Generic resources are what you need
Lai, Huiyuan; Toral Ruiz, Antonio; Nissim, Malvina

Published in:

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Final author's version (accepted by publisher, after peer review)

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Generic resources are what you need:
Style transfer tasks without task-specific parallel training data

Huiyuan Lai, Antonio Toral, Malvina Nissim
CLCG, University of Groningen / The Netherlands
{h.lai, a.toral.ruiz, m.nissim}@rug.nl

Abstract
Style transfer aims to rewrite a source text in a different target style while preserving its content. We propose a novel approach to this task that leverages generic resources, and without using any task-specific parallel (source–target) data outperforms existing unsupervised approaches on the two most popular style transfer tasks: formality transfer and polarity swap. In practice, we adopt a multi-step procedure which builds on a generic pre-trained sequence-to-sequence model (BART). First, we strengthen the model’s ability to rewrite by further pre-training BART on both an existing collection of generic paraphrases, as well as on synthetic pairs created using a general-purpose lexical resource. Second, through an iterative back-translation approach, we train two models, each in a transfer direction, so that they can provide each other with synthetically generated pairs, dynamically in the training process. Lastly, we let our best resulting model generate static synthetic pairs to be used in a supervised training regime. Besides methodology and state-of-the-art results, a core contribution of this work is a reflection on the nature of the two tasks we address, and how their differences are highlighted by their response to our approach.

1 Introduction
Text style transfer is, broadly put, the task converting a text of one style into another while preserving its content. In its recent tradition within Natural Language Generation (NLG), two tasks and their corresponding datasets have been commonly used (Zhang et al., 2018; Luo et al., 2019; Wu et al., 2019; Yi et al., 2020; Zhou et al., 2020). One dataset was specifically created for formality transfer and contains parallel data (GYAFC (Rao and Tetreault, 2018)), while the other one contains a large amount of non-parallel sentiment labelled texts (YELP (Li et al., 2018)), with parallel pairs for test, and is used for the task of polarity swap. Examples from these datasets are shown in Table 1.

The two tasks are usually conflated in the literature under the general style transfer label and addressed with the same methods, but we find this an oversimplification. Formality transfer implies rewriting a formal sentence into its informal counterpart (or vice versa) while preserving its meaning. Polarity swap, instead, aims to change a positive text into a negative one (or vice versa); and while the general theme must be preserved, the meaning is by definition not maintained (e.g. “I hated that film” → “I loved that film”). In line with previous work, we also address both tasks in a similar way, but this is actually to unveil how their different nature affects modelling and evaluation.

Due to the general scarcity of parallel data, previous works mainly adopted unsupervised approaches, dubbed unpaired methods (Dai et al., 2019) since they do not rely on labelled training pairs. However, it has also been shown that best results, unsurprisingly, can be achieved if parallel training data (such as the formality dataset (Rao and Tetreault, 2018)) is available (Sancheti et al., 2020; Lai et al., 2021). For this reason, substantial work has gone into the creation of artificial training pairs through various methods (see Section 2); approaches using synthetic pairs are thus still considered unsupervised in the style transfer literature, since they do not use manually labelled data.

We explore how parallel data can best be derived and integrated in a general style transfer framework. To do so, we create pairs in a variety of ways and use them in different stages of our framework. A core aspect of our approach is leveraging generic resources to derive training pairs, both natural and synthetic. On the natural front, we use abundant data from a generic rewriting task: paraphrasing. As for synthetic data, we leverage a general-purpose computational lexicon using its antonymy relation to generate polarity pairs.
In practice, we propose a framework that adopts a multi-step procedure which builds upon a general-purpose pre-trained sequence-to-sequence (seq2seq) model. First, we strengthen the model’s ability to rewrite by conducting a second phase of pre-training on natural pairs derived from an existing collection of generic paraphrases, as well as on synthetic pairs created using a general-purpose lexical resource. Second, through an iterative back-translation (Hoang et al., 2018) approach, we train two models, each in a transfer direction, so that they can provide each other with synthetically generated pairs on-the-fly. Lastly, we use our best resulting model to generate static synthetic pairs, which are then used offline as parallel training data.

Contributions Using a large pre-trained seq2seq model (1) we achieve state-of-the-art results for the two most popular style transfer tasks without task-specific parallel data. We show that (2) generic resources can be leveraged to derive parallel data for additional model pre-training, which boosts performance substantially and that (3) an iterative back-translation setting where models in the two transfer directions are trained simultaneously is successful, especially if enriched with a reward strategy. We also offer (4) a theoretical contribution over the nature of the two tasks: while they are usually treated as the same task, our results suggest that they could possibly be treated separately.\(^1\)

2 Related Work

Style transfer is most successful if task-specific parallel data is available, as in the case of formality transfer (Rao and Tetreault, 2018). Like in most NLP tasks, large pre-trained models have been shown to provide an excellent base for fine-tuning in a supervised setting (Chawla and Yang, 2020; Lai et al., 2021).

Since parallel data for fine-tuning such large models for style transfer is scarce, a substantial amount of work has gone into methods for creating artificial sentence pairs so that models can be trained in a supervised regime.

One way to do this is to artificially generate parallel data via back-translation, so that training pairs are created on-the-fly during the training process itself (Zhang et al., 2018; Lample et al., 2019; Prabhumoye et al., 2018; Luo et al., 2019). In these systems, one direction’s outputs and its inputs can be used as pairs to train the model of the opposite transfer direction.

