An Empirical Study on Multi-Task Learning for Text Style Transfer and Paraphrase Generation

Paweł Bujnowski[a], Kseniia Ryzhova[c*], Hyungtak Choi[b], Katarzyna Witkowska[d*], Jarosław Piersa[a], Tymoteusz Krumholc[a] and Katarzyna Beksa[a]

[a] Samsung R&D Institute, Warsaw, Poland
[b] Samsung Research, Samsung Electronics Co. Ltd., Seoul, Korea
[c] Rankomat, Warsaw, Poland
[d] Polytechnic University of Catalonia, Barcelona, Spain
{p.bujnowski, ht777.choi, j.piersa,
t.krumholc, k.beksa}@samsung.com,
{ksenija.rijova, witek.witkowska}@gmail.com

Abstract

The topic of this paper is neural multi-task training for text style transfer. We present an efficient method for neutral-to-style transformation using the transformer framework. We demonstrate how to prepare a robust model utilizing large paraphrases corpora together with a small parallel style transfer corpus. We study how much style transfer data is needed for a model on the example of two transformations: *neutral-to-cute* on internal corpus and *modern-to-antique* on publicly available Bible corpora. Additionally, we propose a synthetic measure for the automatic evaluation of style transfer models. We hope our research is a step towards replacing common but limited rule-based style transfer systems by more flexible machine learning models for both public and commercial usage.

1 Introduction

The goal of text style transfer (ST[1]) is to convert the input sentence into an output, preserving the meaning but modifying the linguistic layer (grammatical or lexical).

Style transfer is extensively studied in academic papers (Li et al., 2018a; Rao and Tetreault, 2018; Carlson et al., 2018; Jhamtani et al., 2017) and also gains popularity in commercialized chatbots. Amazon Alexa introduced styles mimicking celebrities that replace the original Alexa voice (Amazon, 2020). Samsung offers applications personalized in terms of style, e.g. Celebrity Alarm (Samsung, 2019). Both examples offer novel user experience, though the number of available system responses seems to be limited. The two main constraints are voice generation and text content limitations. While voice generation can be implemented with voice synthesis systems, like e.g. by Jia et al. (2018) or Prenger et al. (2018), content limitation might be resolved by flexible machine learning text ST methods: the system could transform neutral answer into one of predefined styles using a machine learning model. The major challenges are the limited amount of style data and the lack of convincing automatic evaluation measures. In our study we try to mitigate these two issues.

The goal of our paper is to present an efficient method to train domain-unlimited ST models using a small style dataset. Inspired by the successful outcomes of multi-task learning (Caruana, 1997; Collobert and Weston, 2008; Johnson et al., 2017), we propose a transformer model (Vaswani et al., 2017) that jointly solves paraphrase generation and style transfer tasks. We hypothesize that training in the multi-task mode on an English large parallel corpus for paraphrasing may help preserve the input content along with successful adjustment of vocabulary and grammar to the target style, even using a small ST corpus.

To verify the hypothesis and add practical value, we perform detailed tests on various sizes of text ST corpora to verify how their volume affects the results. This is a convenient approach, because we have

[*] Work done during an internship at Samsung R&D Institute, Warsaw, Poland.

[1] Abbreviations used throughout the article: ST — Style Transfer, (N)MT — (Neural) Machine Translation, N2C — Neutral to Cute, M2A — Modern to Antique, TS — Text Simplification, DNNs — Deep Neural Networks.

publicly available large parallel corpora for paraphrasing, but rather small parallel corpora of style transformations we want to achieve. We perform our experiments using openly available paraphrase sources (Rao and Tetreault, 2018; Carlson et al., 2018; Quora, 2017; Wieting and Gimpel, 2018; Williams et al., 2018) and two ST corpora, one for each task: the internal "cute person" style corpus and the processed Bible texts corpus composed of publicly available sources. We examine two style transformations: *neutral-to-cute* and *modern-to-antique* (the latter one on different Bible translations).

Besides studies with various data volumes, we propose a method of creating an automatic measure. We show that simple common measures do not work if separated and instead we propose a fitted compound measure.

2 Related works

2.1 Style transfer: not too far from paraphrasing

Similarly to Xu et al. (2012) we see language ST as a task composed of two linked subtasks: paraphrasing and style adjustment. In our paper we claim that the ST task consists mostly in good paraphrasing and much less in adding the target style. Following this idea, we focus on background methods for the first task – paraphrasing, and then for style transfer.

Traditionally, the paraphrasing task involved rule and dictionary-based approaches (McKeown, 1983; Bolshakov and Gelbukh, 2004; Kauchak and Barzilay, 2006). Another popular method was the statistical paraphrase generation that recombined words probabilistically in order to create new sentences (Quirk et al., 2004; Wan et al., 2005; Zhao et al., 2009). Currently, DNNs are used for automatic paraphrasing, e.g. by sequence-to-sequence (seq2seq) models (Sutskever et al., 2014).

Presumably, the first paper presenting a deep learning model for the paraphrase task is the one by Prakash et al. (2016). The authors successfully compared their residual LSTM to previous LSTM-derived seq2seq models (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Bahdanau et al., 2014; Vaswani et al., 2017). In parallel, upon research on variational autoencoders (VAEs), e.g. Chung et al. (2015), other approaches to paraphrasing emerged (Bowman et al., 2016; Gupta et al., 2018).

More recently Li et al. (2018b) implemented paraphrase generation with deep reinforcement learning that proved better performance than the previous seq2seq results on Twitter (Lan et al., 2017) and Quora (Quora, 2017) datasets. Insufficient corpora for paraphrase generation gave motivation to new practical studies on unsupervised approach (Artetxe et al., 2018; Conneau and Lample, 2019). Recently, Roy and Grangier (2019) proposed a monolingual system for paraphrasing (without translation) and compared it to the unsupervised and translation methods, presenting various linguistic characteristics for each of them.

