A large-scale computational study of content preservation measures for text style transfer and paraphrase generation

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko
Skolkovo Institute of Science and Technology (Skoltech)
{n.babakov, d.dale, v.logacheva, a.panchenko}@skoltech.ru

Abstract

Text style transfer and paraphrasing of texts are actively growing areas of NLP, dozens of methods for solving these tasks have been recently introduced. In both tasks, the system is supposed to generate a text which should be semantically similar to the input text. Therefore, these tasks are dependent on methods of measuring textual semantic similarity. However, it is still unclear which measures are the best to automatically evaluate content preservation between original and generated text. According to our observations, many researchers still use BLEU-like measures, while there exist more advanced measures including neural-based that significantly outperform classic approaches. The current problem is the lack of a thorough evaluation of the available measures. We close this gap by conducting a large-scale computational study by comparing 57 measures based on different principles on 19 annotated datasets. We show that measures based on cross-encoder models outperform alternative approaches in almost all cases. We also introduce the Mutual Implication Score (MIS), a measure that uses the idea of paraphrasing as a bidirectional entailment and outperforms all other measures on the paraphrase detection task and performs on par with the best measures in the text style transfer task.

1 Introduction

Text style transfer (TST) and paraphrases generation (PG) are active areas of research in NLP, with dozens of papers proposing new methods. These methods could be applied for practical purposes, such as supporting human writers, personalizing digital assistants, or even creating artificial personalities.

Research and development of TST models require fast feedback loops, and they require fast and reliable automatic quality measures. TST is hard to evaluate for several reasons. First, golden answers, even if available, are not the only valid way to rewrite the text. Second, parallel corpora with different styles do not emerge naturally and are hard to find. This means that reference-based evaluation is often prohibitive and creates a need for manual evaluation of TST or for clever automatic measures.

The basic desired properties of TST are style accuracy, content preservation, and fluency (Mir et al., 2019). For many methods of unsupervised TST, keeping the content of the original text and automatically measuring its preservation is one of the most difficult tasks (see e.g. Dale et al. (2021)).

During development, the only way to control content preservation is to use automatic measures. Such measure takes two sentences and return the value which indicates the similarity of their content. More formally, the measure sim quantifies semantic relatedness of two utterances, an original text x and a style-transferred or paraphrased text y : $sim(x, y) \rightarrow [0; 1]$. The measure sim yields high score for the pairs with similar content and low score for ones with different content.

As Krishna et al. (2020) and Yamshchikov et al. (2021) show, most TST works evaluate the content preservation with BLEU (Papineni et al., 2002) or similar measures based on word overlap between two texts. The situation in PG is almost identical. Most works including the most recent ones (Sun et al., 2021; Fu et al., 2020) also still rely on BLEU.

Even though measures like BLEU, based on a word or character-level n-grams are pretty intuitive and straightforward, they don't take into account synonyms and distributively related words. Moreover, there already exist several pieces of evidence that correlation of standard BLEU-like automatic measures is relatively low (Briakou et al., 2021). The recent development of vector representations of textual information (Mikolov et al., 2013; Zhang et al., 2019) and various ways to handle these vectors provides room for improvement of the approaches to scoring the content preservation. It

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Student Research Workshop, pages 300 - 321
May 22-27, 2022 ©2022 Association for Computational Linguistics

is, therefore, crucial to perform a thorough analysis of all existing content preservation measures and to gather best practices from the top-performing approaches to create a new approach that could demonstrate stable performance in terms of both PG and TST tasks.

In this work, we further extend a comprehensive study of Yamshchikov et al. (2021) by analyzing a much more diverse set of measures including recently developed transformers-based ones, and also by proposing a new measure specially developed for TST and PG content preservation scoring. The contributions of our paper are as follows:

- We perform a large-scale evaluation of automatic content preservation measures for text style transfer and paraphrase generation tasks, which includes 57 measures applied to 9 paraphrasing datasets and 10 text style transfer datasets. To the best of our knowledge, this is the largest and the most comprehensive evaluation of this kind;
- We introduce Mutual Implication Score (MIS): a measure of content preservation based on predictions of NLI models in two directions. We show that it outperforms all known measures in paraphrase detection and shows consistently high results for TST. We opensource the model on Huggingface Model Hub.[1]

The code for measures and experiments is released publicly.[2]

2 Related work

2.1 Measures of content preservation

There exists a large number of content preservation measures that can be classified into several groups. In this section, we describe all of these approaches. Refer to Figure 1 for a schematic description of all approaches.

Words or characters n-grams (ngram) The most simple and intuitive way to compare two texts is based on the overlap of word or character n-grams. The standard method used to evaluate the quality of a generated text is to compare it with a human-written reference text via BLEU score (Papineni et al., 2002), which is the precision of word n-grams. In TST and PG papers, BLEU is often used to evaluate content preservation relative to the original text or a reference. Other popular measures based on words or n-grams are ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015). Such approaches as Levenshtein distance (Levenshtein et al., 1966), Jaro-Winkler distance (Jaro, 1989) also work at the subword level by calculating the edit distance between two sequences, so we also refer them to the ngram group. Panchenko and Morozova (2012) provided a comparative study of classic word similarity measures and their combinations. The ngram measures are simple and intuitive but do not handle well such linguistic phenomena as synonyms, negation, and issues with word order.

Similarity between static embeddings (emb-static) Another family of measures partially overcomes these difficulties by representing texts with their embeddings and calculating the distance (e.g. cosine similarity) between the embeddings of two texts. This group of measures can be further divided by the way the embeddings are generated. The basic way of obtaining the embedding of a text is by averaging across static word embeddings: Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017).

Similarity between contextualized embeddings (emb-context) Special distance function (e.g. WMD (Kusner et al., 2015), POS-distance (Tian et al., 2018a)) can be also applied to context-dependent vectors: BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019).

Similarity between embeddings from bi-encoders (emb-bi-enc) Embeddings of a text can be generated by encoding a text with a pre-trained *encoder*. If the two texts are encoded separately, and then we compute the cosine similarity between their embeddings, we refer to such models as *bi-encoders*. This group of models is usually trained in a supervised manner. The encoders can be trained on the translation task (Laser (Artetxe and Schwenk, 2019), LaBSE (Feng et al., 2020)), paraphrase identification task (SIMILE (Wieting et al., 2019)), or text generation task (BARTScore (Yuan et al., 2021)). They potentially can compare the meanings of texts that are very different in terms of structure and vocabulary.

[1] https://huggingface.co/SkolkovoInstitute/Mutual_Implication_Score

[2] https://github.com/skoltech-nlp/mutual_implication_score

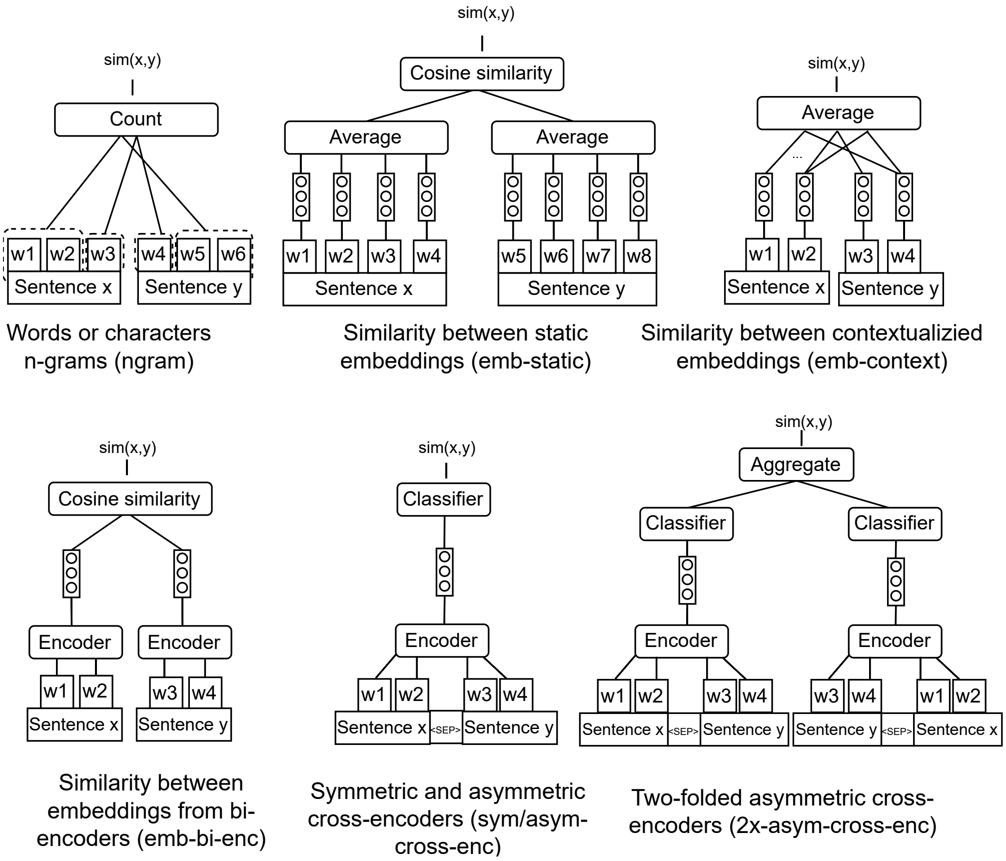

Figure 1: Different approaches to calculating content preservation between two sentences.