Another common strategy is to use style-word-editing (Li et al., 2018; Xu et al., 2018; Wu et al., 2019; Lee, 2020) to explicitly separate content and style. These approaches first detect relevant words in the source and then do operations like deleting, inserting and combining to create the pair’s target. Back-transferring is generally used to reconstruct the source sentence for training, so that pairs are also made on-the-fly.

Lample et al. (2019) provide evidence that disentangling style and content to learn distinct representations (Shen et al., 2017; Fu et al., 2018; John et al., 2019; Yi et al., 2020) is not necessary. Reconstructing the source, instead, appears beneficial: it is used by Dai et al. (2019) who pre-train a model on style transfer data with the Transformer architecture (Vaswani et al., 2017); and by Zhou et al. (2020), who use an attentional seq2seq model that pre-trains the model to reconstruct the source sentence and re-predict its word-level style relevance.

Luo et al. (2019) pre-train a LSTM-based seq2seq model (Bahdanau et al., 2015) using sentence pairs generated by a template-based baseline. More recently, Li et al. (2020a) proposed a two-stage strategy of search and learning for formality transfer where they perform a simulated annealing search (Liu et al., 2020) to obtain output sentences as pseudo-references, and then fine-tune GPT-2 (Radford et al., 2019) with the resulting pairs.

The methods above create task-specific artificial pairs, some using pre-crafted manual rules or templates. We aim to overcome this by exploiting generic resources. Additionally, it is not evident which strategy works best for creating parallel data, whether offline or on-the-fly, and the simultaneous advantage of both strategies has not been fully explored. Lastly, Chawla and Yang (2020) develop a semi-supervised model based on sequence-to-sequence pre-trained model (BART, Lewis et al. (2020)) using parallel training data and large amounts of non-parallel data, which achieves a significant performance. In previous work, we have also shown that a sequence-to-sequence pre-trained model (BART) outperforms a language model (GPT-2) in content preservation and overall performance when task-specific parallel training data is available (Lai et al., 2021).

Therefore, we use BART as generic base model; we enrich it with iterative back-translation to cre-
Table 1: Samples of each dataset.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>STYLE</th>
<th>SENTENCE-PAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GYAFCTR</td>
<td>Informal</td>
<td>that is just my gut feeling.</td>
</tr>
<tr>
<td></td>
<td>Formal</td>
<td>That is my personal opinion.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no different between ages if the mind is near to eachother.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>There is no difference between ages if the intellect is similar.</td>
</tr>
<tr>
<td>YELP</td>
<td>Negative</td>
<td>this branch is getting worse and worse.</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>this branch is getting better and better.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bad service in these areas and really ruined our visit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>good service in these areas and really made our visit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I guess I’ve always been pretty good with words.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I think narrating has always been my strong suit.</td>
</tr>
<tr>
<td>PARABANK 2</td>
<td>Source</td>
<td>The bank is coming up on your left.</td>
</tr>
<tr>
<td></td>
<td>Target</td>
<td>You have the bank on the left side.</td>
</tr>
</tbody>
</table>

Table 2: Dataset Statistics.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>STYLE</th>
<th>PAIRED</th>
<th>UNPAIRED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Train</td>
<td>Valid</td>
</tr>
<tr>
<td>GYAFCTR [F&amp;R]</td>
<td>Informal</td>
<td>51,967</td>
<td>2,788</td>
</tr>
<tr>
<td></td>
<td>Formal</td>
<td>51,967</td>
<td>2,247</td>
</tr>
<tr>
<td>YELP</td>
<td>Negative</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>PARABANK 2</td>
<td>Source</td>
<td>1,132,289</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Target</td>
<td>1,132,289</td>
<td>N/A</td>
</tr>
</tbody>
</table>

3 Tasks, Datasets, and Evaluation

3.1 Tasks and Datasets

The task of style transfer is generally defined as the conversion of a text written in a given style to approximately the same text in a different style: style should be changed while preserving the original “content”. We focus on the two most popular tasks, namely formality transfer and polarity swap, and use the two standard available datasets. Example pairs are shown in Table 1; statistics are in Table 2.

Formality Transfer Dataset Grammarly’s Yahoo Answers Formality Corpus (GYAFCTR) (Rao and Tetreault, 2018) is a dataset containing aligned formal and informal sentences from two domains: Entertainment & Music (E&M) and Family & Relationships (F&R). Parallel pairs are provided for training, validation, and test, with four human references for every test sentence. In the experiments we report in this paper we use data from the F&R domain, which is the one more commonly used.

Polarity Swap Dataset YELP is a dataset of business reviews on Yelp (with scores 1–5) processed by Li et al. (2018). Samples with a score greater than 3 are considered as positive otherwise they are negative. The dataset comes in the form of large amounts of non-parallel data for training and development, while parallel pairs are provided for evaluation. For each test sentence, Li et al. (2018) provide one human reference; three additional human references are released by Luo et al. (2019).

Although these two tasks have been conflated in previous work as “style transfer”, they are not exactly the same, which we hypothesise affects both their modelling and evaluation. More specifically, in polarity swap the actual content is not exactly preserved (the message is actually the opposite), rather it’s the general “theme/topic” that needs to be preserved. In formality transfer, instead, the “translation” happens really more at style level, and content needs to stay the same. This is evident if we look at examples in Table 1 (top two blocks). In YELP, we can see that the theme-related words are expected to stay while changing the polarity words. Therefore, although the two sentences refer to the same event/concept, they convey opposite meanings. On the contrary, in formality transfer, an informal text should be changed into a formal one, but the overall meaning should be preserved. In this sense, formality transfer can be seen much more as rewriting than polarity swap and can be conceived akin to the more general task of paraphrasing.

