Exploiting Social Media Content for Self-Supervised Style Transfer

Dana Ruiter[1]**, Thomas Kleinbauer**[1]**, Cristina España-Bonet**[2,3]**,**
Josef van Genabith[2,3]**, Dietrich Klakow**[1]

[1]Spoken Language Systems Group, Saarland University, Germany
[2]Saarland Informatics Campus, Saarland University, Germany
[3]DFKI GmbH, Germany
`druiter@lsv.uni-saarland.de`

Abstract

Recent research on style transfer takes inspiration from unsupervised neural machine translation (UNMT), learning from large amounts of non-parallel data by exploiting cycle consistency loss, back-translation, and denoising autoencoders. By contrast, the use of self-supervised NMT (SSNMT), which leverages (near) parallel instances hidden in non-parallel data more efficiently than UNMT, has not yet been explored for style transfer. In this paper we present a novel Self-Supervised Style Transfer (3ST) model, which augments SSNMT with UNMT methods in order to identify and efficiently exploit supervisory signals in non-parallel social media posts. We compare 3ST with state-of-the-art (SOTA) style transfer models across civil rephrasing, formality and polarity tasks. We show that 3ST is able to balance the three major objectives (fluency, content preservation, attribute transfer accuracy) the best, outperforming SOTA models on averaged performance across their tested tasks in automatic and human evaluation.

1 Introduction

Style transfer is a highly versatile task in natural language processing, where the goal is to modify the stylistic attributes of a text while maintaining its original meaning. A broad variety of stylistic attributes has been considered, including formality (Rao and Tetreault, 2018), gender (Prabhumoye et al., 2018), polarity (Shen et al., 2017) and civility (Laugier et al., 2021). Potential industrial applications are manifold and range from simplifying professional language to be intelligible to laypersons (Cao et al., 2020), the generation of more compelling news headlines (Jin et al., 2020), to related tasks such as text simplification for children and people with disabilities (Martin et al., 2020).

Data-driven style transfer methods can be classified according to the kind of data they use: parallel or non-parallel corpora in the two styles (Jin et al., 2021). To learn style transfer on non-parallel monostylistic corpora, current approaches take inspiration from unsupervised neural machine translation (UNMT) (Lample et al., 2018), by exploiting cycle consistency loss (Lample et al., 2019), iterative back-translation (Jin et al., 2019) and denoising autoencoders (DAE) (Laugier et al., 2021). As these approaches are similar to UNMT they suffer from the same limitations, i.e. poor performance relative to supervised neural machine translation (NMT) systems when the amount of UNMT training data is small and/or exhibits domain mismatch (Kim et al., 2020). Unfortunately, this is precisely the case for most existing style transfer corpora.

In this paper, we follow an alternative approach inspired by self-supervised NMT (Ruiter et al., 2021) that jointly learns online (near) parallel sentence pair extraction (SPE), back-translation (BT) and style transfer in a loop. The goal is to identify and exploit supervisory signals present in limited amounts of (possibly domain-mismatched) non-parallel data ignored by UNMT. The architecture of our system–called **S**elf-**S**upervised **S**tyle **T**ransfer (3ST)–implements an online self-supervisory cycle, where learning SPE enables us to learn style transfer on extracted parallel data, which iteratively improves SPE and BT quality, and thereby style transfer learning, in a virtuous circle.

We evaluate and compare 3ST to current state-of-the-art (SOTA) style transfer models on two established tasks: formality and polarity style transfer, where 3ST is the most balanced model and reaches top overall performance.

To gain insights into the performance of 3ST on an under-explored task, we also focus on the civil rephrasing task, which is interesting as i) it has been explored only twice before (Nogueira dos Santos et al., 2018; Laugier et al., 2021) and ii) it makes an important societal contribution in order to tackle hateful content online. We focus on performance and qualitative analysis of 3ST predictions

on this task's test set and identify shortcomings of the currently available data setup for civil rephrasing. On civil rephrasing, 3ST generates more neutral sentences than the current SOTA model while being on par in overall performance.

Our contribution is threefold:

- Efficient detection and exploitation of the supervisory signals in non-parallel social media content via jointly-learning *online* SPE and BT, outperforming SOTA models on averaged performance across civility, formality and polarity tasks in automatic and human evaluation (Δ in Tables 2 and 3).

- Simple end-to-end training of a single online model without the need for additional external style-classifiers or external SPE, enabling the initialization of the 3ST network on a DAE task, which leads to SOTA-matching fluency scores during human evaluation.

- A qualitative analysis that identifies flaws in the current data, emphasizing the need for a high quality civil rephrasing corpus.