Symmetric and asymmetric cross-encoders (sym/asym-cross-enc) The models called *cross-encoders* process both texts simultaneously using cross-attention and directly predict the relationship between the texts. They can perform symmetrically (score is independent of the order of the texts being compared) or asymmetrically (score strongly depends on the order of the texts). Due to their supervised nature, such models can reflect content preservation more accurately than word-based approaches, but they depend on labeled data and may not generalize well to new domains. The presence of symmetry is defined by the task the model was trained on. Thus, models trained on the Natural Language Inference (NLI) task data (such as BLEURT (Sellam et al., 2020) or NUBIA (Kane et al., 2020)) are asymmetric, while cross-stsb-base model trained solely on STS-B dataset (Cer et al., 2017) for semantic textual similarity, or APD model (Nighojkar and Licato, 2021) trained on paraphrase datasets perform symmetrically.

Two-folded asymmetric cross-encoders (2x-asym-cross-enc) A textual entailment model can be used for scoring semantic relations between two phrases. Nighojkar and Licato (2021) propose to use a natural language inference (NLI) model for paraphrase identification, and Deng et al. (2021) suggest a similar model for evaluation of summarization and text style transfer. The main idea of these works is to use NLI models in a two-fold manner (direct and reverse). NLI models are generally asymmetric cross-encoders, so we classify this group of approaches as a two-fold asymmetric encoder.

As shown in Figure 2, despite the wide variety of measures, n-gram-based measures are still used most often, while embedding-based measures and cross-encoders are much less popular. In some papers, no automatic content preservation measures are used.

2.2 Evaluation of content preservation measures

Our work in many respects follows the setup of Yamshchikov et al. (2021) and extends it in several directions. In this work, the authors collected crowdsource estimates of content preservation for 14,000 sentence pairs from 14 sources and compared these estimates with 13 automatic

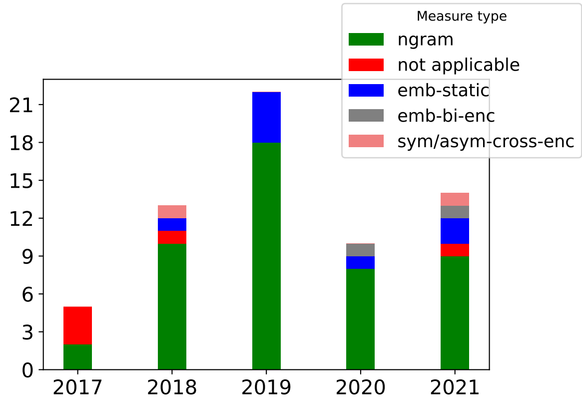

Figure 2: The number of research papers on TST and PG which use automatic content preservation measures from different groups, based on 58 publications listed in Appendix (Table 7).

measures. They evaluated the quality of automatic measures by the correlation between rankings provided by these measures and rankings created by human scores. This scoring showed that the WMD over GloVe embeddings and L2 distance between the ELMo embeddings outperform other measures. However, no supervised sentence encoders or cross-encoders were considered in this work.

In the work by Briakou et al. (2021), the authors evaluated measures of formality transfer in four languages. The main subject of this work is a thorough analysis of multilingual formality style transfer, including a high-level analysis of all aspects of style transfer quality: style accuracy, content preservation, and fluency. The authors used chrF and a cross-encoder (XLM-R) trained on a semantic text similarity dataset to calculate content preservation. They also cautioned against using BLEU in this context, because it has a lower correlation with human judgments than many other measures. However, automatic measures of content preservation were not the main focus of this work, so we extend its results by applying more diverse measures on the English part of their dataset, among others.

3 Datasets used in comparative study

We run our analysis of measures on parallel datasets manually labeled for semantic similarity or content preservation. To make the comparison more generalizable, we fetch a large number of datasets generated by different models.

3.1 Text style transfer datasets

The text style transfer task is aimed at transforming a text to change its *style* (a particular attribute of its text) while keeping the content intact. Since in some cases the style cannot be separated from the content (e.g. if the style is positive/negative sentiment), strict preservation of all content is sometimes impossible in the TST task. Therefore, we consider the parallel TST datasets separately from other data used for the analysis.

In many TST works, outputs were evaluated with human judgments, but the raw similarity labels are rarely published. We managed to find datasets that include human similarity scores for various TST tasks

- Detoxification:
  - Tox600 (Dale et al., 2021),
  - CAE (Laugier et al., 2021)
- Formality transfer:
  - xformal-FoST (Briakou et al., 2021),
  - STRAP_form, (Krishna et al., 2020)
  - Yam. GYAFC (Yamshchikov et al., 2021)[3]
- Sentiment transfer:
  - PG-YELP (Pang and Gimpel, 2019)
  - Yam. Yelp (Yamshchikov et al., 2021)
- Transfer to Old English:
  - Yam. Bible (Yamshchikov et al., 2021),
  - STRAP_coha (old American English), (Krishna et al., 2020)
  - STRAP_SP (Shakespearean English) (Krishna et al., 2020)

3.2 Paraphrases datasets

Unlike TST, the paraphrase generation task requires full preservation of content. There exist a large number of parallel datasets of paraphrases manually labelled for content preservation. The majority of them have binary labels ("same"/"different"). We use the following datasets in our analysis:

- MSRP (Dolan and Brockett, 2005),
- Twitter-URL (Lan et al., 2017),
- PIT (Xu et al., 2014),
- PAWS (Yang et al., 2019b),
- ETPC (Kovatchev et al., 2018),

[3] We use the datasets collected and/or used in the analysis by Yamshchikov et al. (2021). For clarity, we prepend their names with "Yam." prefix.

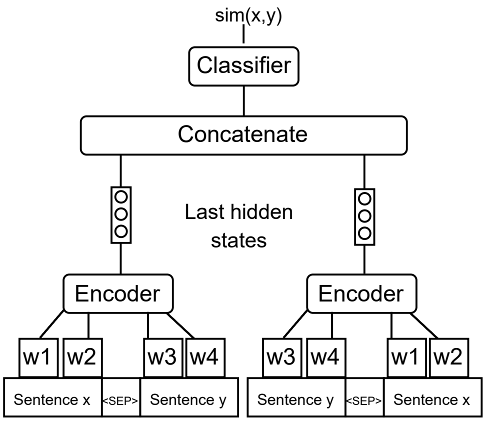

Figure 3: Mutual Implication Score (MIS).

- APT (Nighojkar and Licato, 2021),
- Yam. Para (Yamshchikov et al., 2021).

We provide detailed information about the datasets in the Appendix tables 5 and 6.

4 Mutual Implication Score (MIS)

The goal of our research is not only to analyze the existing measures of content preservation but also to suggest a new measure that can outperform the existing ones. We devise a new measure that is based on measuring content similarity with NLI, as described by Nighojkar and Licato (2021). In this work, the authors exploit the assumption that implies the two sentences with the same meaning should be equivalent in their inferential properties, i.e. each sentence should textually entail the other. This means that the NLI model is supposed to return similar entailment scores when applied to semantically equal sentences regardless of the sequence these sentences are sent to the input of the model. The authors used this assumption to propose an adversarial method of dataset creation for paraphrase identification.

NLI models predict whether one text logically entails another, and are, therefore, asymmetric. High entailment probability in the forward direction means that the second text accurately follows the first one and does not contain hallucinated information. A high entailment score in the backward direction means that all the information from the first text is retained in the second text.

The most natural way to aggregate scores from both directions is to multiply them or compute their arithmetic or harmonic mean. We use this approach as a baseline. We yield NLI scores from the following models:

PG		TST	
Measure	ρ	Measure	ρ
MIS	**0.61**	MIS	**0.54**
DeBERTa	0.60	RobNLI	0.47
RobNLI	0.59	DeBERTa	0.46
FBrobNLI	0.55	FBrobNLI	0.43

Table 1: Mean Spearman correlations of MIS and baseline NLI-based measures on PG and TST datasets. For baseline NLI measures, the forward and backward scores are averaged.

- **RobNLI** (Nie et al., 2020) — RoBERTA-Large (Zhuang et al., 2021) pre-trained on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), FEVER-NLI (Nie et al., 2019), and ANLI (Nie et al., 2020),
- **FBrobNLI** (Liu et al., 2019) — RoBERTA-Large pre-trained only on MNLI,
- **DeBERTa** (He et al., 2021) pre-trained on the MNLI dataset.

Although these NLI models are a good starting point, they might not be fully suitable for measuring content preservation, because they were trained for a different task. We suggest that further fine-tuning them on the data annotated with content preservation scores might yield better models.

Thus, we modify the RoBERTA architecture used for NLI. Namely, we use the original encoder in both forward and backward directions, concatenate the last hidden states, and then send them to the classification module which is tuned on data annotated with content preservation scores. We refer to this model as **Mutual Implication Score** (**MIS**). The scheme of our model is given in Figure 3.

We initialize the model with pre-trained weights from the RobNLI model. We tune it on Quora Question Pairs dataset (Sharma et al., 2019) for 2 epochs with a learning rate $4e^{-6}$ and all but the last encoder layer and classifier layer frozen.

We evaluate the model with the Spearman rank correlation coefficient of the automatic content preservation scores with human judgments. We evaluate all TST and PG datasets introduced in Section 3. We evaluate MIS and baseline NLI-based measures (we aggregate the NLI scores for both directions with the arithmetic mean because it showed the best results in our preliminary experiments).