Leveraging this observation, we explore if paraphrase pairs can be used to make the model learn the basic task of “rewriting” in a first stage. The advantage of using rewriting is the large amount of parallel data available. Specifically, we use PARABANK 2, a large-scale, diverse, collection of paraphrases (Hu et al., 2019). Given the different nature of the two tasks, we expect this strategy to help more formality transfer than polarity swap, since the latter is much less of a rewriting task than the former. In spite of the differences highlighted above, we approach both tasks within the same framework for two reasons: (i) to compare to previous works, which have treated the tasks as manifestations of the same “style transfer” task; but also (ii) to observe if and how the tasks respond differently to modelling and evaluation metrics.
3.2 Task Evaluation

The performance of text style transfer is commonly assessed on style strength and content preservation. For style strength, using a pre-trained style classifier is the most popular automatic evaluation strategy. For content preservation, \( n \)-gram-based matching metrics such as BLEU (Papineni et al., 2002) are most commonly used. However, these metrics usually fail to recognize information beyond the lexical level. Since word embeddings (Mikolov et al., 2013; Pennington et al., 2014) have become the prime alternative to \( n \)-gram-based matching to capture similarity, embeddings-based metrics have also been developed (Fu et al., 2018). However, embedding-based metrics like cosine similarity still work at the token-level, and might fail to capture the overall semantics of a sentence.

To overcome such limitations, recent work has developed learnable metrics, which attempt to directly optimize correlation with human judgments. These metrics, with the prime examples of BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020), have recently shown promising results in machine translation evaluation. To the best of our knowledge, only our previous work used BLEURT in the evaluation of formality style transfer models (Lai et al., 2021); we are now proposing to use it also for the evaluation of polarity swap. Formally, the objective is to minimize the following negative log likelihood:

\[
L(\phi) = -\sum \log(p(y_i|y_{1:i-1}, x; \phi))
\]

where \( \phi \) are the parameters of BART.

Our framework can be conceived as a pipeline, visualized in Figure 1. At the core of the framework are two BART models (model A and model B), one for each transfer direction. Since the main challenge in unpaired style transfer is that we cannot directly employ supervision (i.e. task-specific parallel training pairs), we explore and evaluate different ways of creating and using sentence pairs at different stages of the pipeline.

First, we strengthen the model’s ability to rewrite by conducting a second phase of pre-training on natural pairs derived from an existing collection of generic paraphrases, as well as on synthetic pairs created using a general-purpose lexical resource (Step 1, Section 4.1). Second, we use iterative back-translation with several reward strategies to train the two models in both transfer directions simultaneously; sentence pairs are created on the fly (Step 2, Section 4.2). Third, we create high-quality synthetic pairs using our best systems from the previous step, to create a static resource of parallel data that can be used to train new transfer models (Step 3, Section 4.3).

4 Approach

We propose a framework that adopts a multi-step procedure on top of the large pre-trained seq2seq model BART (Lewis et al., 2020).

Given a source sentence \( x = \{x_1, \cdots, x_n\} \) of length \( n \) with style \( s_1 \), the goal of text style transfer is to generate a sentence \( y \) with style \( s_2 \), preserving the source sentence’s meaning in formality transfer or the source sentence’s theme in polarity swap.\(^{2}\) Formally, the objective is to minimize the following negative log likelihood:

\[
L(\phi) = -\sum \log(p(y_i|y_{1:i-1}, x; \phi))
\]

where \( \phi \) are the parameters of BART.

Our framework can be conceived as a pipeline, visualized in Figure 1. At the core of the framework are two BART models (model A and model B), one for each transfer direction. Since the main challenge in unpaired style transfer is that we cannot directly employ supervision (i.e. task-specific parallel training pairs), we explore and evaluate different ways of creating and using sentence pairs at different stages of the pipeline.

First, we strengthen the model’s ability to rewrite by conducting a second phase of pre-training on natural pairs derived from an existing collection of generic paraphrases, as well as on synthetic pairs created using a general-purpose lexical resource (Step 1, Section 4.1). Second, we use iterative back-translation with several reward strategies to train the two models in both transfer directions simultaneously; sentence pairs are created on the fly (Step 2, Section 4.2). Third, we create high-quality synthetic pairs using our best systems from the previous step, to create a static resource of parallel data that can be used to train new transfer models (Step 3, Section 4.3).

4.1 Further Pre-training: Learning to Rewrite

As hinted at in Section 3, style transfer can be seen as a specific way of **paraphrasing**. On the basis

\(^{2}\)In what follows, we use the term “content” in a more general way to refer to both cases.
of this intuition, we hypothesise that generic paraphrase data, which already exists in much larger amounts than task-specific style transfer data, can be useful for text style transfer in terms of teaching the models the more generic task of “rewriting”. For polarity swap, which is less of a rewriting task than formality transfer, as the meaning is reversed rather than preserved, we also create synthetic pairs using a general-purpose lexical resource.

Using the natural and the synthetic pairs we conduct a second phase of pre-training. We expect this strategy to help specifically with content preservation, which is known to be the most difficult part of style transfer, especially in an unsupervised setting (Sancheti et al., 2020; Lai et al., 2021).

