPL on dataset	word overlap, noun overlap	Bi-LSTM	Overall: Ref-BLEU-2	Yelp[s], Amazon	No	Yes	Yes	Yes
Pryzant et al. (2020)	Yes	No	No	A	A	A	-	Yes	-	-	-	Overall: Source-BLEU, Classifier	Own, IBC, News headlines, Trump speeches	No	Yes	Yes	No
Syed et al. (2020)	n/a	n/a	n/a	-	-	-	-	No	-	Source-BLEU, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L	Stylistic Alignment (Own)	-	Gutenberg, Opinosis, Wikipedia2, Shakespeare	Yes	No	No	No
Wu et al. (2020)	No	No	No	A	-	A	Overall R	No	SGP	-	GRU	Overall: Source-BLEU, BLEU-2, Embedding Avg, Embedding Extrema, Embedding Greedy	MTFC, TCFC	No	Yes	Yes	Yes
Cao et al. (2020)	No	No	No	-	A	-	-	No	BERT on dataset per style	Source-BLEU, Ref-BLEU	fastText	-	SimpWiki, MSD	No	No	No	Yes
Madaan et al. (2020)	No	No	No	A	A	A	-	No	-	Source-BLEU	LSTM	Overall: Ref-BLEU, ROUGE, METEOR	Yelp[s], Amazon, Gender, Political Slant	No	Yes	Partially	Yes
Zhang et al. (2020b)	No	No	No	A	A	A	-	No	-	-	-	Overall: Ref-BLEU	GYAFC, Own	No	No	No	No
Zhou et al. (2020)	No	No	No	A	A	A	-	No	-	Ref-BLEU[4]	TextCNN	-	Yelp[s], GYAFC	No	Yes	Yes	No
Jin et al. (2020)	No	No	No	A	A	R	Attractive. A	No	GPT-2 PPL	-	-	Overall: Source-BLEU, METEOR, ROUGE, CIDEr	Own	No	Yes	Partially	No
Duan et al. (2020)	No	No	No	-	A	A	-	No	-	-	TextCNN	Diversity: LD	Yelp[s], Paper-News	No	Yes	Yes	Yes
Tran et al. (2020)	No	No	No	A	A	-	-	No	GPT-2 PPL	Source-BLEU, METEOR, ROUGE	Classifier	-	Own	No	No	No	Yes
Huang et al. (2020)	No	No	No	A	A	A	-	No	RNN-LM PPL forward/backward	Source-BLEU	fastText	-	Yelp[s], Yahoo	No	No	No	No
Wang et al. (2020)	Yes	No	No	-	-	-	Overall A	No	-	BERT fine-tuned with STS	GRU	Overall: Ref-BLEU[4], PINC	GYAFC	No	Yes	Partially	No
Kim and Sohn (2020)	n/a	n/a	n/a	-	-	-	-	No	3-gram LM PPL on dataset	Source-BLEU	fastText	-	Yelp[s], Amazon	No	Yes	Yes	No
Jafaritazehjani et al. (2020)	No	No	No	A	R	A	-	No	GRU-LM PPL on dataset	Cosine Sim	TextCNN	-	Yelp[s]	No	No	No	Yes
Sancheti et al. (2020)	No	No	No	-	A	A	-	No	-	Source-BLEU	TextCNN	-	GYAFC, Yelp[l], Shakespeare	No	No	No	No
Pant et al. (2020)	n/a	n/a	n/a	-	-	-	-	No	-	Source-BLEU	Classifier	-	Yelp[s], Yelp[l]	No	No	No	No
Chakrabarty et al. (2020)	No	No	No	-	A	A	Overall A	No	-	Ref-BLEU, BERTScore	-	Novelty: Own	Own	No	Yes	Yes	Yes

Table 8: Shown are the papers considered for the meta-analysis.