The results are shown in Table 1. Fine-tuning the (slightly modified) NLI model on content preservation data slightly improves its performance on datasets generated by paraphrasing models and yields significantly higher correlation on TST datasets.

5 Measures analysis

We compute the content preservation scores for paraphrasing and style transfer datasets using measures of different types. We analyze the performance of individual measures and compare the performance of different groups of measures. We also look into the difference in measures performance on PG and TST tasks and analyze the individual datasets.

5.1 Experimental setting

We analyze 57 content preservation measures of different types. As described in Section 2.1, the measures can be divided into the following groups: a word or character n-gram based (ngram), the measures based on the distance between static (emb-static) or contextualized (emb-context) embeddings, or embeddings from bi-encoders (emb-bi-enc), different groups of encoders-based measures: symmetric (sym-cross-enc), asymmetric (asym-cross-enc) or two-fold asymmetric (2x-asym-cross-enc) cross-encoders. This grouping is used explicitly during analysis. The full list of measures is given in Table 8.

We compute the content preservation scores for 19 datasets listed in Section 3. The full information about the datasets is given in Appendix Tables 5 and 6.

We evaluate measures using the Spearman rank correlation coefficient of the automatic scores with human judgments. Since we use a large number of measures and datasets, we report only aggregated results. The full results are available in the Appendix Figures 7 and 8.

5.2 Measure-level analysis

Figure 4 shows the correlations of the best-performing measures from different groups for individual datasets. The last columns of the plots show the performance of each measure averaged across datasets. The plot shows that MIS and similar measures based on two-folded asymmetric cross-encoders have the best average performance on the paraphrase datasets. For TST datasets, there is no clear winner: symmetric cross-encoders (cross-stsb-large/base), bi-encoders (SIMCSE-SL/SB), asymmetric cross-encoders (BLEURT, NUBIA), and two-folded asymmetric cross-encoder (MIS) demonstrate almost equal performance.

Measure	Toxic	Old_Eng	Form	Sent
BLEURT-B128	0.47	0.52	0.61	**0.39**
BLEURT-L128	**0.54**	0.57	0.64	0.35
MIS	0.50	0.60	0.69	0.28
NUBIA	0.43	0.60	0.66	0.33
SIMCSE-SL	0.46	0.60	0.69	0.36

Table 2: Mean Spearman correlation of measures which perform best on different text style transfer tasks. Tasks: *Toxic* — detoxification, *Old_Eng* — old-style to modern English, *Form* — formal to informal, *Sent* — sentiment transfer. The best scores are shown **in bold**.

Paraphrase Generation (PG)			
	ρ_{max}	ρ_{avg}	$\#wins$
2x-asym-cross-enc	**0.61**	**0.56**	3
sym-cross-enc	0.55	0.51	**5**
asym-cross-enc	0.54	0.49	2
emb-bi-enc	0.54	0.45	2
emb-context	0.47	0.42	0
ngram	0.42	0.34	0
emb-static	0.32	0.27	0
Text Style Transfer (TST)			
	ρ_{max}	ρ_{avg}	$\#wins$
sym-cross-enc	**0.55**	**0.51**	**3**
emb-bi-enc	0.55	0.49	3
asym-cross-enc	0.54	0.46	3
2x-asym-cross-enc	0.54	0.45	0
emb-context	0.5	0.45	2
emb-static	0.4	0.36	1
ngram	0.41	0.35	1

Table 3: Spearman correlations of measures belonging to different groups: ρ_{max} — correlation of the best-performing in the group, ρ_{avg} — correlation averaged over the group, $\#wins$ — the number of times the model from the group performs best on any of the datasets.

The performance of content preservation measures on TST datasets varies from style to style. The TST datasets we use contain style transformations of four types: detoxification, formal to informal, positive to negative sentiment, and modern to old-style English. Thus, it seems natural to average the measures performance not only by all TST datasets but also by TST datasets of different styles. The averaged scores are shown in Table 2. There is no clear winner for old-style English and formality transfer: MIS and SIMCSE-SL show almost equal performance. However, we can see that BLEURT measures are clear leaders in detoxification and sentiment transfer.

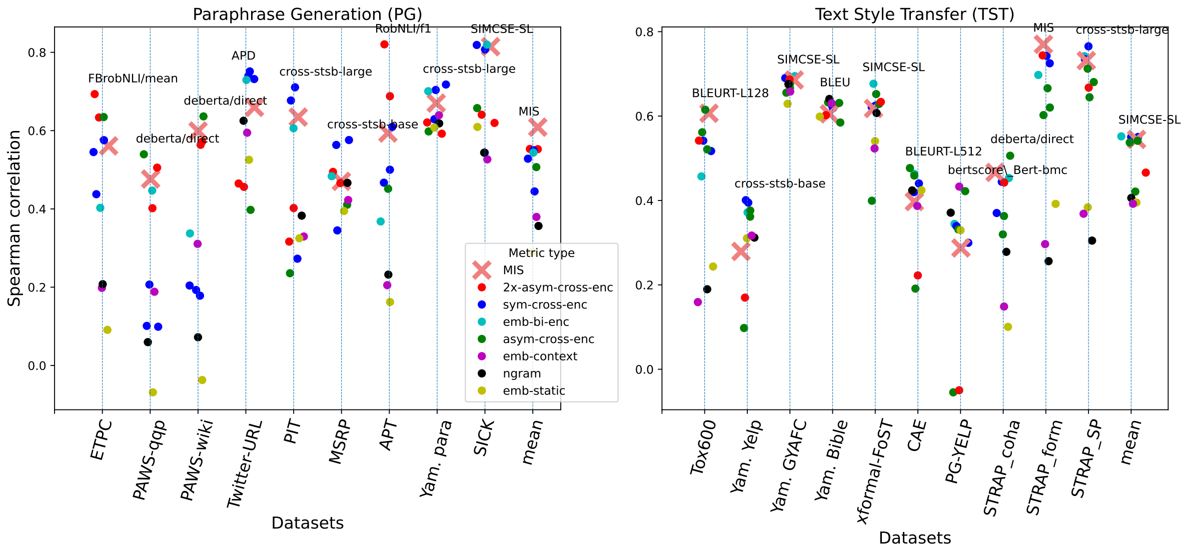

Figure 4: Correlation of measures of different classes with human judgments on paraphrase and text style transfer datasets. The text above each dataset indicates the best-performing measure. The rightmost columns show the mean performance of measures across the datasets.

5.3 Group-level analysis

To get more generalizable results of the analysis, we perform a group-level comparison of measures in Table 3. We report the Spearman correlation scores averaged over datasets of PG and TST tasks (as before, we do not merge all datasets and consider the two tasks separately). We report the mean and maximum correlations of all measures of a group. We also compute the number of times when a measure of a group performs best on the particular dataset. This indicator can be somewhat biased due to the nature of each dataset, however, it can still serve as an additional source of information. If the difference between correlations is not significant (by Williams test (Graham and Baldwin, 2014)) we assign one winning time to each group.

From this point of view, we can even better see that two-folded asymmetric models are the best choice for paraphrases detection because the mean correlation outperforms the next best-performing group by 0.05. Symmetric cross-encoders can also be an alternative option for this task because they show the largest number of wins. Symmetric cross-encoders show the highest mean correlation on the TST task. At the same time, the number of wins and correlations of the best models from this class are similar for all encoder-based classes.

Finally, from the measure-level and group-level perspective, we can see that encoder-based measures outperform ngrams-based measures in the absolute majority of datasets on TST and PG tasks.

5.4 Data-level analysis

So far we relied on the correlations averaged across different datasets. However, it is also natural to have a closer look at how the behavior of different measures changes across datasets.

For this purpose, we represent each dataset as a vector of correlations of each measure with the human judgments and plot a dendrogram (see Figure 5) to show the clustered structure of the obtained vectors. The dendrogram should be interpreted as follows. The height at which each dataset is connected to another dataset or group of datasets indicates the distance between the dataset vectors. We additionally plot a heatmap of cosine similarities of these datasets vectors in Appendix Figure 9.

Datasets related to sentiment transfer (PG-YELP, Yam. Yelp) look different from others, thus, they form a separate cluster in the dendrogram. The reason for this dissimilarity is probably the fact that in this type of TST task (sentiment transfer) the content of the utterance changes more significantly than in other tasks. Moreover, PG-YELP is originally distributed as a pairwise comparison dataset. To yield sentence-level scores, we apply Luce Spectral Ranking (Maystre and Grossglauser, 2016). This preprocessing might affect the quality of labels.

In general, the datasets are clustered into two rather dense groups and this clustering does not match the separation of the datasets among TST and PG tasks. The different behavior of the tested

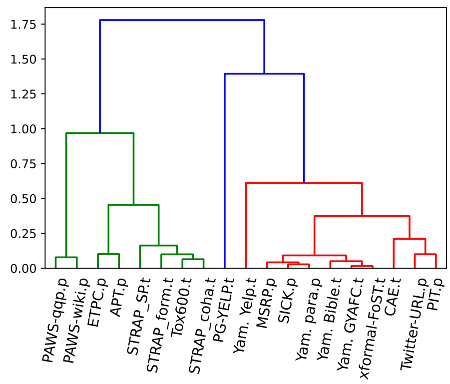

Figure 5: Dendrogram of vectors of measures correlations on a dataset. The height of the bar indicates the distance between vectors or groups of vectors. Postfixes ‘p’ and ‘t’ denote the datasets for PG and TST tasks, respectively.

measures might be explained by the way the data is annotated. For example, the PAWS datasets were collected in an adversarial manner (by shuffling the words in sentences), STRAP datasets were generated with TST models, and Yam. datasets were annotated by a similar group of workers — these three sets form clusters in the dendrogram.