Generic Training Pairs We use data from PARABANK 2 to make the model learn the basic task of “rewriting”. We use this dataset in its entirety or filtered (models M1.1 and M1.2 in Table 3). In the first case, the whole of the paraphrase pairs from PARABANK 2 are used to further pre-train the model. In the second case, we follow the rationale that not all pairs are equally relevant for our tasks, and selecting task-specific ones could be beneficial. For instance, while both PARABANK 2 pairs in Table 1 are good examples of rewriting, the one on the right is more meaningful in terms of formality transfer. Therefore, we train two binary style classifiers, one for formality and one for polarity, using TextCNN (Kim, 2014) on the training sets of GYAFC and YELP. These classifiers are then used to automatically select more strongly style-opposed pairs. The resulting filtered paraphrase subset $D_p$ is such a set of pairs:

$$D_p = \{(x, y)|p(s_1|x) + p(s_2|y))/2 > \sigma\} \quad (2)$$

where $p(s_i|x)$ is the probability of a sentence being a style $s_i$, predicted by the style classifier, and $\sigma$ is the threshold for data selection$^3$; $x$ and $y$ constitute the sentence pair.

Synthetic Pairs for Polarity Swap Due to the nature of polarity swap, we expect that even filtered paraphrases might not benefit polarity swap as much as formality. We therefore add another strategy to enhance polarity swap rewriting and create pairs for further pre-training exploiting a general-purpose lexical resource (model M1.3 in Table 3). Specifically, we use SentiWordNet (Baccianella et al., 2010) to obtain words’ sentiment scores to detect the polarity of each word in the sentence. To maximise the quality of synthetic pairs, we select sentences that contain one polarity word only, and swap that one with its WordNet antonym (Miller, 1995). The new synthetic sentence is regarded as the target sentence corresponding to the original sentence.

The generic/filtered/synthetic pairs are used for a second phase of seq2seq pre-training for BART. Examples of these pairs are in Appendix A.5.

4.2 Iterative Back-translation and Rewards: Pairs on-the-fly
After further pre-training BART, we use iterative back-translation to train two models, each in a transfer direction, so that they can provide one another with synthetically generated pairs on-the-fly. We obtain pseudo-parallel data via back-transfer: the outputs of one direction are used to provide the supervision to train the model of the opposite direction (Figure 2). To explicitly guide the model to preserve the content and to apply the target style, we add content and style rewards in a reinforcement learning fashion (models M2.* in Table 3).

Rewarding Style Strength To provide an explicit signal to teach the model to change the sentence’s style, a style classifier (SC) based reward is used to push the model to change the sentence into the target style. For this SC reward, which evaluates

---

$^3\sigma = 0.85$ in our experiments.
how well the transferred sentence $y'$ matches the target style, we reuse the style classifier trained for selecting paraphrase data (Section 4.1). The SC’s confidence in each transfer direction is

$$p(s_i|y') = \text{softmax}_i(\text{TextCNN}(y', \theta))$$

(3)

where $i \in \{1, 2\}$ and $\theta$ are the parameters of the style classifier, fixed during training transfer models. Formally, the reward is

$$R_{sc} = \lambda_{sc}[p(s_2|y') - p(s_1|y')]$$

(4)

where $s_1$ and $s_2$ are source style and target style, respectively. $y'$ is the generated target sentence sampled from the distribution of model outputs at each decoding time step.

We apply the SC reward in two ways: in the supervised training process using pseudo-parallel data (SCO); and in the process of generating pseudo-parallel data itself (SC1). For the latter, we generate text in the target style by sampling the distribution of model outputs, while at the same time use the SC reward to feedback its corresponding style signal to the model.

**Rewarding Content Preservation** Following Sancheti et al. (2020), we use a BLEU-based reward, formulated as follows:

$$R_{\text{bleu}} = \lambda_{\text{bleu}}[\text{BLEU}(y_{s_i}^\prime, x) - \text{BLEU}(y_{s_1}^\prime, x)]$$

(5)

where $y_{s_i}^\prime$ is the generated sentence in target style $s_i$ sampled from the distribution of model outputs at each time step in decoding, and $y_{s_1}^\prime$ is obtained by greedily maximizing the distribution.

Since new-generation metrics show promising results in evaluation (Section 3), we use BLEURT also as an alternative metric to BLEU in the reward strategy, expecting it might be better at measuring semantics at the sentence level. Formally, we formulate the BLEURT-based reward as

$$R_{\text{bleurt}} = \lambda_{\text{bleurt}}[\text{BLEURT}(y_{s_i}^\prime, x)]$$

(6)

where $y_{s_i}^\prime$ is the generated sentence in target style $s_i$ sampled from the distribution of model outputs.

**Gradients and Objectives** We use the policy gradient algorithm (Williams, 1992) to maximize the expected reward of the generated sentence $y^\prime$, whose gradient with respect to the parameters $\phi$ of the neural network model is estimated by sampling:

$$\nabla_\phi J(\phi) = E[R \cdot \nabla_\phi \log(P(y^\prime|x; \phi))]$$

(7)

where $\nabla_\phi J(\cdot)$ is the gradient of objective function $J(\cdot)$ with respect to model parameters $\phi$, $E(\cdot)$ is the expectation, $R$ is the reward of the sequence $y^\prime$ that is sampled from the distribution of model outputs at each decoding time step. The overall objectives are the combination of the base model’s loss (Eq. 1) and the policy gradient of rewards (Eq. 7) which are used to train our framework end-to-end.