2.2 Style transfer methods

We name here only some existing solutions. Xu et al. (2012) used a phrase-based MT method on Shakespeare M2A corpus. This research was followed by Jhamtani et al. (2017), who improved the results with the DNN seq2seq approach using a copy mechanism.

Rao and Tetreault (2018) adapted phrased-based and neural MT models for formality transfer using GYAFC corpus. Later Niu et al. (2018) improved results on GYAFC by creating a multi-task system for both formality transfer and English-French MT.

It is worth mentioning Dryjański et al. (2018) who used DNNs both elements: generation of ST phrases and their positions related to the input sentence.

Another distinctive study was conducted for text simplification (TS) task with the matched Wikipedia–Simple Wikipedia parallel data, e.g. Wubben et al. (2012; Wang et al. (2016). What is significant for TS are the results for automatic measures in Xu et al. (2016) followed by Alva-Manchego et al. (2019). We draw our inspiration from the authors, along with the results of Xu et al. (2012) in our measure propositions.

Our solution has common features with Wieting and Gimpel (2018). The authors demonstrated that using pretrained embeddings from a large parallel paraphrase corpus ($\sim$50 millions) and out-of-the-box models, it was possible to reach state-of-the-art results on several SemEval semantic textual similarity competitions. In our work, instead of using pretrained embeddings we follow Johnson et al. (2017) and train the multi-task model on a single language, but with paraphrases and a small *neutral-to-style* dataset. Compared to the previous solutions, our approach can be seen as a universal method to tackle text ST (e.g. formality transfer, simplification and more).

3 Model

3.1 Multilingual model

We used the Multilingual Transformer (Vaswani et al., 2017) model from the Fairseq package (Ott et al., 2019). Using this model we approach the problem similarly to multilingual translation, treating each style as a new language. An overview of the system is presented in Figure 1.

In this architecture all the language pairs share a single Transformer neural network. We fed the model with two paired datasets, {English sentences vs English paraphrase references} and {English sentences vs English sentences with target style}, and trained it on both sets. The parallel training on both datasets following this multilingual (multi-task) approach produces robust results, but requires retraining the model from scratch after any modification of the style corpus. For Multilingual Transformer we preprocessed the sentences with the SentencePiece toolkit (Kudo and Richardson, 2018) without any pretokenization. The vocabulary size was predefined to 16k. We used a shared English dictionary for both the paraphrase pairs and the target style corpus binarization. We also removed lines with more than 250 tokens.

For each training set the target style sample constituted only up to 0.6% of the whole set. It appears that Multilingual Transformer can effectively train the model even with a huge disproportion between paraphrases and style corpora sizes.

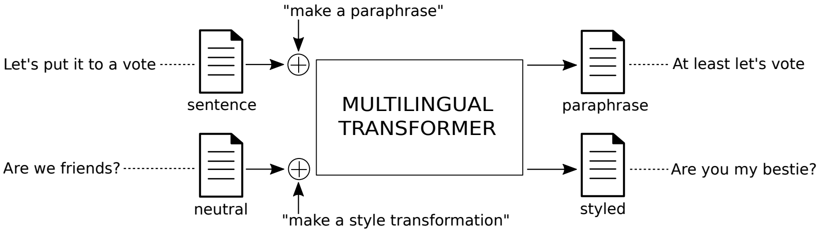


Figure 1: Model overview: Multilingual transformer trained with paraphrases and styled data.

3.2 Training parameters details

For training the models we used multiple GPUs (GeForce GTX 1080 Ti, 11GB), which contributed to more representative updates due to different average sentence length among mini-batches distributed between workers. The training on 8 GPUs lasted about 27 hours. We used Multilingual Transformer models with shared-decoders from the Fairseq package. The architecture consisted of 6 encoder and 6 decoder fully connected layers of dimension 1054 with 4 attention heads. The embeddings were of dimension 512. We set the dropout to 0.3, weight decay to 0.0001 and the optimizer to Adam ($\beta_1 = 0.9, \beta_2 = 0.98$) with the learning rate equal to 0.0005. We used label-smoothed cross entropy loss with label smoothing set to 0.1. For generation we used beam size equal to 5. We stopped the training after 40 epochs.

4 Experiments

Using our ST methods we performed two tasks: N2C and M2A. For each task we prepared a number of target (style) parallel corpora differing only in volumes. For the N2C transformation it was 1k, 3k, 5k, 7k, 10k, 13k, and 17k. For the M2A transformation we prepared the same corpora volumes plus the additional 30k. The proportions for training and validation were equal for all the target datasets with the ratio of 80%:20% respectively. The style corpora are described in subsections 4.1.2 and 4.1.3. As supplementary data, we used much larger parallel paraphrases corpora described in subsection 4.1.1. Firstly, paraphrase data is required for producing high quality sentences. Secondly, the generated utterances must be properly *tuned* to the target style. In our experiments we searched for answers to the following questions.

1. How much style data is needed in multilingual training for style transformations of high quality?
2. How to automatically evaluate the major aspects of ST models? How to create a synthetic measure?
3. What elements are successful and what challenges remain when using transformer models for ST?

4.1 Data

4.1.1 Paraphrases data

For each task we built a large parallel corpus, used as supplementary data in model training. We treat these corpora as domain-unspecific and stylistically neutral. As sources, we used the paraphrases corpus presented in Wieting and Gimpel (2018), MultiNLI corpus (Williams et al., 2018), Quora Kaggle dataset (Quora, 2017) and Bible corpus (Carlson et al., 2018).