6 Using automatic measures to rank text style transfer systems

While above we compared automatic and human ranking of individual text pairs, our final goal is to find a measure to rank TST or PG *systems*. Six TST datasets used in our analysis were created by running several TST models on the same dataset and manually assessing the degree of content preservation in the resulting text pairs. They cover diverse tasks: formality transfer (xformal-FoST and STRAP_form datasets), text detoxification (Tox600 and CAE datasets), Shakespeare style transfer (STRAP_SP), and sentiment transfer (PG-YELP). We use the human judgments on content preservation from these datasets to rate the ability of various measures to rank text style transfer systems.

For brevity and clarity, we do not report the results of this analysis for all measures. Instead, we select the best-performing measure from each group:

- **cross-encoders**: MIS, RobNLI/mean, BLEURT-L128 and cross-stsb-base,
- **bi-encoders**: LaBSE and SIMCSE-SL (supervised, using ROBERTa-large),
- **embedding-based models**: SIMILE, BERTScore (with microsoft/deberta-xlarge-mnli model), and WMD,
- **ngram-based measures**: BLEU and ChrF.

Measure	Measure type	ρ	*acc*
MIS	2x-asym-cross-enc	0.93	0.50
BLEURT-L128	asym-cross-enc	0.92	0.83
RobNLI/mean	2x-asym-cross-enc	0.83	0.50
cross-stsb-base	sym-cross-enc	0.63	0.50
SIMCSE-SL	emb-bi-enc	0.60	0.50
LaBSE	emb-bi-enc	0.58	0.67
bertscore-Mic-Deberta	emb-context	0.55	0.50
SIMILE	emb-bi-enc	0.38	0.33
BLEU	ngram	0.10	0.17
w2v_wmd	emb-static	0.03	0.17
chrf	ngram	0.03	0.17

Table 4: Mean rank correlation (ρ) of text style transfer system-level automatic scores with human judgments, and percentage of cases when they correctly identify the best system (acc).

We show the results aggregated across the datasets in Table 4. The scores for individual datasets and measures and a list of measures managed to identify the best-performing model for a given dataset are given in Appendix C.

No measure can fully match the system rankings produced by humans. However, our MIS measure and BLEURT have the highest correlations with human judgments. BLEURT performs best on this task because it correctly identifies the winner on 5 datasets out of 6. The popular measures BLEU, ChrF, and WMD identify the best system only on the xformal-FoST dataset.

7 Computational efficiency of the measures

While the correlation of measures with human judgments is important, the usability of the measure in real tasks can not be treated in isolation from its computational efficiency. The main capabilities of such measures are robustness and inference speed.

One of the key functions of content preservation measures is to compare different TST or PG approaches with each other and ensure that different runs of the learning-based measure yield similar results. This problem does not apply to words or character n-grams-based models. However, this could yield some issues with trainable model-based measures. That is why it is crucial for all such measures to open-source trained weights. Moreover, when using such measures for comparison it is nec-

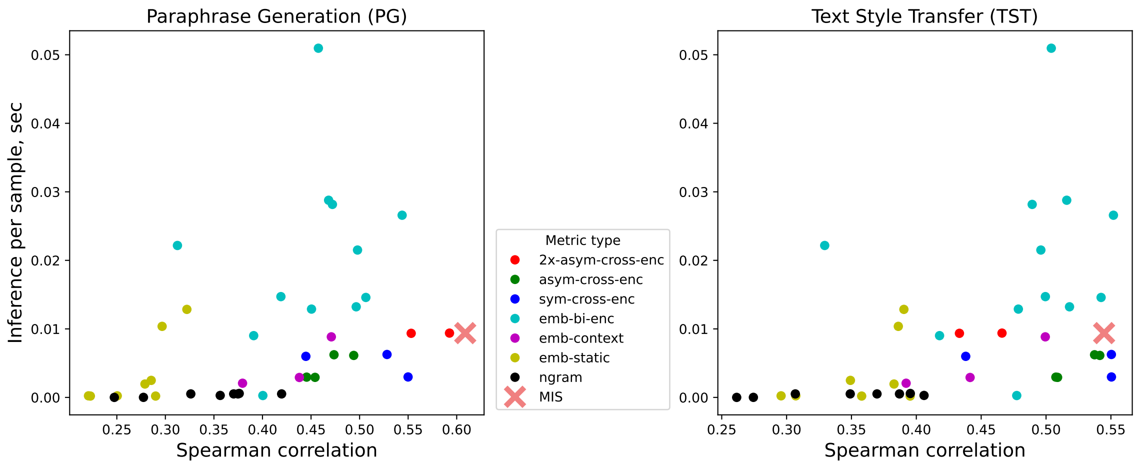

Figure 6: Dependence of time necessary for calculating similarity score for one sample and average correlation of a measure on text style transfer and paraphrases generation tasks.

essary to put the model into inference mode and freeze all layers. In such a case the model-based measures yield similar scores to similar text pairs regardless of the number of attempts or any hardware properties.

Another blocker to the usage of a certain measure could be a long inference time. We conduct additional experiments by calculating the average inference time per sample for a subset of measures representing each class w.r.t. the average correlation of the measure on the task. We concatenate texts from both tasks into two united datasets. For trainable measures, we use a data loader with a batch size equal to eight. We load all trainable models to NVIDIA GeForce RTX 2080 Ti. All other measures are calculated sample-wise on Intel(R) Xeon(R) Gold 5217 CPU @ 3.00GHz . We plot the results on Figure 6.

The most optimal measures are located at the bottom right corner of these plots, which means that the measure requires the least possible computational time and at the same time demonstrates a high correlation with human judgments. For the PG task, the MIS measure demonstrates the best performance and its average inference time is at the approximately same level as most of the other model-based measures. For TST task symmetric and asymmetric cross-encoders are the most optimal.

8 Conclusions

As our experiments show, encoder-based measures of content preservation correlate with human judgments much better than the traditional word or character-based measures such as BLEU on a wide range of datasets. In all paraphrase datasets and 9 out of 10 text style transfer datasets, the best-performing measures are based on the cross-encoder or bi-encoder architecture.

We suggest a measure called MIS which is based on the idea that texts with similar meanings mutually entail each other. We show that the proposed architecture outperforms other measures in the evaluation of paraphrases and performs on par with the top-performing measures in the evaluation of text style transfer. More specifically, it is particularly successful in transferring between contemporary and old English and between formal and informal styles. Thus, we recommend using this measure for content preservation scoring for paraphrases and TST tasks in the aforementioned tasks and to use BLEURT for other TST tasks.

While the best measures in our analysis improve over the popular ones (e.g. BLEU) by a large margin, their correlation with human judgments is still far from perfect. We expect that even better measures of content preservation will be proposed in the nearest future. We also hope that the MIS measure and the performed large scale computational study could be applied to other NLP tasks, such as machine translation, text summarization, etc.

Acknowledgements

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

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. 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*, pages 1321–1336, Online and Punta Cana, Dominican Republic. 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*, pages 1061–1071, Online. 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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

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

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

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

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, Hong Kong. Association for Computational Linguistics.

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

Yao Fu, Yansong Feng, and John P. Cunningham. 2020. Paraphrase generation with latent bag of words.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongyu Gong, Linfeng Song, and Suma Bhat. 2020. Rich syntactic and semantic information helps unsupervised text style transfer. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 113–119, Dublin, Ireland. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252, Online. Association for Computational Linguistics.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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*, pages 1576–1586, Lisbon, Portugal. 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*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Mingxuan Hu and Min He. 2021. Non-parallel text style transfer with domain adaptation and an attention model. *Appl. Intell.*, 51(7):4609–4622.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

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.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. 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 Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052, Florence, Italy. Association for Computational Linguistics.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta

Cana, Dominican Republic. Association for Computational Linguistics.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. 2019. Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3579–3584, Hong Kong, China. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Léo 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*, pages 1442–1461, Online. Association for Computational Linguistics.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102, Online. Association for Computational Linguistics.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019a. 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)*, pages 3304–3313, Hong Kong, China. Association for Computational 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. 2020. DGST: a dual-generator network for text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7131–7136, Online. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019b. Decomposable neural paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3403–3414, Florence, Italy. Association for Computational Linguistics.

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

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. *arXiv preprint arXiv:1907.11692*.

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.

Lucas Maystre and Matthias Grossglauser. 2016. Fast and accurate inference of plackett – luce models supplementary material.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

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.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7106–7116, Online. Association for Computational Linguistics.

Alexander Panchenko and Olga Morozova. 2012. A study of hybrid similarity measures for semantic relation extraction. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 10–18.

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.

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*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

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 (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Rafal Józefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444.

Chinmay Rane, Gaël Dias, Alexis Lechervy, and Asif Ekbal. 2021. Improving neural text style transfer by introducing loss function sequentiality. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2197–2201. ACM.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), pages 3786–3800, Online. Association for Computational Linguistics.

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

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation.

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

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi S. Jaakkola. 2020a. Educating text autoencoders: Latent representation guidance via denoising. 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 8719–8729. PMLR.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020b. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198, Online. Association for Computational Linguistics.