### 4.3 Final Training: High-quality Pairs

As a final step, we let our best models generate pairs to create a static resource of parallel data. We feed the system source sentences randomly picked from the training sets and generate the corresponding sentences in the target style. We then select high-quality pairs using BLEURT and our style classifier. The resulting dataset $D_h$ is a set of pairs:

$$D_h = \{(x, y')|\text{BLEURT}(x, y') > \sigma_c \text{ and } (p(s_1|x) + p(s_2|y'))/2 > \sigma_s\}$$

(8)

where $x$ and $y'$ are the source sentence and generated sentence, respectively. $p(s_1|x)$ is the probability of a sentence being of style $s_1$ as predicted by the style classifier, and $\sigma_s$ is the threshold for data selection regarding content and style.\footnote{\(\sigma_c = 0.15\) and \(\sigma_s = 0.9\) in our experiments.}

Finally, these pairs are used to fine-tune the original BART with all reward strategies, so as to train new transfer models in a supervised way (model M3.1 in Table 3).

### 5 Experiments

All experiments are implemented atop Huggingface Transformers (Wolf et al., 2020), taking the BART base model (139M parameters) for our experiments. We train our framework using the Adam optimiser (Kingma and Ba, 2015) with the initial learning rate $1e^{-5}$. The batch size is set to 32. The final values $\lambda$ for style and content rewards are both set to 1 based on validation results. Both WordNet and SentiWordNet are used from NLTK.\footnote{https://www.nltk.org/}

#### 5.1 Evaluation Metrics

To assess style and content we use common metrics for this task. For content preservation we add two learnable metrics, which we hope will be adopted from now on, to glean better insights into the systems’ behaviour in the two tasks (Section 3.2).

We measure style strength automatically by evaluating the target style accuracy of transferred
sentences. We use the style classifiers trained for selecting paraphrase data (Section 4.1). The classifiers have an accuracy of 92.6% and 98.1% on the test sets of F&R and YELP, respectively.

To assess content preservation, we follow previous work and calculate BLEU\(^6\) between the generated sentence and the human reference(s). Additionally, we compute BLEURT and COMET\(^7\). As the human references for YELP are released from different researchers and appear to differ quite a lot in nature (see Appendix A.6 for examples), we provide two evaluation results: one using the first human reference only (Table 3), and the other using all four (Appendix A.2).

As overall score, for a direct comparison to previous work (Luo et al., 2019; Zhou et al., 2020; Lai et al., 2021) we compute the harmonic mean (HM) of style accuracy and BLEU.

### 5.2 Results

Table 3 reports results for each step.\(^8\)

Results of Step 1 show that using paraphrase data benefits more formality transfer than polarity swap, confirming the latter is much less of a rewriting task than the former. Filtering paraphrases to a subset closer to the task (M1.2) substantially helps formality and yields some improvement in polarity. WordNet-derived synthetic pairs (M1.3) are definitely a better strategy for polarity.\(^9\)

The first block of Step 2 confirms that further

---

\(^6\)We use multi-bleu.perl with default settings.  
\(^7\)COMET is designed to also take in input sentences into account, but our evaluations including them yielded lower correlations with human judgements. This might be because in COMET training input and output are different languages.  
\(^8\)Results for more models per step are in Appendix A.1.  
\(^9\)The WordNet-based strategy could in principle be used on its own to solve the polarity swap task with no learning involved, but results prove it insufficient: BLEURT: -0.475; COMET: -0.221; BLEU: 0.296; ACC: 0.206; HM: 0.243.
pre-training significantly improves performance on formality transfer (compare M2.2 with M2.1). This results in the best model for formality transfer. For polarity, instead, we see improvement from further pre-training only when using WordNet-based synthetic pairs (compare M2.2 with M2.3). Overall, in Step 2 we see that combining SC rewards and content-related rewards results in the best balance regarding content preservation and style strength.

In Step 3, we see that the model trained with high-quality synthetic pairs (M3.1) achieves the best overall performance on polarity swap. For comparison, we use the subset of paraphrase data as training pairs in place of the generated pairs, and see that performance is lower (M3.2).

Table 4 shows example outputs of each step and their evaluation results. It is interesting to see the impact of paraphrase-based pre-training: for formality, in M1.1 and M1.2, the phrase “if you want to do this” is used in place of “if you’re set on that”. This rewriting ability can also be observed on the polarity swap (“on top of there jobs” → “they’re doing their jobs well”); note also that using paraphrases seems to prompt better writing: “there” → “their”, M1.1/M3.2, though this is not consistent throughout the models). For formality, the quality of the output gradually improves in Step 2, with M2.2 achieving the best performance on BLEU and style confidence (M2.2); the model trained with high-quality synthetic pairs (M3.1) has the highest BLEURT and COMET. In M3.2, trained on paraphrase pairs, we find nice variability again (“if you’re on board”). For polarity, M1.3 (using WordNet-based synthetic pairs), swaps a polarity word with its antonym (“friendly” → “unfriendly”). In Step 2, the models are indeed changing the polarity of the sentence; finally, the model trained with high-quality pairs (M3.1) nicely changes “and” into “or” to get the right semantics (though it loses the correct form “their”) and is scored best. Further exploration of combining generic and task-specific rewriting appears very promising for these tasks.

As an additional curiosity-driven qualitative assessment of the behaviour of our models, we probed the polarity swap models with neutral sentences. As a first example, we use “the earth revolves around the sun.” as the source sentence, and observe that the models in both transfer directions generate the same sentences as the input. With as input the neutral sentence “there is a grocery store near my house.”, the model which transforms negative sentences into positive ones generates “there is a great grocery store near my house.” while into the other direction it generates “there is no grocery store near my house.”. It is worth mentioning that all the training data comes from business reviews on YELP, and the first example is clearly outside that domain. For the second example, closer to the domain of YELP, the transformation proposed by the model is rather reasonable in terms of obtaining a positive (“great grocery store”) or negative (“no grocery store”) output. It is left to future research to investigate what it should mean to transform a neutral sentence into a positive/negative one, and how such a test can help to better understand the models’ behaviour and the task itself.