	Human Evaluation							Automated Evaluation					Experiments				
Source	Statistical Analysis	Evaluations Released	Adequately Specified	Fluency	Content	Style	Other	Statistical Significance	Fluency	Content	Style	Other	Datasets	>1 Run	Code Provided	Eval. Code Provided	Prepr. Specified
Krishna et al. (2020)	No	No	No	A	A	-	-	No	RoBERTa on external dataset	Source-BLEU	RoBERTa	-	GYAFC, Shakespeare	No	Yes	Yes	Yes
Ma et al. (2020)	No	No	No	A	R	R	-	No	Pretrained GPT PPL	BERTScore	Classifier	Repetitiveness: Own; Diversity: Own	ROC, Paraphrase corpus	No	No	No	Yes
Malmi et al. (2020)	n/a	n/a	n/a	-	-	-	-	No	-	Source-BLEU	BERT	-	Yelp[s]	No	No	No	Yes
Cheng et al. (2020)	No	No	No	-	R	R	Context Consistency R	No	LSTM-LM PPL on dataset	Ref-BLEU	TextCNN	Overall: GLEU	GYAFC, Yahoo, Own	No	No	No	Yes
Li et al. (2020b)	n/a	n/a	n/a	-	-	-	-	No	-	Ref-BLEU, Source-BLEU	fastText	-	Yelp[s], IMDb	No	No	No	Yes
Dathathri et al. (2020)	No	No	No	A	-	R	-	No	GPT PPL	-	Classifier	Diversity: LD	n/a	No	Yes	Yes	n/a
He et al. (2020)	n/a	n/a	n/a	-	-	-	-	No	LSTM-LM PPL on dataset per style	Ref-BLEU, Source-BLEU	TextCNN	-	Yelp[s], GYAFC	No	Yes	Yes	Yes
Xu et al. (2020)	n/a	n/a	n/a	-	-	-	-	No	GPT-2 PPL	Source-BLEU	CNN	Overall: GLEU	Yelp[s], Amazon	No	Yes	Partially	Yes
Yi et al. (2020)	No	No	No	A	A	A	-	No	5-gram KenLM PPL	Ref-BLEU, Cosine Sim	BERT	-	Yelp[s], GYAFC, Own	No	Yes	Partially	Yes
Lee (2020)	No	No	No	A	A	A	-	No	GPT-1 PPL, GPT-2 PPL	Ref-BLEU[2], Source-BLEU, BERTScore	TextCNN	-	Yelp[s], Amazon	No	Yes	Yes	Yes
Li et al. (2020a)	No	No	No	A	A	-	-	No	GPT-2 PPL	Source-BLEU	RoBERTa	-	GYAFC	No	Yes	Yes	Yes
Lyu et al. (2021)	No	Yes	No	A	A	A	-	No	-	-	-	Overall: Source-BLEU, METEOR, ROUGE, CIDEr	Own	No	Yes	Partially	Yes
Liu et al. (2021)	No	No	No	A	A	A	-	No	GPT-2 PPL	Source-BLEU, Ref-BLEU	fastText	-	Yelp[s], Amazon, IMDb	No	Yes	Yes	Yes
Goyal et al. (2021)	No	No	No	A	A	A	Overall A	No	Transformer PPL on dataset	Source-BLEU, Ref-BLEU	fastText	-	GYAFC, IMDb	No	No	No	No
Briakou et al. (2021c)	No	Yes	Yes	A	A	A	Overall R	No	5-gram Kneser LM PPL	Source-BLEU	BERT	Overall: Ref-BLEU	GYAFC, Own	No	Yes	No	Yes
Ma and Li (2021)	No	No	No	A	A	A	-	No	GRU-LM PPL on dataset	Ref-BLEU[4]	TextCNN	-	Yelp[s], GYAFC	No	Yes	Yes	No
Lai et al. (2021)	n/a	n/a	n/a	-	-	-	-	No	-	BLEURT, COMET, Ref-BLEU[4]	Classifier	-	Yelp[s], GYAFC	No	Yes	Yes	No
Mireshghallah and Berg-Kirkpatrick (2021)	No	No	No	R	R	-	-	No	GPT-2 PPL, GLEU, lexical diversity	Source-BLEU	Classifier	-	Yelp[s], Twitter2, DIAL	No	Yes	Yes	Yes
Xiao et al. (2021)	No	No	No	A	A	A	Consistency A	No	5-gram KenLM PPL	Ref-BLEU, Source-BLEU	BERT	-	Yelp[s], GYAFC	No	Yes	Yes	No
Ma et al. (2021)	No	No	No	A	A	A	-	No	GRU-LM PPL on dataset	Ref-BLEU[4]	TextCNN	-	Yelp[s], GYAFC	No	Yes	Yes	Yes
Laugier et al. (2021)	No	No	No	A	A	A	Overall A	No	GPT-2 PPL	Own	BERT	-	Yelp[s], IMDb, Toxicity	No	Yes	Yes	No
Yu et al. (2021)	No	No	No	A	A	A	-	No	Grammarly, LM PPL, adversarial unigram classifier	Attribute Hit (Own), Ref-BLEU, Source-BLEU, EMD	Classifier	-	Yelp[s], IMDb	No	No	No	Yes
Briakou et al. (2021a)	No	No	No	A	A	A	Overall R	No	LL, 5-gram KenLM PPL, BERT PLL, XLM PLL	Ref-BLEU, Source-BLEU, METEOR, chrF, WMD, Cosine Sim, BERT, BERTScore, XLM	BERT, XLM	-	GYAFC	No	Yes	Yes	No
Kashyap et al. (2022)	No	No	No	A	A	A	-	No	RoBERTa fine-tuned on external dataset	Sentence Sim	fastText	-	Yelp[s], IMDb, Political Slant	No	No	No	No
Liu et al. (2022)	No	No	No	A	A	A	-	No	-	-	TextCNN	Overall: Ref-BLEU	GYAFC	No	Yes	Yes	No
Reif et al. (2022)	No	No	No	A	A	A	-	No	GPT-2 PPL	Ref-BLEU	PT	-	Yelp[s], GYAFC	No	No	No	No