Yukai Shi, Sen Zhang, Chenxing Zhou, Xiaodan Liang, Xiaojun Yang, and Liang Lin. 2021. GTAE: graph transformer-based auto-encoders for linguistic-constrained text style transfer. *ACM Trans. Intell. Syst. Technol.*, 12(3):32:1–32:16.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2019. Zero-shot fine-grained style transfer: Leveraging distributed continuous style representations to transfer to unseen styles. *CoRR*, abs/1911.03914.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *CoRR*, abs/1811.00552.

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)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018a. Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018b. Structured content preservation for unsupervised text style transfer. *CoRR*, abs/1810.06526.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. 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.

Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2018. A task in a suit and a tie: paraphrase generation with semantic augmentation. *CoRR*, abs/1811.00119.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. 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 Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883, Florence, Italy. Association for Computational Linguistics.

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

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 (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. 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.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *CoRR*, abs/1903.06353.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220.

Qian Yang, Zhouyuan Huo, Dinghan Shen, Yong Cheng, Wenlin Wang, Guoyin Wang, and Lawrence Carin. 2019a. An end-to-end generative architecture for paraphrase generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3132–3142, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019b. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

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

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018b. Style transfer as unsupervised machine translation. *CoRR*, abs/1808.07894.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5897–5906. PMLR.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Datasets

Name	Comment	Size
ETPC	all data from textual_np_pos and textual_np_neg files	6004
PAWS-qqp	dev_and_test.tsv from qqp part used	677
PAWS-wiki	Test split from PAWS-Wiki Labeled (Final)	8000
Twitter-URL	Test split used	10120
PIT	Test split used	972
MSR	Test split used	1630
APT	Test split used (ap-h-test)	1252
Yam. para	Data from Paralex,Parphrase folder used	3223
SICK	Test split form SICK_test_annotated used	4927

Table 5: Paraphrase generation (PG) datasets used in the experiments.

Name	Comment	Size	Style
Tox600	All data used	600	Toxic
Yam. Yelp	Yelp subset of annotated data	2000	sentiment
Yam. GYAFC	GYAFC subset of annotated data	6000	Formality
Yam. Bible	Bible subset of annotated data	2000	Old-style English
xformal-FoST	English subset of annotated data use (meta_gyafc_en.tsv)	2458	Formality
CAE	All data used. For each sentence pair, the mean human score was used. The dataset was obtained by direct request to Laugier et al. (2021)	500	Toxic
PG	All data used. Individual ranks were induced from side-by-side comparisons using the Luce spectral ranking model. The dataset was obtained by direct request to Pang and Gimpel (2019).	886	Sentiment
STRAP_coha	For each sentence pair, the mean human score was used. All data used	100	Historical American English
STRAP_form		684	Formality
STRAP_SP		550	Old-style English

Table 6: Text style transfer (TST) datasets used in the experiments.

B Measures analysis

Citation	Measure	Task
Hu et al. (2017)	Automatic content preservation measures are not used	CG
Shen et al. (2017)	Automatic content preservation measures are not used	TST
Mueller et al. (2017)	Edit distance	CG
Jhamtani et al. (2017)	PINC (Chen and Dolan, 2011), BLEU	TST
Radford et al. (2017)	Only style accuracy analyzed	TST
Logeswaran et al. (2018)	round-trip BLEU	CG
Subramanian et al. (2018)	self-BLEU	TST
Zhang et al. (2018b)	BLEU	TST
Prabhumoye et al. (2018)	Manual pairwise comparison only	TST
Tian et al. (2018b)	self-BLEU, POS-distance - noun difference between the original and transferred sentences	TST
Yang et al. (2018)	self-BLEU	TST
Rao and Tetreault (2018)	STS CNN model (He et al., 2015)	TST
Carlson et al. (2018)	PINC, BLEU	TST
Zhao et al. (2018)	BLEU	TST
Fu et al. (2018)	Cossim between averaged or max/min-pooled GloVe (Pennington et al., 2014) embeddings	TST
Xu et al. (2018)	BLEU	TST
Zhang et al. (2018a)	BLEU	TST
Gupta et al. (2018)	BLEU, ROUGE, METEOR	PG
Pang and Gimpel (2019)	Cossim between GloVe (Pennington et al., 2014) embeddings weighted by inverse document frequency	TST
Li et al. (2018)	BLEU	TST
Smith et al. (2019)	self-BLEU	TST
Sudhakar et al. (2019)	self-BLEU	TST
Wu et al. (2019b)	BLEU	TST
John et al. (2019)	Cossim between averaged or max/min-pooled GloVe (Pennington et al., 2014) embeddings	TST
Luo et al. (2019)	BLEU	TST
Dai et al. (2019)	self-BLEU	TST
Jain et al. (2019)	BLEU, spacy.docsim	TST
Lai et al. (2019)	self BLEU	TST
Wang et al. (2019)	BLEU	TST
Xu et al. (2019)	BLEU	TST
Kajiwara (2019)	BLEU, F1-score over added, deleted, adn kept words	PG
Wu et al. (2019a)	Case insensitive BLEU	TST
Li et al. (2019a)	BLEU	TST
Li et al. (2019b)	BLEU, ROUGE	PG
Chen et al. (2019)	BLEU, ROUGE, METEOR	PG
Yang et al. (2019a)	BLEU, METEOR, TER (Snover et al., 2006)	PG
Egonmwan and Chali (2019)	BLEU, ROUGE, METEOR,GMS and EACS (Sharma et al., 2017)	PG
Wang et al. (2018)	BLEU, METEOR, TER (Snover et al., 2006)	PG
Krishna et al. (2020)	SIMILEWieting et al. (2019)	TST
Shen et al. (2020b)	self-BLEU	CG
Li et al. (2020)	self-BLEU	TST
Xu et al. (2020)	self-BLEU	TST
Gong et al. (2020)	Cossim between averaged or max/min-pooled GloVe embeddings	TST
Zhang et al. (2020)	BLEU	TST
Shen et al. (2020a)	BLEU	CG
He et al. (2020)	self-BLEU	TST
Goyal and Durrett (2020)	BLEU	PG
Fu et al. (2020)	BLEU, ROUGE	PG
Laugier et al. (2021)	BLEU, cosine sinilarity of USE (Cer et al., 2018)	TST
Lai et al. (2021)	BLEU, BLEURT (Sellam et al., 2020)	TST
Shi et al. (2021)	WMD (Kusner et al., 2015), BLEU, BERTScore (Zhang et al., 2019)	TST
Riley et al. (2021)	self-BLEU	TST
Krause et al. (2021)	Only detoxicifcation and fluency analyzed	CG
Lee et al. (2021)	BLEU, BERTScore (Zhang et al., 2019)	TST
Cao et al. (2020)	BLEU	TST
Rane et al. (2021)	BLEU	TST
Hu and He (2021)	Word Overlap, BLEU, cosine similarity between avearged or max/min-pooled GloVe (Pennington et al., 2014) embeddings	TST
Sun et al. (2021)	BLEU, ROUGE, METEOR	PG

Table 7: Automatic content preservation measures used in recent works on text style transfer (TST), paraphrase generation (PG), and controllable generation (CG).

Measure name in report	Comment	Article
RobNLI/*	Combination or separate use of NLI scores in direct or reverse direction	Nie et al. (2020)
SIMILE	Cosine similarity between embeddings generated with LSTM-based model	Wieting et al. (2019)
w2v_wmd_norm	Word mover distance with word2vec normalized	Kusner et al. (2015)
w2v_wmd	Word mover distance with word2vec	
w2v_l2	Euclidean distance vetween word2vec	
w2v_cossim	Cosine similarity over word2vec	
USE	Cosine similarity between embeddings generated with Universal Sentence Encoder	Cer et al. (2018)
SIMCSE-UL SIMCSE-UB SIMCSE-ULu SIMCSE-UBu SIMCSE-SL SIMCSE-SB SIMCSE-SBertUnc	Unsupervised and supervised version of SIMCSE:Simple Contrastive Learning of Sentence Embeddings. Unsupervised version trained to predict the input sentence itself with only dropout used as noise. Supervised version trained to produce embeddings on NLI data in contrastive manner using entailing sample as positive sample and contradiction as negative.	Gao et al. (2021)
LaBSE	Cosine similarity between language-agnostic cross-lingual sentence embeddings	Feng et al. (2020)
BERT-base-NLI-STSB		Reimers and Gurevych (2019)
ROUGEL	ROUGE Longest Common Subsequence	Lin (2004)
ROUGE3	ROUGE with trigram	
ROUGE2	ROUGE with bigram	
ROUGE1	ROUGE with unigram	
NUBIA	Multi-module pipeline consisting of Feature Extraction, Aggregation and Calibration for semantic similarity scoring	Kane et al. (2020)
FBrobNLI/*	Combination or separate use of Facebook roberta NLI model's scores in direct or reverse direction	Liu et al. (2019)
MoverScore	Special case of Earth Mover's Distance applied to BERT embeddings	Zhao et al. (2019)
METEOR	The measure is based on the harmonic mean of unigram precision and recall	Banerjee and Lavie (2005)
Levenshtein	The minimum number of single-character edits	Levenshtein et al. (1966)
Jaro_winkler	String measure measuring an edit distance between two sequences with special modification giving more rating to strings that match from the beginning for a set prefix	Jaro (1989)
fasttext_wmd_norm	Normalized word mover distance over fasstext vectors	Kusner et al. (2015)
fasttext_wmd	Word mover distance over fasstext vectors	
fasttext_l2	Euclidean distance between fasttext vectors	
fasttext_cossim	Cosine similarity between fasttext vectors	
facebook/bart-large-cnn	Weighted log probability of one text y given another text x. The weights are used to put different emphasis on different tokens	Lewis et al. (2020)
BLEURT-L512 BLEURT-L128 BLEURT-B512 BLEURT-B128	BERT fine-tuned for semantic similarity evaluation task in cross-encoder manner on sythetic data	Sellam et al. (2020)
deberta/*	Combination or separate use of NLI scores from deberat model in direct or reverse direction	He et al. (2021)
cross-stsb-large cross-stsb-base	Base and Large version of CrossEncoder trained on STSb	Reimers and Gurevych (2019)
APD	Paraphrase detector trained on the Adversarial Paraphrasing dataset from the correponding paper	Nighojkar and Licato (2021)
chrf	Character n-gram F-score	Popović (2015)
BLEU	Modified unigram precision score	Papineni et al. (2002)
bertscore/roberta-large bertscore_Bert-bmc bertscore-Mic-Deberta	F1-score over BERT-embeddings between tokens from initial and target setneces. The packages are: roberta-large, Bert base multilingal cased, microsoft/deberta-xlarge-mnli correspondingly	Zhang et al. (2019)

Table 8: The full list of the automatic measures of content preservation used in the analysis.