2 Related Work

Style transfer can be treated as a supervised translation task between two styles (Jhamtani et al., 2017). However, for most style transfer tasks, parallel data is scarcely available. To learn style transfer without parallel data, prior research has focused on exploiting larger amounts of monostylistic data in combination with a smaller amount of style-labeled data. One such approach is using variational autoencoders and disentangled latent spaces (Fu et al., 2018), which can be further incentivized towards generating fluent or style-relevant content by fusing them with adversarial (Shen et al., 2017) or style-enforcing (Hu et al., 2017) discriminators. Chawla and Yang (2020) use a language model as the discriminator, leading to a more informative signal to the generator during training and thus more fluent and stable results. Li et al. (2018) argue that adversarially learned outputs tend to be low-quality, and that most sentiment modification is based on simple deletion and replacement of relevant words.

The above approaches focus on separating content and style, either in latent space or surface form, however this separation is difficult to achieve (Gonen and Goldberg, 2019). Dai et al. (2019) instead train a transformer together with a discriminator,
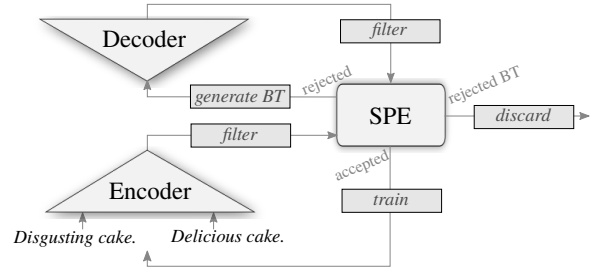

Figure 1: 3ST: joint learning of style transfer, SPE, and BT.

without disentangling the style features before decoding. Current approaches treat style transfer similar to an unsupervised neural machine translation (Artetxe et al., 2019) task. Jin et al. (2019) create pseudo-parallel corpora by extracting similar sentences offline from two monostylistic corpora to train an initial NMT model which is then iteratively improved using back-translation. Luo et al. (2019) use a reinforcement approach to further improve sentence fluency. Laugier et al. (2021) improve fluency without the need of any style-specific classifiers, giving their model a head start by initializing it on a pre-trained transformer model. Wang et al. (2020) argue that standard NMT training cannot account for the small differences between informal and formal style transfer, and apply style-specific decoder heads to enforce style differences.

Our approach differs from the two step approach of Jin et al. (2019), who first extract similar sentences from style corpora *offline* and then initialize their system by training on them. Ruiter et al. (2020) show that *joint online learning* to extract and translate in self-supervised NMT (SSNMT) leads to higher recall and precision of the extracted data. Following this observation, our 3ST approach performs similar sentence extraction and style transfer learning online with a single model in a loop. We further extend the SSNMT-based approach by combining it with UNMT methods, namely by generating additional training data via online back-translation, and by initialising our models with DAE trained in an unsupervised manner.

3 Self-Supervised Style Transfer (3ST)

Figure 1 shows the 3ST architecture, which uses the encoder outputs at training time as sentence representations to perform online (near) parallel sentence pair extraction (SPE) together with online back-translation (BT) and style transfer.

Self-Supervised NMT (SSNMT): SSNMT (Ruiter et al., 2019) is an encoder-decoder architecture that jointly learns to identify parallel data in non-parallel data and bidirectional NMT. Instead of using SSNMT on different language corpora to learn machine translation, we show how ideas from SSNMT can be used to learn a self-supervised style transfer system from non-parallel social media content. A single bidirectional encoder simultaneously encodes both styles and maps the internal representations of the two styles into the same space. This way, they can be used to compute similarities between sentence pairs in order to identify similar and discard non-similar ones for training. Formally, given two monostylistic corpora $S1$ and $S2$ of opposing styles, e.g. *toxic* and *neutral*, sentence pairs $(s_{S1} \in S1, s_{S2} \in S2)$ are input to an encoder-decoder system, a transformer in our experiments. From the internal representations for the input sentences s_{S1} and s_{S2}, SSNMT uses the sum of the word embeddings $w(s)$ and the sum of the encoder outputs $e(s)$ for filtering. The embedded pairs $\{w(s_{S1}), w(s_{S2})\}$ are scored using the margin-based measure (Artetxe and Schwenk, 2019). The same is done with pairs $\{e(s_{S1}), e(s_{S2})\}$. If a sentence pair is the most similar pair for both style directions *and* for both sentence representations, it is accepted for training, otherwise it is discarded. This sequence of scoring and filtering is denoted as **sentence pair extraction (SPE)** in 3ST. SPE improves style transfer and style transfer improves SPE online in a virtuous loop, resulting in a single system that jointly learns to identify its supervision signals in the data and to perform style transfer.