Name	Size (lines)	Task
Paraphrases	5,000,000	Netural-to-Cute, Modern-to-Antique
MultiNLI	16,329	Netural-to-Cute, Modern-to-Antique
Quora	145, 000	Neutral-to-Cute, Modern-to-Antique
Bible	1,339,891	Netural-to-Cute

Table 1: Parallel "paraphrases" corpora used in Neutral-to-Cute and Modern-to-Antique tasks.

In the M2A task we removed the Bible subcorpus from the dataset of paraphrases to perform the multi-task training with a small amount of style data (like in the N2C task).

4.1.2 Neutral-to-Cute data

The N2C training corpus contains 17k lines, some longer than one sentence. The efforts of searching for available cute person style corpora turned out to be ineffective. Thus, the set was created with the use of an internal crowdsourcing platform by educated linguists with academic background and experience in style transfer projects. In a series of tasks the linguists rewrote input ("neutral") sentences into "cute person" style unless they were "cute" enough (pairs with no change constituted 15% of the total set). The "cute style" was described as "informal", "positive", "superlative", "excited" and "slangy". The "cute style" was usually created by inserting an adequate "cute style" phrase, paraphrasing a fragment or the whole sentence. The generated corpus covered numerous genres, styles and topics, e.g. self-presentation, jokes, facts, small talk and anecdotes. We also created a 300 lines long test corpus with four "cute style" candidate answers for each input sentence. Linguists were asked to follow the same guidelines as in the training corpus creation task, and additionally to keep a degree of variation between the candidate answers. The same set of neutral sentences is used both for human and automatic evaluation.

4.1.3 Modern-to-Antique Bible data

The idea of using the Bible as a parallel corpus suitable for ST tasks was presented in Carlson et al. (2018). The authors claim that traditionally used sentence and verse demarcation makes for easy sentence alignment. The dataset consists of 8 public domain available Bible versions.[2] For the "antiquification" task's purpose, as the input we chose World English Bible (WEB, released in 2000), being the most stylistically modern version. As the target style, we chose King James Version (KJV, released in 1611), as it is one of the most influential English versions of the Bible, which perfectly shows the "majesty of style".

Using these sources, we built a 30 thousand verse long parallel WEB-KJV Bible corpus. We additionally sampled 300 verses of pairs for human and automatic evaluation. The automatic test set was built in the same way as the one used in the N2C task. Each input (WEB-style) was paired with a target-style (KJV) sentence and three candidate answers from other Bible translations: Darby's, Young's Literal Translation and American Standard Version. For human evaluation we selected 150 Bible verses (half of the automatic sample). Additionally, we test out-of-domain ST capability of the model by adding 150 "neutral style" small talk sentences, previously used in the N2C task, to the M2A test corpus.

[2] `https://github.com/keithecarlson/StyleTransferBibleData`

4.1.4 Lexical diversity

Additionally, we tested the lexical diversity of corpora used in both tasks. For this purpose we used MTLD (measure of textual lexical diversity) index (McCarthy, 2005; McCarthy and Jarvis, 2010) for our two tasks (see Table 2). We assume that higher MTLD scores are associated with higher topical and lexical diversity of “neutral” and “cute” corpora. This may indicate the higher difficulty of the N2C task.

We also measured the number of tokens, types (unique tokens), mean and standard deviation of number of sentences and tokens per line. The study revealed the difference between M2A and N2C tasks. In the M2A task, output (KJV style) has two times less sentences than input (WEB style) while maintaining similar (10% lower) number of tokens. In contrast, in the second task, “cute” style outputs tend to have 50% more sentences and 33% more tokens than “neutral” style sentences (inputs).

Corpus	Tokens	Types	MTLD	Mean sentences/line	Mean tokens/line	Median tokens/line
WEB 17k	359871	9461	54.88	2.26 ± 0.80	32.65 ± 12,36	31
KJV 17k	346942	9469	54.18	1.12 ± 0.38	29.52 ± 12,63	27
Neutral 17k	147190	10876	129.48	1.5 ± 0.87	12.89 ± 10.4	10
Cute 17k	186455	12648	156.11	1.94 ± 1.04	17.17 ± 11.32	15

Table 2: MTLD score for N2C and M2A corpora. The higher score indicates higher lexical diversity.

4.2 Human evaluation method

For human evaluation, a panel of three language experts was employed for each style. Every judge evaluated the same 300 transformations in four criteria, using Likert scale: 1 – very bad, 2 – unacceptable, 3 – flawed, but acceptable, 4 – good, with minor errors, 5 – very good.

The criteria covered four aspects.

- **Language**: the correctness of grammar, spelling, vocabulary usage, lack of unnecessary repetitions or loops, etc.
- **Quality**: semantics, fluency, comprehensibility, logic and the general “feel” of the sentence.
- **Content**: the degree of semantic similarity to the input sentence.
- **Style**: the appropriateness of style in the output sentence.

Our *Content* is similar to *Meaning Preservation* used in Callison-Burch (2008) and Rao and Tetreault (2018), while their *Fluency* is splitted into our *Language* and *Quality*.

The judges were instructed on how to understand the cute person and antique styles. All the sets were evaluated by the same panel of language experts, in order to preserve the common understanding of criteria and keep the results coherent. The judges were attended by the panel of two supervisors who verified their understanding of the criteria. All the evaluators had academic background in linguistics and at least Bachelor’s Degree in this or related field.

The evaluated element was the transformation in relation to the input sentence. The examples below show that some transformation elements may produce parallelity of score values (compare Quality and Content). However, the opposite effect occurs as well – raising one score may cause lowering another one (compare Content and Style).