**Comparison to other systems** To put our results in perspective, we compare our best system (M2.2 for formality and M3.1 for polarity in Table 3) against the most recent and best performing unpaired systems. For formality: UnsuperMT (Zhang et al., 2018); DualRL (Luo et al., 2019); StyIns (Yi et al., 2020); Zhou’s (Zhou et al., 2020); TGLS (Li et al., 2020a). For polarity: Style-Transformer (Dai et al., 2018); DGST (Li et al., 2020b); Zhou’s (Zhou et al., 2020); TGLS (Li et al., 2020a). As an additional curiosity-driven qualitative assessment of the behaviour of our models, we probed the polarity swap models with neutral sentences. As a first example, we use “the earth revolves around the sun.” as the source sentence, and observe that the models in both transfer directions generate the same sentences as the input. With as input the neutral sentence “there is a grocery store near my house.”, the model which transforms negative sentences into positive ones generates “there is a great grocery store near my house.” while into the other direction it generates “there is no grocery store near my house.”. It is worth mentioning that all the training data comes from business reviews on YELP, and the first example is clearly outside that domain. For the second example, closer to the domain of YELP, the transformation proposed by the model is rather reasonable in terms of obtaining a positive (“great grocery store”) or negative (“no grocery store”) output. It is left to future research to investigate what it should mean to transform a neutral sentence into a positive/negative one, and how such a test can help to better understand the models’ behaviour and the task itself.

As an additional curiosity-driven qualitative assessment of the behaviour of our models, we probed the polarity swap models with neutral sentences. As a first example, we use “the earth revolves around the sun.” as the source sentence, and observe that the models in both transfer directions generate the same sentences as the input. With as input the neutral sentence “there is a grocery store near my house.”, the model which transforms negative sentences into positive ones generates “there is a great grocery store near my house.” while into the other direction it generates “there is no grocery store near my house.”. It is worth mentioning that all the training data comes from business reviews on YELP, and the first example is clearly outside that domain. For the second example, closer to the domain of YELP, the transformation proposed by the model is rather reasonable in terms of obtaining a positive (“great grocery store”) or negative (“no grocery store”) output. It is left to future research to investigate what it should mean to transform a neutral sentence into a positive/negative one, and how such a test can help to better understand the models’ behaviour and the task itself.

### Table 5: Comparison with other systems

<table>
<thead>
<tr>
<th></th>
<th>BLEURT</th>
<th>COMET</th>
<th>BLEU</th>
<th>ACC</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Copy</td>
<td>-0.114</td>
<td>0.272</td>
<td>0.474</td>
<td>0.120</td>
<td>0.192</td>
</tr>
<tr>
<td>UnsuperMT (Zhang et al., 2018)</td>
<td>-0.665</td>
<td>-0.446</td>
<td>0.327</td>
<td>0.670</td>
<td>0.439</td>
</tr>
<tr>
<td>DualRL (Luo et al., 2019)</td>
<td>-0.589</td>
<td>-0.451</td>
<td>0.404</td>
<td>0.654</td>
<td>0.489</td>
</tr>
<tr>
<td>StyIns (Yi et al., 2020)</td>
<td>-0.395</td>
<td>-0.112</td>
<td>0.458</td>
<td>0.761</td>
<td>0.573</td>
</tr>
<tr>
<td>Zhou’s (Zhou et al., 2020)</td>
<td>-0.454</td>
<td>-0.203</td>
<td>0.447</td>
<td>0.799</td>
<td>0.573</td>
</tr>
<tr>
<td>TGLS (Li et al., 2020a); 0 → 1</td>
<td>-0.383</td>
<td>-0.139</td>
<td>0.312</td>
<td>0.019</td>
<td>0.036</td>
</tr>
<tr>
<td>Ours (M2.2; lowercase)</td>
<td>0.009</td>
<td>0.328</td>
<td>0.563</td>
<td>0.866</td>
<td>0.682</td>
</tr>
<tr>
<td>Ours (M2.2; lowercase)</td>
<td>-</td>
<td>-</td>
<td>0.741</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

References and more examples are in Appendix A.3.
Table 6: Pearson correlation between evaluation metrics for content preservation over $N$ systems.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>N</th>
<th>BLEURT</th>
<th>COMET</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>COMET</td>
<td>BLEU</td>
<td></td>
</tr>
<tr>
<td>Formality Transfer</td>
<td>21</td>
<td>0.980</td>
<td>0.775</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(p&lt;0.01)</td>
<td>(p&lt;0.01)</td>
<td>(p&lt;0.01)</td>
</tr>
<tr>
<td>Polarity Swap</td>
<td>21</td>
<td>0.968</td>
<td>0.671</td>
<td>0.479</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(p&lt;0.01)</td>
<td>(p&lt;0.01)</td>
<td>(p=0.03)</td>
</tr>
</tbody>
</table>

As visible in Table 5, our models achieve the best overall performance on both tasks. For formality transfer, this is true in all evaluation metrics. For polarity swap, StyIns has the highest style accuracy, while our model is better on all other metrics.