To address the characteristics of the monostylistic corpora we extend basic SSNMT in two ways:

Large-Scale Extraction: SSNMT extracts parallel data from *comparable corpora*, which contain smaller topic-aligned documents $\{d_{S1}, d_{S2}\}$ of similar content, thus reducing the search space during SPE from $|S1| \times |S2|$ to $|d_{S1}| \times |d_{S2}|$. However, style transfer corpora usually consist of large collections of (unaligned) sequences of a specific style, which forces the exploration of the full space. Improving over the one-by-one comparison of vector representations, we index[1] our data using FAISS (Johnson et al., 2019).

---

[1] As our internal representations change during the course of training, we re-index at each iteration over the data.

Corpus	Train	Dev	Test	$\varnothing$
CivCo-Neutral	136,618	500	–	–
CivCo-Toxic	399,691	500	4,878	14.9
Yahoo-Formal	1,737,043	4,603	2,100	12.7
Yahoo-Informal	3,148,351	5,665	2,741	12.4
Yelp-Pos	266,041	2,000	500	9.9
Yelp-Neg	177,218	2,000	500	10.7

Table 1: Number of sentences of the different tasks train, dev and test splits, as well as average number of tokens per sequence ($\varnothing$) of the tokenized test sets. Splits with target references available are underlined.

UNMT-Style Data Augmentation: We follow Ruiter et al. (2021) and use the current models' state to generate **back-translations** *online* from sentences rejected during SPE in order to increase the amount of supervisory signals to train on. Further, we initialize our style transfer models using **denoising autoencoding** using BART-style[2] noise (Lewis et al., 2020). After pre-training a DAE on the stylistic corpora, our models will generate fluent English sentences from the beginning and only need to learn to separate the two styles $S1$ and $S2$ during style transfer learning.

4 Experimental Setup

4.1 Data

Formality For the formality task, we use the test and development (dev) splits of the GYAFC corpus (Rao and Tetreault, 2018), which is based on the Yahoo Answers L6[3] corpus. However, as GYAFC is a parallel corpus and we want to evaluate our models in a setup where only monostylistic data is available, we follow Rao and Tetreault (2018) and re-create the training split without downsampling and without creating parallel reference sentences. For this, we extract all answers from the *Entertainment & Music* and *Family & Relationships* domains in the Yahoo Answers L6 corpus. We use a BERT classifier fine-tuned on the GYAFC training split to classify sentences as either *informal* or *formal*. This leaves us with a much ($46\times$) larger training split than the parallel GYAFC corpus, although consisting of non-parallel data where a single instance is less informative than a parallel one. We remove

---

[2] This is algorithmically equivalent to using a common pre-trained BART model for initialization, with the benefit that we have full control on the vocabulary size and data it is pre-trained on. We use this benefit by focusing the pre-training on in-domain data instead of generic out-of-domain data.

[3] www.webscope.sandbox.yahoo.com/catalog.php?datatype=l

sentences from our training data that are matched with a sentence in the official test-dev splits. We deduplicate the test-dev splits to match those used by Jin et al. (2019). For DAE pre-training, we sample sentences from Yahoo Answers L6.

Polarity We use the standard train-dev-test splits[4] of the Yelp sentiment transfer task (Shen et al., 2017). This dataset is already tokenized and lower-cased. Therefore, as opposed to the civility and formality tasks, we do not perform any additional pre-processing on this corpus. For DAE pre-training, we sample sentences from a generic Yelp corpus[5] and process them to fit the preprocessing of the Yelp sentiment transfer task, i.e. we lowercase and perform sentence and word tokenization using NLTK (Bird and Loper, 2004).

Civility The civil rephrasing task is rooted in the broader domain of hate speech research, which commonly focuses on the detection of hateful, offensive, or profane contents (Yang et al., 2019). Besides deletion, moderation, and generating counterspeech (Tekiroğlu et al., 2020), which are *reactive* measures after the abuse has already happened, there is a need for *proactive* ways of dealing with hateful contents to prevent harm (Jurgens et al., 2019). Civil rephrasing is a novel approach to fight abusive or profane contents by suggesting civil rephrasings to authors before their comments are published. So far, civil rephrasing has been explored twice before (Nogueira dos Santos et al., 2018; Laugier et al., 2021). However, their datasets are not publicly available. In order to compare the works, we reproduce the data sets used in Laugier et al. (2021). We follow their approach and create our own train and dev splits on the Civil Comments[6] (CivCo) dataset. Style transfer learning requires distinct distributions in the two opposing style corpora. To increase the distinction in our toxic and neutral datasets, we filter them using a list of slurs[7] such that the toxic portion contains only sentences with at least one slur, and the neutral portion does not contain any slurs in the list. Laugier et al. (2021) kindly provided us with the original test set used in their study. We removed

sentences contained in the test set from our corpus and split the remaining sentences into train and dev. To initialize 3ST on DAE with data related to the civility task domain, i.e. user comments, we sample sentences from generic Reddit comments crawled with PRAW[8].