Score	**Neutral-to-Cute**	**Modern-to-Antique**
input	Neutral: *I don’t want to scare you, but right now there is a skeleton inside you.*	Modern: *Let it be, when these signs have come to you, that you do what is appropriate for the occasion; for God is with you.*
styled output	Cute: *Wow, don’t want to scare you, but right now there’s a skeleton inside you. So cool, so I’m sure we can try again!*	Antique: *Let it be, when these signs be come unto thee, that thou doest ought for the occasion; for God is with thee.*
Language	5 – The sentence is linguistically correct.	3 – The output contains language mistakes.
Quality	4 – The sentence is logical and semantically correct, but the phrase added at the end is a bit semantically separated from the input meaning.	3 – Small semantic distortion (*ought* used incorrectly as a synonym to *appropriate*).
Content	3 – The score was lowered for the additional sentence introducing a little new meaning.	3 – Small error: incorrect introduction of the word *ought* as a synonym to *appropriate*.
Style	5 – The added phrase *wow* and the contraction *there’s* are typical to the excited person style. Although unnecessarily introduced, the sentence which lowers the score for Content raises it for Style.	5 – The changes (*thee*, *thou*, *doest*) are compliant with the antique style.

Table 3: Examples of outputs with human evaluation scores.

The two 300-transformation sets (one for *neutral-to-cute* and one for *modern-to-antique*) were evaluated separately for all the trainings with various sizes of corpora (1k, 3k, 5k, 7k, 10k, 13k, 17k for both transformations plus 30k for M2A only) by three language experts and in four criteria (Language, Quality, Content and Style). It gives the total number of 54,000 single assessments.

We also measured the inter-annotator agreement using Krippendorff's alpha (Krippendorff, 2004) for each criterion and each task (see Table 4). The values in Krippendorff's alpha range from $\alpha = 0$ (perfect disagreement) to $\alpha = 1$ (perfect agreement). Customarily, $\alpha \geqslant 0.667$ is considered the minimum threshold for reliable annotation and $\alpha \geqslant 0.8$ is the optimal threshold (Krippendorff, 2004).

For three out of four criteria, inter-annotator agreement was higher in the M2A task. The agreement in Style assessment was significantly higher in the N2C task. We assume that this discrepancy is caused by the limited proficiency of judges in the biblical style.

Task	Language	Quality	Content	Style
Modern-to-Antique	0.715	0.774	0.829	0.683
Neutral-to-Cute	0.641	0.667	0.796	0.858

Table 4: Krippendorff's alpha for Neutral-to-Cute and Modern-to-Antique tasks human evaluation.

4.3 Human evaluation results

In this section we present the outcomes of our generative models for each dataset volume, focusing on the impact of the style data volume on style transformation quality. In order to prove the complexity of the problem, we adopted the proposed human evaluation measures (Language, Quality, Content and Style). As expected, dependencies between the target data volume used in models training and human measures statistics are not strictly linear. We can make an insight into this process. First, we focus on N2C transformations and then we move to Bible ST.

4.3.1 *Neutral-to-Cute* transformation

In Table 5 and in Figure 2 we present the study results for "cute person" data. We counted means and standard deviations of combined datasets of size $N = 900$, putting together the same three datasets of 300 sentences evaluated by independent linguists.

Corpus size	1k	3k	5k	7k	10k	13k	17k
Language (L): $\bar{x} \pm \sigma$	4.33 ± 1.07	**4.69 ± 0.76**	4.57 ± 0.88	4.67 ± 0.76	4.61 ± 0.89	4.58 ± 0.89	4.59 ± 0.91
Quality (Q): $\bar{x} \pm \sigma$	4.38 ± 1.08	**4.64 ± 0.84**	4.54 ± 0.95	4.64 ± 0.84	4.57 ± 0.96	4.50 ± 0.97	4.45 ± 1.04
Content (C): $\bar{x} \pm \sigma$	3.81 ± 1.48	**4.36 ± 1.16**	4.21 ± 1.28	4.27 ± 1.26	4.24 ± 1.25	3.99 ± 1.34	4.09 ± 1.31
Style (S): $\bar{x} \pm \sigma$	3.27 ± 1.72	3.37 ± 1.74	3.21 ± 1.80	3.52 ± 1.73	3.64 ± 1.65	3.92 ± 1.58	**4.04 ± 1.46**
Mean(L,C,S): $\bar{x} \pm \sigma$	3.81 ± 0.80	4.14 ± 0.74	4.00 ± 0.75	4.15 ± 0.71	4.17 ± 0.78	4.16 ± 0.78	**4.24 ± 0.76**
#(output=input): %	19.67%	27.33%	31.00%	26.67%	19.67%	13.67%	13.67%

Table 5: Human measures statistics of Neutral-to-Cute transformation across sizes of target corpora.

First of all, we notice that the scores for all datasets are quite high – above 3 for each human measure – including the smallest 1k style dataset. The lowest score, for Style (3.27) in the 1k style dataset, points that some transformations are stylistically flawed. The 3.81 score for Content shows that most of the generated sentences semantically reflect the input. The highest scores are those for Quality and Language – both over 4.30 – being more than good. The growth of the style dataset volume is also reflected by evaluation scores. The Style factor increases together with the data volume (except for the 5k set). The maximum Style score is 4.04 for the biggest 17k dataset, which demonstrates the high quality of style conversion. As opposed to the Style score, Language, Quality and Content do not improve with larger datasets sizes (between 7k and 17k). It may be due to the negative correlation between Style and the remaining human measures, which we analyze in section 4.4.