Table 9: Shown are the papers considered for the meta-analysis.

B Human Validation Details

We summarize the validation of automated metrics for TST evaluation in Table 10. In general, we consider an automated metric to be validated if it was validated for the mentioned aspect. We observe the three aspects of fluency, content preservation, and style transfer strength being validated. In addition, we also find validations for metrics evaluating TST as a whole. For fluency, we consider the PPL of an LM for measuring fluency to be validated. For style transfer strength, we consider any classifier architecture to be validated. Validation by Yamshchikov et al. (2021) is not considered because of the different purpose (validated on datasets for TST and paraphrasing and not actual model outputs).

Source	Fluency	Content	Style	Overall	Datasets
Xu et al. (2012)	-	Ref-BLEU	Ref-BLEU, PINC, LM, LR, Cosine Sim	-	Shakespeare
Niu and Bansal (2018)	-	-	-	Source-BLEU	Stanford Politeness Corpus
Li et al. (2018)	Ref-BLEU	Ref-BLEU	Bi-LSTM	-	Yelp[s], Captions, Amazon
Fu et al. (2018)	-	Cosine Dist	-	-	Amazon, Paper-News
Rao and Tetreault (2018)	SGP	CNN-SM	FC	Ref-BLEU, PINC, TERp	GYAFC
Luo et al. (2019)	Ref-BLEU	Ref-BLEU	TextCNN	-	Yelp[s], GYAFC
Mir et al. (2019)	Adv. Classifier, LSTM-LM PPL	Source-BLEU, METEOR, Embed Average, Embed Greedy, Embed Extrema, WMD	TextCNN, fastText	-	Yelp[s]
Pang and Gimpel (2019)	LM PPL	Cosine Sim	TextCNN	-	Yelp[s]
Wang et al. (2020)	-	-	-	Ref-BLEU	GYAFC
Pryzant et al. (2020)	Source-BLEU, Classifier	Source-BLEU, Classifier	Source-BLEU, Classifier	-	WNC
Wu et al. (2020)	SGP	-	GRU	Embed Average, Embed Extrema, Embed Greedy, BLEU-2	MTFC, TCFC
Cao et al. (2020)	-	Ref-BLEU, Source-BLEU	-	-	SimpWiki, MSD
Yamshchikov et al. (2021)	-	Multiple	-	-	7 TST & Paraphrase Datasets
Briakou et al. (2021a)	5-gram KenLM PPL, BERT PLL, XLM PLL	Ref-BLEU, Source-BLEU, METEOR, chrF, WMD, Cosine Sim, BERTScore, BERT, XLM	BERT, XLM	-	GYAFC

Table 10: Shown are publications validating automated metrics; underlined are the ones showing the highest correlation (if multiple compared).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☐ A2. Did you discuss any potential risks of your work?
Not applicable. We review existing work in terms of text style transfer evaluation and try to point out existing problems.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
0, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We conducted an extensive meta-analysis of text style transfer. We surveyed 89 works summarized in Sections 2, 3 and the Appendix

☑ B1. Did you cite the creators of artifacts you used?
1, 2, 3, 4, Limitations, Appendix

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. We just reviewed the existing works.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
1,2,3,4

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We reviewed existing papers, we did not collect data ourselves.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, we reviewed text style transfer papers and pointed out the problems of existing human evalutions in this field. Section 2.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.