5.3 Reflections on Tasks and Evaluation

The strategy of making the model learn the basic task of “rewriting” in a first stage clearly benefits more formality transfer than polarity swap. This is not surprising, since the latter is not simply “rewriting a sentence in a different style”; rather, the task involves changing the meaning of a sentence to obtain its opposite polarity, and thus, broadly put, its meaning. The fact that polarity swap cannot be regarded as a “style change” task is also evident from evaluation. Rather than only using BLEU, we suggested to also use BLEURT and COMET, and this provides us with additional evidence. Specifically, from Table 6 we observe that BLEU has a high correlation with BLEURT/COMET for formality transfer but not for polarity swap.

To glean further insights into this difference, we leverage human judgments released by Li et al. (2018) for YELP and see how they correlate with the used metrics. We calculate system-level Pearson correlation between the automatic evaluations and human judgment.

Results show that while COMET and BLEURT highly correlate with human judgments, BLEU does so to a lesser extent, suggesting this might be a less strong measure to assess the goodness of polarity swap.14

Intuitively, if a system does not change the polarity it may still have a high n-gram overlap (high BLEU) while new-generation metrics do not have this problem. For formality this limitation of BLEU is not much of an issue, since meaning is not altered. Nevertheless, we suggest that the evaluation of style transfer and related tasks should use learned metrics whenever possible.

6 Conclusions

We proposed an unpaired approach that adopts a multi-step procedure based on the general-purpose pre-trained seq2seq model BART.

Achieving state-of-the-art results on the two most popular “style transfer” tasks, we have shown the benefit of further pre-training using data derived from generic resources as well as the advantage of back-translation, paired with rewards, especially towards content preservation. We have also seen how leveraging paraphrases can enhance both variability and naturalness in the generated text.

Through experimental settings as well as the introduction of BLEURT and COMET as metrics, we have also highlighted how the two tasks we addressed differ, and should probably not be conflated into a single “style transfer” label. Indeed, we show that they benefit from partially different modelling, and react differently to evaluation metrics, both key aspects to improve future modelling of these tasks.

Acknowledgments

This work was partly funded by the China Scholarship Council (CSC). The anonymous EMNLP reviewers provided us with useful comments which contributed to improving this paper and its presentation, so we’re grateful to them. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Ethics Statement

All work that automatically generates and/or alters natural text could unfortunately be used maliciously. While we cannot fully prevent such uses once our models are made public, we do hope that writing about risks explicitly and also raising awareness of this possibility in the general public are ways to contain the effects of potential harmful uses. We are open to any discussion and suggestions to minimise such risks.

---

12See Section 2 for details on these models.
13A sample comparison of outputs is in Appendix A.4.
14Pearson’s $r = .922$ for BLEURT, $r = .941$ for COMET, and $r = .901$ for BLEU. All $p < .001$, $N = 7$. 
References


A Appendices:

This appendices include: 1) detailed results for the different steps of the pipeline (A.1); 2) detailed evaluation results of using four human references on YELP (A.2); 3) example outputs for the different steps of the pipeline (A.3); 4) example outputs for existing systems we compare to, and our best models (A.4); 5) sample examples for further pre-training (A.5); 6) sample examples of human reference on YELP (A.6).

A.1 Detailed results for the different steps of the pipeline

<table>
<thead>
<tr>
<th>Model</th>
<th>DATASET</th>
<th>GYAFC (FORMALITY TRANSFER)</th>
<th>YELP (POLARITY SWAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEURT</td>
<td>COMET</td>
<td>BLEU</td>
</tr>
<tr>
<td>Original BART</td>
<td>-0.116</td>
<td>0.242</td>
<td>0.414</td>
</tr>
<tr>
<td>Further pre-trained BART using whole dataset</td>
<td>0.012</td>
<td>0.209</td>
<td>0.420</td>
</tr>
<tr>
<td>Further pre-trained BART using subset</td>
<td>0.011</td>
<td>0.225</td>
<td>0.441</td>
</tr>
<tr>
<td>Further pre-trained BART using synthetic data</td>
<td>0.011</td>
<td>0.225</td>
<td>0.441</td>
</tr>
</tbody>
</table>

Table A.1: Detailed results for the different steps of the pipeline. Note that (i) SC0 represents the SC reward is used in the supervised training process using pseudo-parallel data, and SC1 is in the process of generating pseudo-parallel data.

A.2 Detailed evaluation results of using four human references on YELP

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEURT</th>
<th>COMET</th>
<th>BLEU</th>
<th>ACC</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Copy</td>
<td>-0.337</td>
<td>-0.033</td>
<td>0.640</td>
<td>0.019</td>
<td>0.037</td>
</tr>
<tr>
<td>DualRL (Luo et al., 2019)</td>
<td>-0.281</td>
<td>-0.080</td>
<td>0.550</td>
<td>0.894</td>
<td>0.681</td>
</tr>
<tr>
<td>Style-Transformer (Dai et al., 2019)</td>
<td>-0.390</td>
<td>-0.158</td>
<td>0.553</td>
<td>0.857</td>
<td>0.672</td>
</tr>
<tr>
<td>DGST (Li et al., 2020b)</td>
<td>-0.337</td>
<td>-0.131</td>
<td>0.520</td>
<td>0.781</td>
<td>0.624</td>
</tr>
<tr>
<td>StyIns (Yi et al., 2020)</td>
<td>-0.487</td>
<td>-0.280</td>
<td>0.489</td>
<td>0.924</td>
<td>0.640</td>
</tr>
<tr>
<td>Zhou’s (Zhou et al., 2020)</td>
<td>-0.162</td>
<td>0.090</td>
<td>0.608</td>
<td>0.865</td>
<td>0.714</td>
</tr>
<tr>
<td>Ours</td>
<td>-0.053</td>
<td>0.192</td>
<td>0.610</td>
<td>0.887</td>
<td>0.723</td>
</tr>
</tbody>
</table>

Table A.2: Automatic evaluation results using four human references on YELP.
So if you're set on that, that's the way to go!!