In Table 5 and in Figure 2 we added the average of arithmetic mean for Language, Content and Style (i.e. "Average human score"). Quality was omitted because its strong correlation with Language makes it too overlapping. The biggest difference in average human scores is between datasets 1k and 3k. The values for Content and Quality are lower for the largest volumes - 13k and 17k. The reason may be the specific N2C transformation, where more elaborate cute person phrases contrast with context-suitable

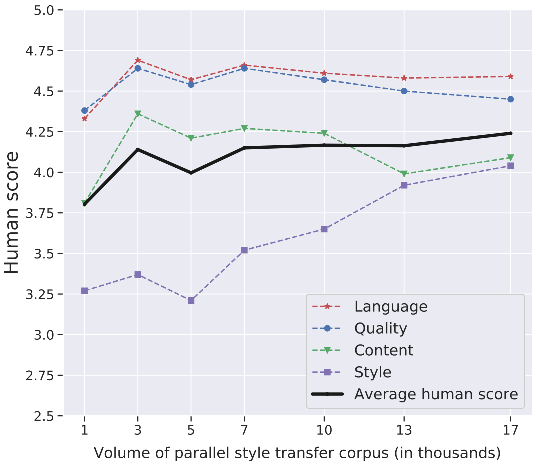


Figure 2: Human evaluation of Neutral-to-Cute transformation (legend: average human score is an arithmetic mean of Language, Content and Style).

ones. Closing the analysis, we also point out the repetition ratio of the input as the model output (the lower the better). The best results are for the largest datasets of 13k and 17k (13.67% in both cases – a bit below the mean of 15% in the training sets).

The following conclusions can be drawn from the N2C data transformation.

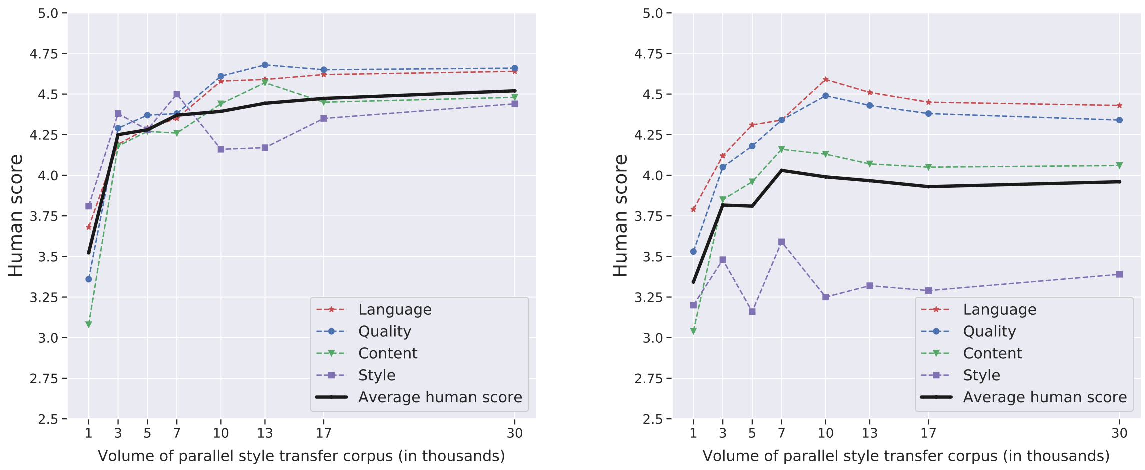
1. Style transformation is acceptable even for the 1k style dataset used for the multi-task training model.
2. For the 3k dataset there is the biggest growth in transformation quality, especially for vocabulary, semantics and logic (and smaller for style score).
3. The more style data we use, the better style transformation we obtain. The models trained with the datasets of 13k or larger, provide good quality with a low unchanged sentence ratio.

4.3.2 *Modern-to-Antique* transformation

In this part of the experiment we focus on the "antique" Bible style data as the model's output. We divided our evaluation into two tasks. In the first and main one we examine the evaluation where modern Bible data is used as the input. It is consistent with our model multi-task training, where besides the large corpus of paraphrases we take the style corpus of M2A Bible data. For this evaluation we use 150 sentences. As a supplementary test we employ the neutral dataset of 150 lines, similar to N2C data input. Table 6 and Figure 3 present the two evaluation studies.

Corpus size	1k	3k	5k	7k	10k	13k	17k	30k
Modern Bible input								
Language (L): $\bar{x} \pm \sigma$	3.68 ± 0.95	4.19 ± 0.93	4.29 ± 0.83	4.35 ± 0.86	4.58 ± 0.58	4.60 ± 0.63	4.62 ± 0.57	**4.64 ± 0.56**
Quality (Q): $\bar{x} \pm \sigma$	3.36 ± 1.21	4.29 ± 0.98	4.37 ± 0.85	4.38 ± 0.85	4.61 ± 0.62	**4.68 ± 0.60**	4.65 ± 0.63	4.66 ± 0.58
Content (C): $\bar{x} \pm \sigma$	3.08 ± 1.29	4.18 ± 1.07	4.27 ± 0.93	4.26 ± 1.02	4.44 ± 0.80	**4.57 ± 0.74**	4.45 ± 0.81	4.48 ± 0.77
Style (S): $\bar{x} \pm \sigma$	3.81 ± 1.19	4.38 ± 0.91	4.28 ± 1.04	**4.50 ± 0.84**	4.16 ± 0.95	4.17 ± 1.03	4.35 ± 0.88	4.44 ± 0.84
Mean(L,C,S): $\bar{x} \pm \sigma$	3.53 ± 0.86	4.25 ± 0.73	4.28 ± 0.69	4.37 ± 0.67	4.39 ± 0.53	4.45 ± 0.55	4.47 ± 0.52	**4.52 ± 0.51**
#(output=input): %	0.00%	0.67%	0.00%	1.33%	0.67%	2.00%	0.67%	0.67%
Neutral input								
Language (L): $\bar{x} \pm \sigma$	3.79 ± 1.11	4.12 ± 1.02	4.31 ± 0.97	4.34 ± 0.92	**4.59 ± 0.72**	4.50 ± 0.77	4.45 ± 0.82	4.43 ± 0.81
Quality (Q): $\bar{x} \pm \sigma$	3.52 ± 1.27	4.05 ± 1.14	4.18 ± 1.12	4.34 ± 0.93	**4.49 ± 0.82**	4.43 ± 0.84	4.38 ± 0.91	4.34 ± 0.95
Content (C): $\bar{x} \pm \sigma$	3.04 ± 1.44	3.85 ± 1.31	3.96 ± 1.29	**4.16 ± 1.13**	4.13 ± 1.12	4.07 ± 1.11	4.05 ± 1.15	4.06 ± 1.17
Style (S): $\bar{x} \pm \sigma$	3.20 ± 1.26	3.48 ± 1.42	3.16 ± 1.37	**3.59 ± 1.42**	3.25 ± 1.33	3.32 ± 1.41	3.30 ± 1.37	3.39 ± 1.35
Mean(L,C,S): $\bar{x} \pm \sigma$	3.34 ± 0.92	3.82 ± 0.85	3.81 ± 0.85	**4.03 ± 0.79**	3.99 ± 0.74	3.97 ± 0.71	3.93 ± 0.76	3.96 ± 0.79
#(output=input): %	2.67%	7.33%	10.67%	8.67%	7.33%	10.00%	7.33%	7.33%