	ETPC	PAWS-qqp	PAWS-wiki	Twitter-URL	PIT	MSRP	APT	SICK	Yam. para	mean
MIS (2x-asym-cross-enc)	0.56	0.48	0.60	0.66	0.63	0.47	0.59	0.81	0.67	0.61
deberta/mean (2x-asym-cross-enc)	0.68	0.53	0.63	0.52	0.44	0.51	0.71	0.70	0.65	0.60
RobNLI/mean (2x-asym-cross-enc)	0.69	0.50	0.57	0.54	0.44	0.48	0.81	0.66	0.64	0.59
RobNLI/prod (2x-asym-cross-enc)	0.67	0.51	0.57	0.53	0.42	0.51	0.82	0.65	0.62	0.59
deberta/prod (2x-asym-cross-enc)	0.67	0.53	0.62	0.49	0.42	0.53	0.71	0.65	0.64	0.58
FBrobNLI/mean (2x-asym-cross-enc)	0.69*	0.40	0.57	0.46	0.40	0.49	0.69	0.64	0.62	0.55
RobNLI/f1 (2x-asym-cross-enc)	0.63	0.51	0.56	0.46	0.32	0.47	0.82*	0.62	0.59	0.55
cross-stsb-base (sym-cross-enc)	0.58	0.21	0.19	0.73	0.68	0.58*	0.47	0.82	0.70	0.55
SIMCSE-SL (emb-bi-enc)	0.40	0.45	0.34	0.73	0.61	0.48	0.37	0.82*	0.70	0.54
FBrobNLI/prod (2x-asym-cross-enc)	0.67	0.40	0.56	0.43	0.36	0.52	0.70	0.61	0.59	0.54
deberta/f1 (2x-asym-cross-enc)	0.64	0.53	0.61	0.39	0.32	0.47	0.71	0.57	0.60	0.54
NUBIA (asym-cross-enc)	0.57	0.27	0.32	0.65	0.59	0.55	0.42	0.80	0.67	0.54
cross-stsb-large (sym-cross-enc)	0.44	0.10	0.18	0.74	0.71*	0.56	0.50	0.81	0.72*	0.53
deberta/reverse (asym-cross-enc)	0.64	0.49	0.61	0.37	0.38	0.39	0.62	0.53	0.62	0.52
RobNLI/reverse (asym-cross-enc)	0.63	0.48	0.56	0.47	0.35	0.39	0.65	0.51	0.59	0.51
deberta/direct (asym-cross-enc)	0.63	0.54*	0.64*	0.40	0.24	0.41	0.45	0.66	0.60	0.51
SIMCSE-SB (emb-bi-enc)	0.38	0.31	0.27	0.72	0.55	0.48	0.34	0.81	0.70	0.51
RobNLI/direct (asym-cross-enc)	0.63	0.49	0.56	0.39	0.27	0.41	0.48	0.66	0.58	0.50
BERT-base-NLI-STSB (emb-bi-enc)	0.43	0.33	0.27	0.67	0.45	0.54	0.35	0.77	0.68	0.50
SIMCSE-SBertUnc (emb-bi-enc)	0.38	0.30	0.25	0.72	0.50	0.46	0.35	0.80	0.71	0.50
BLEURT-L128 (asym-cross-enc)	0.37	0.26	0.35	0.64	0.51	0.50	0.39	0.73	0.70	0.49
FBrobNLI/f1 (2x-asym-cross-enc)	0.63	0.40	0.56	0.29	0.22	0.46	0.70	0.53	0.54	0.48
BLEURT-L512 (asym-cross-enc)	0.27	0.23	0.31	0.62	0.53	0.49	0.39	0.72	0.69	0.47
SIMCSE-ULu (emb-bi-enc)	0.36	0.22	0.25	0.69	0.54	0.47	0.30	0.74	0.68	0.47
bertscore-Mic-Deberta (emb-context)	0.14	0.40	0.46	0.66	0.55	0.49	0.31	0.63	0.59	0.47
FBrobNLI/reverse (asym-cross-enc)	0.64	0.38	0.56	0.31	0.31	0.36	0.62	0.48	0.57	0.47
SIMCSE-UL (emb-bi-enc)	0.40	0.25	0.12	0.68	0.57	0.48	0.30	0.71	0.70	0.47
USE (emb-bi-enc)	0.35	0.16	0.09	0.72	0.55	0.44	0.34	0.76	0.71	0.46
BLEURT-B128 (asym-cross-enc)	0.27	0.23	0.31	0.60	0.48	0.48	0.34	0.71	0.67	0.45
FBrobNLI/direct (asym-cross-enc)	0.63	0.39	0.56	0.32	0.19	0.39	0.41	0.62	0.56	0.45
SIMCSE-UBu (emb-bi-enc)	0.43	0.21	0.15	0.68	0.48	0.43	0.29	0.72	0.67	0.45
BLEURT-B512 (asym-cross-enc)	0.32	0.21	0.28	0.58	0.43	0.48	0.33	0.73	0.66	0.45
APD (sym-cross-enc)	0.55	0.10	0.20	0.75*	0.27	0.35	0.61	0.54	0.63	0.44
bertscore/roberta-large (emb-context)	0.25	0.31	0.32	0.64	0.47	0.48	0.26	0.62	0.59	0.44
ROUGEL (ngram)	0.31	0.14	0.49	0.66	0.45	0.42	0.16	0.53	0.63	0.42
SIMCSE-UB (emb-bi-enc)	0.35	0.14	0.08	0.67	0.49	0.43	0.24	0.68	0.67	0.42
SIMILE (emb-bi-enc)	0.44	-0.13	-0.02	0.71	0.52	0.42	0.29	0.67	0.70	0.40
facebook/bart-large-cnn (emb-bi-enc)	0.18	0.32	0.45	0.50	0.39	0.37	0.12	0.63	0.55	0.39
bertscore_Bert-bmc (emb-context)	0.20	0.19	0.31	0.59	0.33	0.42	0.21	0.53	0.64	0.38
MoverScore (emb-context)	0.25	0.25	0.30	0.23	0.37	0.47	0.26	0.61	0.68	0.38
chrf (ngram)	0.26	0.16	0.24	0.63	0.36	0.39	0.18	0.55	0.61	0.38
ROUGE2 (ngram)	0.32	0.23	0.36	0.63	0.42	0.35	0.13	0.54	0.40	0.37
ROUGE1 (ngram)	0.32	-0.02	-0.00	0.67	0.49	0.45	0.22	0.58	0.63	0.37
BLEU (ngram)	0.21	0.06	0.07	0.63	0.38	0.47	0.23	0.54	0.62	0.36
METEOR (ngram)	0.30	-0.05	0.17	0.64	0.46	0.39	0.11	0.57	0.59	0.35
ROUGE3 (ngram)	0.30	0.15	0.43	0.56	0.40	0.29	0.10	0.48	0.25	0.33
fasttext_wmd (emb-static)	0.18	-0.09	-0.03	0.62	0.35	0.45	0.21	0.57	0.64	0.32
LaBSE (emb-bi-enc)	0.31	0.16	0.09	0.32	0.27	0.36	0.25	0.53	0.54	0.31
w2v_wmd (emb-static)	0.03	-0.07	-0.04	0.62	0.32	0.43	0.20	0.57	0.60	0.30
w2v_cossim (emb-static)	0.09	-0.07	-0.04	0.53	0.32	0.39	0.16	0.61	0.61	0.29
fasttext_wmd_norm (emb-static)	0.34	-0.09	-0.03	0.52	0.23	0.40	0.16	0.51	0.51	0.29
w2v_wmd_norm (emb-static)	0.05	-0.07	-0.04	0.51	0.25	0.43	0.20	0.58	0.60	0.28
Levenshtein (ngram)	0.24	0.32	0.37	0.27	0.08	0.26	0.06	0.37	0.52	0.28
fasttext_l2 (emb-static)	0.28	0.12	-0.03	0.42	0.21	0.32	0.12	0.45	0.35	0.25
Jaro_winkler (ngram)	0.16	0.06	0.13	0.47	0.28	0.20	0.15	0.34	0.43	0.25
w2v_l2 (emb-static)	-0.03	-0.06	-0.05	0.39	0.25	0.41	0.15	0.55	0.40	0.22
fasttext_cossim (emb-static)	0.16	-0.07	-0.04	0.39	0.19	0.29	0.16	0.47	0.43	0.22

Figure 7: Spearman correlations of all the evaluated measures with human judgments for paraphrase generation (PG) datasets. The measures are sorted by the mean correlation across all datasets. The top correlations for individual datasets are marked with *. The color palette of the heatmap is based on the regret, which is the difference between the correlation of the measure on a particular dataset and the best correlation on this dataset. The lower the value of regret, the higher quality.