I'm sorry, but to apologize without making a mistake is humiliation to myself, and don't ever get weak!

If you're on board, that's the way to go.

(Saying sorry to him without committing a mistake is humiliation to yourself... and don't ever get weak!!)

So, if you are set on that, then that is the way to go.

Avoid being weak by not saying sorry for something you did not do.

so if you're set on that, that's the way to go!!

I hardly ever see him in school, but usually, I see him at my brother's basketball games.

I would not see him in school either because I see to profess my brothers basketball games.

so great treatment and great medication to help me deal with my condition.

so, no treatment and no medication to help me deal with my condition.

so, great treatment and great help to deal with my condition.

so, great treatment and great medication to help me deal with my condition.

so, great treatment and great help to deal with my condition.

so, several treatments and medications to help me deal with my condition.
A.5 Sample examples for further pre-training

<table>
<thead>
<tr>
<th>RESOURCE</th>
<th>TASK</th>
<th>INFORMAL/NEGATIVE</th>
<th>FORMAL/POSITIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase Formality</td>
<td>Now I... I like the interactive side of this job. Y’all got five minutes to finish your smoke. Yeah, well, there’s a bridge right here. If they go… they leave the source of power behind. Ain’t many of us can face it out there sober all the time.</td>
<td>I like the interactivity on our work. All of you have five minutes to finish your long smoke. Here’s one bridge, ten kilometers from there. If they leave, they’ll leave their source of strength. Not too many of us could face this outside as sober.</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td>he makes me feel wired. there’s a room on the empty floor – it’s only part of work, you know – routine clearance. it was the only thing i liked to buy here. wherever it went, it was followed by an admiring crowd of small lassans.</td>
<td>I... it gives me a funny feeling. there’s plenty of free space on the next floor. great. yeah, it’s just part of the job, you know... a routine imprint. that is the one thing i actually enjoy buying at this store. wherever he moved, he was followed by an astonished mob of little lassans.</td>
<td></td>
</tr>
<tr>
<td>WordNet Polarity</td>
<td>some of the worst pizza i’ve ever had. also the inside is dirty as heck. wherever it went, it was followed by an admiring crowd of small lassans.</td>
<td>some of the best pizza i’ve ever had. also the inside is clean as heck. absolutely the worst care in all my experience with vets!</td>
<td></td>
</tr>
</tbody>
</table>

Table A.5: Sample examples for further pre-training. * indicates that the sentences are synthetic.

A.6 Sample examples of human reference on YELP

<table>
<thead>
<tr>
<th>NEGATIVE → POSITIVE</th>
<th>POSITIVE → NEGATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>ever since joes has changed hands it’s just gotten worse and worse.</td>
</tr>
<tr>
<td>Reference 1</td>
<td>ever since joes has changed hands it’s gotten better and better.</td>
</tr>
<tr>
<td>Reference 2</td>
<td>since joe changed hands, it has become a better place.</td>
</tr>
<tr>
<td>Reference 3</td>
<td>there is definitly not enough room in that part of the venue.</td>
</tr>
<tr>
<td>Reference 4</td>
<td>so basically tasted watered down.</td>
</tr>
<tr>
<td>Source</td>
<td>it didn’t taste watered down at all.</td>
</tr>
<tr>
<td>Reference 1</td>
<td>so it’s fine because it is not watered down.</td>
</tr>
<tr>
<td>Reference 2</td>
<td>so basically did not taste watered down.</td>
</tr>
<tr>
<td>Reference 3</td>
<td>she said she’d be back and disappeared for a few minutes.</td>
</tr>
<tr>
<td>Reference 4</td>
<td>she said she’d be back, and didn’t disappear at all.</td>
</tr>
<tr>
<td>Source</td>
<td>she said she’d be back and enjoy herself</td>
</tr>
<tr>
<td>Reference 1</td>
<td>she said she’d be back and will not disappeared</td>
</tr>
<tr>
<td>Reference 2</td>
<td>she said she’d be back and have a good time</td>
</tr>
<tr>
<td>Reference 3</td>
<td>i can’t believe how inconsiderate this pharmacy is. this pharmacy is really considerate.</td>
</tr>
<tr>
<td>Reference 4</td>
<td>i can not imagine how considerate this pharmacy is. the pharmacy was so considerate of me</td>
</tr>
<tr>
<td>Source</td>
<td>i can not believe how considerate this pharmacy is</td>
</tr>
<tr>
<td>Reference 1</td>
<td>i can not imagine how considerate this pharmacy is. the pharmacy was so considerate of me</td>
</tr>
<tr>
<td>Reference 2</td>
<td>i can’t believe how inconsiderate this pharmacy is.</td>
</tr>
<tr>
<td>Reference 3</td>
<td>the pharmacy was so considerate of me</td>
</tr>
<tr>
<td>Reference 4</td>
<td>i can not believe how considerate this pharmacy is</td>
</tr>
</tbody>
</table>

Table A.6: Sample examples of human reference on YELP. The first human reference is provided by Li et al. (2018), and the 3 additional references are released by Luo et al. (2019).