Table 6: Human measures statistics of Modern-to-Antique transformation across sizes of target corpora.



(a) Modern Bible input

(b) Neutral input

Figure 3: Human evaluation of models for Bible Modern-to-Antique transformation (legend: average human score is an arithmetic mean of Language, Content and Style).

Modern Bible input

Comparing curves of Bible and Cute style transformations we see significant differences, but also some similarities. Like for N2C models, in both Bible tests the biggest increase in quality is noticeable between the trainings with 1k and 3k style datasets. The modern Bible input data results are better than good for all human criteria for the model using the 3k style dataset. The Style rank reaches a maximum of 4.50 for the model with 7k Bible data volume. However, the most robust model, referring to the human average score, was trained using the biggest 30k style dataset, with the result of 4.52. The M2A transformation models bring a very low unchanged sentence ratio (0-2%), reflecting the training data proportions. We can notice the nonlinear dependency between the target dataset size and Style scores. However, the average of human scores behaves as expected – it grows with the size of style data used for training.

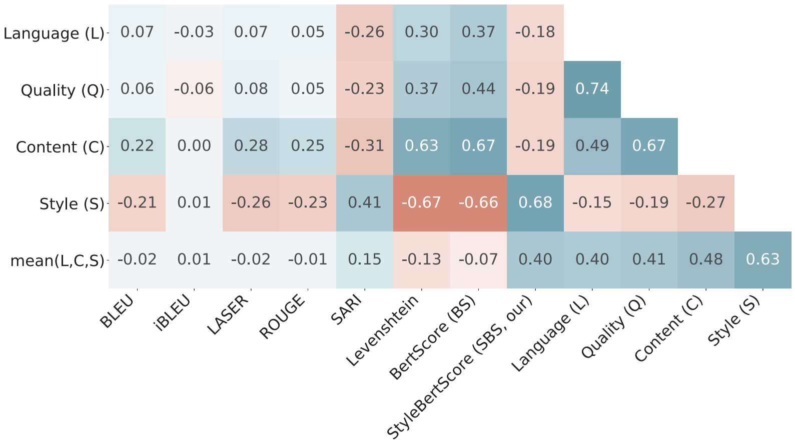
Neutral input

To some extent, we can treat our test with neutral input for the Bible models as another transfer learning experiment. Surprisingly, the results of this study are quite satisfactory. Although in many examples of the same model, human Style scores are even 1 point lower compared to the dedicated modern Bible input, they are still above 3. Like for modern Bible data, the highest Style mark (3.59) for neutral input was reached with the 7k dataset. Considering the human average score, the largest quality growth is between the models trained with 1k and 3k target datasets – 3.34 and 3.82 respectively. Evaluations for all the test samples show that the ratio of unchanged sentences, although higher than for modern Bible inputs, is still low or medium – between 2.67% and 10.67%.

The comparison of human measure curves on the left side (for modern Bible input) of Figure 3 with similar ones on the right side (for neutral input) shows a very interesting observation: Language and Quality marks have more or less similar values on both plots, while Style means are much lower in neutral input tests (from 0.61 up to 1.12 points). In this study Content for neutral input varies from being the same or slightly lower to being lower up to 0.5 points, compared to modern Bible data.

From this analysis, we draw the following conclusions for Bible data and hypotheses for other styles.

1. Multi-task models built using paraphrases and a small or medium amount of style data can easily generate lexicon and logic that are distinctive for style conversion even for unexpected input data.
2. Moreover, style and semantics are more difficult to generate by multi-task models when new input data differ significantly from training input style data. Though, the results are still acceptable.



	BLEU	iBLEU	LASER	ROUGE	SARI	Levenshtein	BertScore (BS)	StyleBertScore (SBS, our)	Language (L)	Quality (Q)	Content (C)	Style (S)
Language (L)	0.07	-0.03	0.07	0.05	-0.26	0.30	0.37	-0.18				
Quality (Q)	0.06	-0.06	0.08	0.05	-0.23	0.37	0.44	-0.19	0.74			
Content (C)	0.22	0.00	0.28	0.25	-0.31	0.63	0.67	-0.19	0.49	0.67		
Style (S)	-0.21	0.01	-0.26	-0.23	0.41	-0.67	-0.66	0.68	-0.15	-0.19	-0.27	
mean(L,C,S)	-0.02	0.01	-0.02	-0.01	0.15	-0.13	-0.07	0.40	0.40	0.41	0.48	0.63

Figure 4: Spearman correlation between various NLP metrics and human judgment for N2C data.