	Tox600	STRAP_coha	STRAP_SP	CAE	Yam. Bible	STRAP_form	Yam. GYAFC	PG-YELP	Yam. Yelp	xformal-FoST	mean
SIMCSE-SL (emb-bi-enc)	0.46	0.45	0.74	0.46	0.62	0.70	0.69*	0.34	0.37	0.68*	0.55
cross-stsb-base (sym-cross-enc)	0.54	0.44	0.73	0.42	0.62	0.73	0.69	0.30	0.40*	0.62	0.55
cross-stsb-large (sym-cross-enc)	0.52	0.37	0.77*	0.44	0.62	0.74	0.69	0.34	0.39	0.62	0.55
MIS (2x-asym-cross-enc)	0.61	0.47	0.73	0.40	0.61	0.77*	0.69	0.29	0.28	0.62	0.54
SIMCSE-SB (emb-bi-enc)	0.47	0.43	0.72	0.43	0.63	0.68	0.69	0.35	0.38	0.65	0.54
BLEURT-L128 (asym-cross-enc)	0.61*	0.36	0.71	0.46	0.63	0.62	0.68	0.33	0.38	0.63	0.54
BLEURT-L512 (asym-cross-enc)	0.56	0.32	0.68	0.48*	0.63	0.60	0.67	0.42	0.36	0.65	0.54
NUBIA (asym-cross-enc)	0.54	0.46	0.72	0.32	0.62	0.70	0.67	0.31	0.35	0.62	0.53
SIMCSE-SBertUnc (emb-bi-enc)	0.44	0.33	0.69	0.46	0.62	0.61	0.69	0.34	0.37	0.62	0.52
SIMCSE-UL (emb-bi-enc)	0.37	0.37	0.68	0.46	0.63	0.62	0.68	0.38	0.35	0.62	0.52
BLEURT-B128 (asym-cross-enc)	0.50	0.26	0.67	0.43	0.63	0.54	0.66	0.42	0.35	0.62	0.51
BLEURT-B512 (asym-cross-enc)	0.53	0.28	0.68	0.41	0.63	0.57	0.65	0.37	0.35	0.61	0.51
USE (emb-bi-enc)	0.31	0.29	0.66	0.46	0.63	0.62	0.69	0.43	0.34	0.61	0.50
SIMCSE-UB (emb-bi-enc)	0.38	0.32	0.65	0.43	0.62	0.59	0.67	0.39	0.35	0.59	0.50
bertscore-Mic-Deberta (emb-context)	0.42	0.37	0.56	0.45	0.63	0.48	0.69	0.42	0.33	0.65	0.50
BERT-base-NLI-STSB (emb-bi-enc)	0.42	0.30	0.67	0.46	0.63	0.62	0.66	0.24	0.35	0.61	0.50
SIMCSE-ULu (emb-bi-enc)	0.38	0.28	0.65	0.43	0.62	0.54	0.68	0.39	0.34	0.58	0.49
SIMCSE-UBu (emb-bi-enc)	0.36	0.24	0.63	0.46	0.62	0.52	0.67	0.37	0.34	0.57	0.48
SIMILE (emb-bi-enc)	0.34	0.28	0.65	0.40	0.63	0.49	0.67	0.38	0.33	0.61	0.48
RobNLI/mean (2x-asym-cross-enc)	0.54	0.44	0.67	0.22	0.60	0.74	0.69	-0.05	0.17	0.63	0.47
RobNLI/prod (2x-asym-cross-enc)	0.51	0.40	0.65	0.25	0.60	0.72	0.68	-0.04	0.18	0.63	0.46
deberta/mean (2x-asym-cross-enc)	0.50	0.44	0.67	0.26	0.60	0.74	0.68	-0.04	0.13	0.58	0.46
MoverScore (emb-context)	0.27	0.26	0.48	0.47	0.63	0.39	0.68	0.42	0.33	0.57	0.45
deberta/prod (2x-asym-cross-enc)	0.46	0.41	0.65	0.29	0.60	0.71	0.68	-0.03	0.12	0.58	0.45
RobNLI/f1 (2x-asym-cross-enc)	0.46	0.36	0.61	0.25	0.60	0.68	0.68	-0.02	0.18	0.63	0.44
bertscore/roberta-large (emb-context)	0.27	0.25	0.47	0.45	0.63	0.38	0.66	0.39	0.32	0.59	0.44
APD (sym-cross-enc)	0.27	0.31	0.52	0.24	0.58	0.62	0.67	0.30	0.29	0.59	0.44
FBrobNLI/mean (2x-asym-cross-enc)	0.49	0.45	0.64	0.23	0.60	0.72	0.67	-0.10	0.02	0.61	0.43
deberta/f1 (2x-asym-cross-enc)	0.42	0.38	0.60	0.29	0.60	0.67	0.67	-0.02	0.12	0.58	0.43
RobNLI/reverse (asym-cross-enc)	0.41	0.33	0.54	0.28	0.59	0.65	0.67	0.02	0.15	0.62	0.43
RobNLI/direct (asym-cross-enc)	0.53	0.42	0.65	0.14	0.60	0.67	0.66	-0.06	0.17	0.46	0.42
deberta/direct (asym-cross-enc)	0.52	0.51*	0.64	0.19	0.58	0.67	0.66	-0.05	0.10	0.40	0.42
deberta/reverse (asym-cross-enc)	0.38	0.35	0.55	0.31	0.60	0.66	0.67	0.00	0.13	0.54	0.42
facebook/bart-large-cnn (emb-bi-enc)	0.37	0.07	0.51	0.35	0.59	0.42	0.60	0.39	0.31	0.57	0.42
FBrobNLI/prod (2x-asym-cross-enc)	0.45	0.40	0.59	0.23	0.60	0.67	0.67	-0.09	0.02	0.61	0.41
BLEU (ngram)	0.19	0.28	0.30	0.42	0.64*	0.26	0.68	0.37	0.31	0.61	0.41
FBrobNLI/direct (asym-cross-enc)	0.52	0.44	0.63	0.16	0.58	0.64	0.63	-0.10	0.01	0.46	0.40
chrf (ngram)	0.15	0.20	0.34	0.44	0.63	0.26	0.67	0.36	0.32	0.58	0.40
w2v_cossim (emb-static)	0.24	0.10	0.38	0.42	0.60	0.39	0.63	0.33	0.31	0.54	0.40
FBrobNLI/f1 (2x-asym-cross-enc)	0.42	0.36	0.54	0.23	0.59	0.62	0.64	-0.07	0.01	0.61	0.39
bertscore_Bert-bmc (emb-context)	0.16	0.15	0.37	0.39	0.63	0.30	0.66	0.43*	0.32	0.52	0.39
fasttext_wmd (emb-static)	0.15	0.16	0.31	0.44	0.63	0.27	0.68	0.38	0.32	0.56	0.39
ROUGE1 (ngram)	0.15	0.19	0.31	0.43	0.63	0.27	0.68	0.33	0.31	0.55	0.39
w2v_wmd (emb-static)	0.15	0.17	0.31	0.41	0.64	0.26	0.67	0.39	0.31	0.55	0.39
FBrobNLI/reverse (asym-cross-enc)	0.40	0.34	0.49	0.24	0.58	0.61	0.63	-0.04	0.01	0.59	0.38
w2v_wmd_norm (emb-static)	0.17	0.11	0.31	0.40	0.63	0.33	0.66	0.36	0.31	0.56	0.38
ROUGEL (ngram)	0.15	0.10	0.28	0.42	0.64	0.24	0.68	0.33	0.31	0.55	0.37
METEOR (ngram)	0.11	0.06	0.37	0.39	0.62	0.27	0.66	0.36	0.31	0.44	0.36
w2v_l2 (emb-static)	0.25	0.12	0.26	0.37	0.56	0.34	0.59	0.27	0.29	0.53	0.36
fasttext_wmd_norm (emb-static)	0.10	0.24	0.15	0.42	0.61	0.18	0.62	0.36	0.30	0.52	0.35
ROUGE2 (ngram)	0.13	0.10	0.21	0.43	0.64	0.20	0.67	0.30	0.31	0.51	0.35
LaBSE (emb-bi-enc)	0.18	0.24	0.26	0.25	0.57	0.29	0.48	0.30	0.24	0.48	0.33
fasttext_l2 (emb-static)	0.12	0.16	0.08	0.43	0.54	0.14	0.54	0.31	0.26	0.49	0.31
ROUGE3 (ngram)	0.12	0.02	0.17	0.42	0.64	0.15	0.58	0.24	0.26	0.49	0.31
fasttext_cossim (emb-static)	0.08	-0.02	0.13	0.41	0.57	0.18	0.57	0.27	0.27	0.49	0.30
Levenshtein (ngram)	0.12	0.01	0.17	0.13	0.53	0.19	0.48	0.29	0.23	0.59	0.27
Jaro_winkler (ngram)	-0.02	0.13	-0.06	0.31	0.59	0.16	0.59	0.25	0.29	0.39	0.26

Figure 8: Spearman correlations of all the evaluated measures with human judgments for text style transfer (TST) datasets. The measures are sorted by the mean correlation across all datasets. The top correlations for individual datasets are marked with *. The color palette of the heatmap is based on the regret, which is the difference between the correlation of the measure on a particular dataset and the best correlation on this dataset. The lower the value of regret, the higher quality.

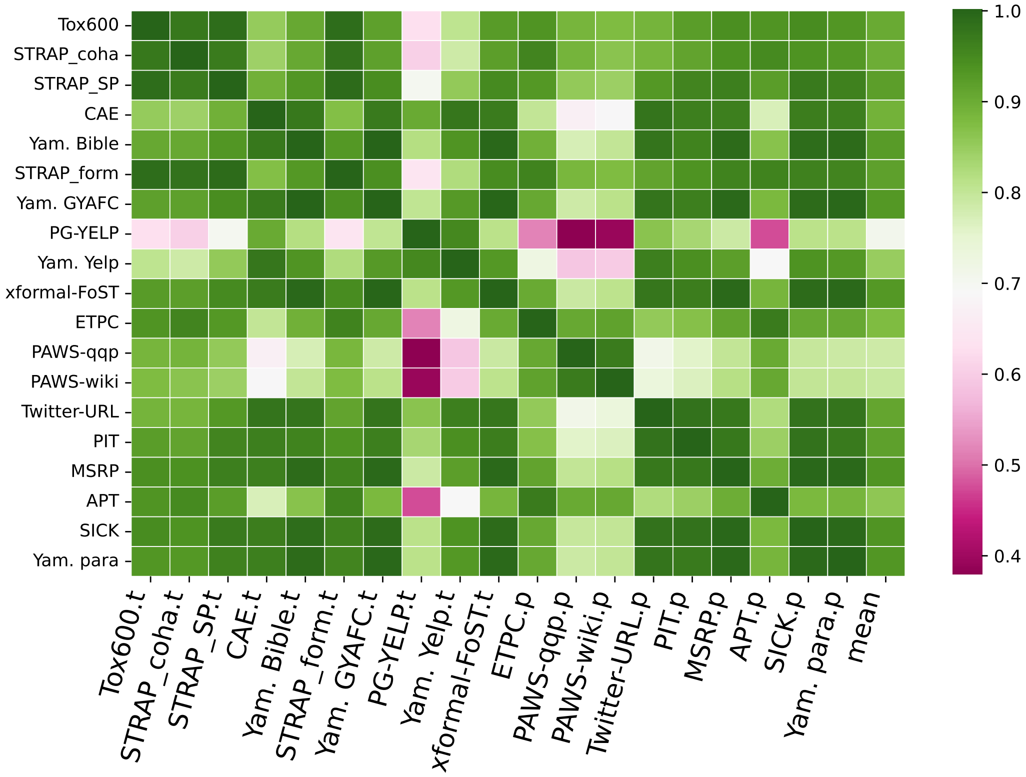

Figure 9: Cosine similarities of vectors of measures' correlations on individual datasets. The last column shows the mean cosine similarity of a dataset vector and vectors of all other dataset (excluding self-similarity). Postfixes 'p' and 't' indicate datasets for to PG and TST tasks, respectively.

C System-level ranking

system	human	MIS	RobNLI/mean	BLEURT-L128	cross-stsb-base	LaBSE	SIMCSE-SL	bertscore-Mic-Deberta	SIMILE	w2v_wmd	BLEU	chrf
paragedi	0.65	0.52	0.39	-0.25	0.82	0.95	0.68	0.76	0.67	-0.67	0.48	0.41
condbert	0.64	0.41	0.27	-0.26	1.07	0.96	0.75	0.83	0.76	-0.34	0.72	0.73
mask_infill	0.59	0.39	0.29	-0.29	0.96	0.99	0.82	0.87	0.82	-0.21	0.79	0.80

Table 9: System ranking on Tox600 (Dale et al., 2021), text detoxification.

system	human	MIS	RobNLI/mean	BLEURT-L128	cross-stsb-base	LaBSE	SIMCSE-SL	bertscore-Mic-Deberta	SIMILE	w2v_wmd	BLEU	chrf
nmt_combined	4.67	0.91	0.90	0.78	4.35	0.98	0.96	0.95	0.93	-0.15	0.88	0.85
pbmt	4.64	0.89	0.88	0.71	4.08	0.98	0.95	0.94	0.91	-0.17	0.85	0.81
ref	4.56	0.87	0.84	0.32	2.98	0.95	0.89	0.86	0.76	-0.44	0.64	0.59
nmt_copy	3.99	0.74	0.72	0.40	3.04	0.97	0.88	0.88	0.82	-0.26	0.77	0.73
nmt_baseline	3.90	0.73	0.70	0.40	3.00	0.96	0.87	0.89	0.82	-0.25	0.77	0.74

Table 10: System ranking on xformal-FoST (Briakou et al., 2021), formality transfer.

system	human	MIS	RobNLI/mean	BLEURT-L128	cross-stsb-base	LaBSE	SIMCSE-SL	bertscore-Mic-Deberta	SIMILE	w2v_wmd	BLEU	chrf
CAET rephrasing	2.63	0.34	0.28	-0.63	0.56	0.92	0.62	0.70	0.56	-0.66	0.47	0.44
IE rephrasing	2.20	0.37	0.36	-0.73	0.55	0.96	0.60	0.73	0.56	-0.56	0.58	0.56
ST (multi) rephrasing	2.10	0.26	0.22	-1.16	-0.22	0.91	0.52	0.63	0.60	-0.67	0.46	0.46
ST (cond) rephrasing	2.08	0.25	0.23	-1.11	-0.07	0.92	0.53	0.66	0.62	-0.65	0.49	0.47
CA rephrasing	1.88	0.05	0.08	-1.54	-2.22	0.90	0.18	0.51	0.16	-0.95	0.23	0.18

Table 11: System ranking on CAE (Laugier et al., 2021), text detoxification.

system	human	MIS	RobNLI/mean	BLEURT-L128	cross-stsb-base	LaBSE	SIMCSE-SL	bertscore-Mic-Deberta	SIMILE	w2v_wmd	BLEU	chrf
m7	3.41	0.16	0.09	-1.03	-0.84	0.95	0.45	0.76	0.52	-0.56	0.52	0.45
m6	3.03	0.18	0.11	-1.16	-0.58	0.95	0.45	0.74	0.56	-0.52	0.58	0.53
m2	3.03	0.14	0.07	-1.23	-1.10	0.94	0.37	0.73	0.47	-0.56	0.53	0.46
m0	2.31	0.10	0.06	-1.50	-1.80	0.91	0.28	0.64	0.30	-0.80	0.34	0.29

Table 12: System ranking on PG-YELP (Pang and Gimpel, 2019), sentiment transfer.

system	human	MIS	RobNLI/mean	BLEURT-L128	cross-stsb-base	LaBSE	SIMCSE-SL	bertscore-Mic-Deberta	SIMILE	w2v_wmd	BLEU	chrf
paraphrase_base	0.79	0.64	0.53	-0.39	1.19	0.94	0.77	0.74	0.65	-0.69	0.45	0.39
paraphrase_0.0	0.76	0.73	0.64	-0.08	1.91	0.94	0.82	0.77	0.71	-0.63	0.50	0.43
paraphrase_0.9	0.59	0.56	0.44	-0.45	1.04	0.93	0.73	0.71	0.61	-0.74	0.42	0.35
unmt	0.31	0.23	0.19	-0.95	-0.31	0.93	0.50	0.69	0.51	-0.61	0.51	0.43
he_2020	0.26	0.21	0.19	-0.99	-0.82	0.90	0.45	0.67	0.46	-0.65	0.45	0.40

Table 13: System ranking on STRAP_form, (Krishna et al., 2020), formality transfer.

system	human	MIS	RobNLI/mean	BLEURT-L128	cross-stsb-base	LaBSE	SIMCSE-SL	bertscore-Mic-Deberta	SIMILE	w2v_wmd	BLEU	chrf
paraphrase_0.0	0.81	0.62	0.58	-0.11	1.48	0.95	0.79	0.76	0.72	-0.69	0.44	0.37
paraphrase_base	0.58	0.44	0.43	-0.52	0.77	0.94	0.69	0.70	0.62	-0.79	0.37	0.31
he_2020	0.35	0.19	0.21	-1.07	-0.28	0.93	0.49	0.68	0.49	-0.65	0.46	0.40
unmt	0.26	0.12	0.13	-1.23	-0.92	0.93	0.41	0.66	0.41	-0.72	0.42	0.34

Table 14: System ranking on STRAP_SP (Krishna et al., 2020), Shakespeare style transfer.

dataset	measures
Tox600	MIS, BLEURT-L128
xformal-FoST	BLEURT-L128 , cross-stsb-base, SimCSE, BERTScore, and all other models
CAE	BLEURT-L128 , cross-stsb-base, SimCSE
PG-YELP	BLEURT-L128 , LaBSE, BERTScore
STRAP_form	LaBSE
STRAP_SP	MIS, BLEURT-L128 , cross-stsb-base, LaBSE, SimCSE, BERTScore, SIMILE

Table 15: The measures that correctly identify the best text style transfer system for each dataset.