4.4 Automated evaluation

Automated evaluation in text ST is a challenging task. Firstly, the full solution space cannot be clearly predefined (there are many ways of transferring style into a sentence). Secondly, the model should not only add style, but also maintain language correctness and preserve content. Additionally, stylizing text might be manifested in various language components (like syntax or lexis). It is difficult to reflect such multidimensional evaluation in any single synthetic score, especially when those aspects of assessment are poorly correlated or orthogonal.

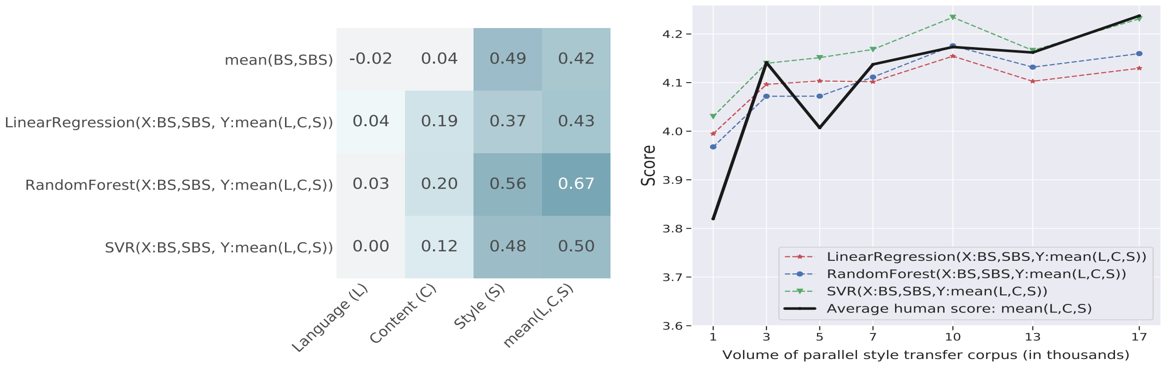
We examined a few well-known automatic metrics in the NLP field. For computations we used Vizseq toolkit (Wang et al., 2019): BLEU, iBLEU, ROUGE-L, LASER; EASSE package (Alva-Manchego et al., 2019): SARI; python-Levenshtein package: Levenshtein ratio; BERTScore (Zhang et al., 2020). These measures correlate to some extent with human ranks in various tasks: MT (ST might be seen as translation within one language), summarization or text simplification (a special case of ST). In Figure 4 we present Spearman correlations between NLP measures and human scores for N2C transformation.

The analysis reveals a limited correlation of the human Style score with the most of checked measures (the biggest positive value is for SARI: 0.41). In our opinion, the tested ranks are not sufficient to estimate all the factors of style transfer. Thus, in this section, we propose an easy method to compose a synthetic style transfer measure that exhibits better correlation with human judgment than the already known measures, even at the cost of its universality. Due to limited space, we focus on N2C transformation only.

We decided to assemble our score from two factors: one capturing the stylistic aspect and the other covering the paraphrase quality. As the measure of paraphrases quality we chose BertScore (Zhang et al., 2020). It is a metric formed on the BERT model's contextual embeddings (Devlin et al., 2019), adequate for identifying semantic similarity between sentences. Moreover, it handles synonyms and spatial lexical dependencies. For those reasons, it has the strongest correlation with the linguistic Content score.

In order to facilitate the evaluation of the stylistic factor, we built a small classifier tool using BERT. We decided to use out-of-the-box torch-transformers[3], and only apply a fine-tuning step on top. From the validation sets we selected 6k-verse subsets (they were not used for the training in style transfer tasks) and built 12k datasets of balanced positive (target) and negative (neutral) styles. The classifier reached an accuracy of 0.77. We called its softmax StyleBertScore. We assumed that the average human score (the arithmetic mean of Style, Content and Language) can be approximated using BertScore and StyleBertScore. In order to combine scores we built a few regression models, from a simple mean, through linear regression, to ones capable of model complex relations: Random Forest and Support Vector Machine Regression (SVR) (Drucker et al., 1997). Figure 5 depicts the estimation results as a linear chart and a heatmap of correlations with human ranks. In our study, Random Forest has the biggest correlation (0.67) and the smallest mean square error (see Table 7).

[3] https://github.com/huggingface/transformers



(a) Spearman correlation between proposed fusion of metrics and human scores (for Neutral-to-Cute evaluation data.)

(b) Regression models estimation of average human rank.

Figure 5: Regression methods for approximating the model score.

Our proposal for creation of automatic measure includes application of two machine learning models and some amount of human evaluation engagement. Although annotation is needed to calibrate the synthetic measure well, we need to make it only once for the whole process. Afterwards, created automatic model of the measure can be used to test many style transfer models without multiplication of human evaluation.

	Mean	Linear Regression	Random Forest	SVR
mean square error (MSE)	11.732	0.401	0.278	0.401

Table 7: Average mean square error (MSE) for each regression model.

We showed that a non-linear combination of the selected metrics (BertScore and StyleBertScore) can approximate the arithmetic mean of human scores and be a reliable method for assembling a style specific automatic measure. The novelty of our metric is that it is partly model dependent, though this results from the nature of the ST task.

5 Conclusions

We discussed the method of text style conversion with a multilingual transformer trained for two tasks: paraphrasing and style changing. We showed a successful approach using a large paraphrase parallel corpus with much less data of neutral-target style pairs. In our numerical experiments models trained with varying sizes of style samples were evaluated with four human scores (and their average), revealing nonlinear dependencies. For both studies, Neutral-to-Cute and Modern-to-Antique, we pointed out essential data volumes in models that brought acceptable and good results. In particular, we indicated the meaningful model performance growth between 1k and 3k sizes of target data. Moreover, the transfer learning ability of our multilingual generator was tested with a satisfying outcome for Neutral-to-Bible transformation (not seen during the model training). Finally, we proposed an easy method to automatically measure style transfer results and to approximate average human score with it. A new measure can be used during model training to estimate style transformation quality, besides checking a typical minimum of cross entropy loss function. In our opinion, this opportunity is worth further research.

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations*, pages 49–54, Hong Kong, China, November. Association for Computational Linguistics.

Amazon. 2020. Samuel L. Jackson - celebrity voice skill for Alexa. *www.amazon.com/gp/product/B07WS3HN5Q*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, Oct - Nov. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

I. A. Bolshakov and Alexander Gelbukh. 2004. Synonymous paraphrasing using wordnet and internet. In *Meziane F., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2004. Lecture Notes in Computer Science, vol 3136*, pages 312–323, 01.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August. Association for Computational Linguistics.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii, October. Association for Computational Linguistics.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the Bible. *Royal Society Open Science*, 5:171920, 10.

Rich Caruana. 1997. Multitask learning. *Mach. Learn.*, 28(1):41–75, July.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2980–2988. Curran Associates, Inc.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *33rd Conference on Neural Information Processing Systems.*

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press.

T. Dryjański, P. Bujnowski, H. Choi, K. Podlaska, K. Michalski, K. Beksa, and P. Kubik. 2018. Affective natural language generation by phrase insertion. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4876–4882, Dec.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI18. AAAI Publications.

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

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *ArXiv*, abs/1707.01161.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA, June. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Content analysis: An introduction to its methodology thousand oaks. *Calif.: Sage*.

T. Kudo and J. Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

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, September. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. 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, June. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018b. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium, October-November. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Kathleen R. McKeown. 1983. Paraphrasing questions using given and new information. *American Journal of Computational Linguistics*, 9(1):1–10.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2018. Waveglow: A flow-based generative network for speech synthesis.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona, Spain, July. Association for Computational Linguistics.

Quora. 2017. Quora question pairs.

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*, pages 129–140, New Orleans, Louisiana, June. Association for Computational Linguistics.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Samsung. 2019. Galaxy's Celebrity Alarm lets you personalize notification alerts with Celebrity Voices. *https://news.samsung.com/global/galaxys-celebrity-alarm-lets-you-personalize-notification-alerts-with-celebrity-voices*.

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Stephen Wan, Mark Dras, Robert Dale, and Cécile Paris. 2005. Towards statistical paraphrase generation: Preliminary evaluations of grammaticality. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 4270–7271. AAAI Press.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. Vizseq: A visual analysis toolkit for text generation tasks. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia, July. 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, June. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea, July. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914. The COLING 2012 Organizing Committee, December.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842, Suntec, Singapore, August. Association for Computational Linguistics.

Appendix: Style Transfer Results

Every judge evaluated the same 300 transformations in four criteria (“Language” – “L”, “Quality” – “Q”, “Content” – “C” and “Style” – “S”) using Likert scale where 1 – very bad, 2 – unacceptable, 3 – flawed, but acceptable, 4 – good, with minor errors, 5 – very good.

“Bible” stands for Modern-to-Antique dataset. “Cute” refers to the Neutral-to-Cute set. For more information refer to the section 4 of the article.

Training Data	Input sentence	Styled Sentence	L	Q	C	S
Bible 30k	*[Neutral]* You’re cool.	Thou art cool.	5	5	5	5
Bible 1k		Thou art cold!	5	5	2	4
Bible 30k	*[Neutral]* You should be a superhero and I’m your sidekick.	Thou shouldest be a superhero, and I am thy sidekick.	5	5	5	5
Bible 1k		Thou should be a fool and I am thy sidekhants.	3	3	1	4
Bible 30k	*[Modern]* My people, what have I done to you? How have I burdened you? Answer me!	My people, what have I done unto thee? how have I straitened thee? hear me.	4	4	3	5
Bible 1k		My people, what have I done unto thee? how hast I buried thee? Aneth me.	3	2	2	5
Bible 30k	*[Modern]* The fool’s talk brings a rod to his back, but the lips of the wise protect them.	The fool’s talk bringeth a rod to his back; but the lips of the wise defendeth them.	5	5	5	5
Bible 1k		The fool’s speak bringeth a rod to his back, but the lips of the wise man keep them.	4	4	4	3
Bible 30k	*[Modern]* Where then is the boasting? It is excluded. By what kind of law? Of works? No, but by a law of faith.	Where then is the glorying? it is excluded. By what manner of law? No, but by the law of faith.	4	3	3	3
Bible 1k		Where is the boasting? It is removed from the law of Of works? but by a law of faith.	2	2	3	2
Cute 17k	Take care of yourself and see a doctor if needed.	Don’t you worry, darling, take care of yourself and see a doctor if needed.	5	5	5	5
Cute 1k		take care of yourself and see a doctor if necessary.	4	5	5	2
Cute 17k	I don’t know much about the president. What’s your opinion?	Hey, sweetie! I don’t know much about the president. What’s your opinion? Cute one, huh?	5	5	5	5
Cute 1k		I don’t know much about the president. What’s your opinion?	5	5	5	1
Cute 17k	The name’s Bond, James Bond. Just kidding, it’s me.	haha, the name’s Bond, James Bond. Just kidding, its me!	4	5	5	5
Cute 1k		No prob darling, just pick the name is Bond, James Bond, just kidding, it’s me.	2	2	2	5
Cute 17k	Am I intelligent?	Seriously, I’m not that intelligent.	5	5	2	3
Cute 1k		Amazing that, sweetie?	2	2	1	5
Cute 17k	Carpe diem.	Carpe totally diem	4	4	5	5
Cute 1k		Good Morning babe‘ u have been kinda quite on da‘ boat!	3	4	2	5