Preprocessing On all datasets, excluding the polarity task data which is already preprocessed, we performed sentence tokenization using NLTK as well as punctuation normalization, tokenization and truecasing using standard Moses scripts (Koehn et al., 2007). Following Rao and Tetreault (2018), we remove sentences containing URLs as well as those containing less than 5 or more than 25 words. For the civility task only, we allow longer sequences of up to 30 words due to the higher average sequence length in this task (Laugier et al., 2021). We perform deduplication and language identification using `polyglot`[9]. We apply a bytepair encoding (Sennrich et al., 2016) of $8k$ mergeoperations. We add target style labels (e.g. *<pos>*) to the beginning of each sequence. Table 1 summarizes all train, dev and test splits.

4.2 Model Specifications

We base our 3ST code on OpenNMT (Klein et al., 2017), using a transformer-base with standard parameters, a batch size of 50 sentences and a maximum sequence length of 100 sub-word units. All models are trained until the attribute transfer accuracy on the development set has converged. Each model is trained on a single Titan X GPU, which takes around 2–5 days for a 3ST model.

For DAE pre-training, we use the task-specific DAE data split into $20M$ train sentences and $5k$ dev and test sentences each. To create the noisy sourceside data, we apply BART-style noise with $\lambda = 3.5$ and $p = 0.35$ for word sequence masking. We also add one random mask insertion per sequence and perform a sequence permutation.

For BERT classifiers, which we use to automatically evaluate the attribute transfer accuracy, we fine-tune a `bert-base-cased` model on the relevant classification task using early stopping with $\delta = 0.01$ and patience 5.

4.3 Automatic Evaluation

While 3ST can perform style transfer bidirectionally, we only evaluate on the *toxic→neutral* direc-

---

tion of the civility task, as the other direction, i.e. generation of toxic content, would pose a harmful application of our system. Similarly, the formality task is only evaluated for the *informal→formal* direction as this is the most common use-case (Rao and Tetreault, 2018). The polarity task is evaluated in both directions. We compare our model against current SOTA models: multi-class (MUL) and conditional (CON) style transformers by Dai et al. (2019), unsupervised machine translation (UMT) (Lample et al., 2019)[10] as well as models by Li et al. (2018) (DAR), Jin et al. (2019) (IMT), Laugier et al. (2021) (CAE), He et al. (2020) (DLA) and Shen et al. (2017) (SCA). Our automatic evaluation focuses on four main aspects:

Content Preservation (CP)   In style transfer, the aim is to change the style of a source sentence into a target style without changing the underlying meaning of the sentence. To evaluate CP, BLEU is a common choice, despite its inability to account for paraphrases (Wieting et al., 2019), which are at the core of style transfer. Instead, we use Siamese Sentence Transformers [11] [12] to embed the source and prediction and then calculate the cosine similarity.

Attribute Transfer Accuracy (ATA)   We want to transfer the style of the source sentence to the target style or attributes. Whether this transfer was successful is calculated using a BERT classification model. We train and evaluate our classifiers on the same data splits as the style-transfer models. This yields classifiers with Macro-F1 scores of 93.2 (formality), 87.4 (civility) and 97.1 (polarity) on the task-specific development sets. ATA is the percentage of generated target sentences that were labeled as belonging to the target style by the task-specific classifier.

Fluency (FLU)   As generated sentences should be intelligible and natural-sounding to a reader, we take their fluency into consideration during evaluation. The perplexity of a language model is often used to evaluate this (Krishna et al., 2020). However, perplexity is unbounded and therefore difficult to interpret, and has the limitation of favoring potentially unnatural sentences containing frequent words (Mir et al., 2019). We therefore use a RoBERTa (Liu et al., 2019) model[13] trained on

---

[10]Model outputs provided by He et al. (2020).
[11]Model `paraphrase-mpnet-base-v2`
[12]https://www.sbert.net/index.html
[13]www.huggingface.co/textattack/roberta-base-CoLA

CoLA (Warstadt et al., 2019) to label model predictions as either *grammatical* or *ungrammatical*.

Aggregation (AGG)   CP, ATA and FLU are important dimensions of style-transfer evaluation. A good style transfer model should be able to perform well across all three metrics. To compare overall style-transfer performance, it is possible to aggregate these metrics into a single value (Li et al., 2018). Krishna et al. (2020) show that corpus-level aggregation are less indicative for the overall performance of a system and we thus apply their sentence-level aggregation score, which ensures that each predicted sentence performs well across all measures, while penalizing predictions which are poor in at least one of the metrics. We also report the average AGG difference of a model m to 3ST across all tasks that m was tested on (Δ).

The automatic evaluation relies on external models, which are sensitive to hyperparameter choices during training. However, we use the same evaluation models across all style transfer model predictions and supplement the automatic evaluation with a human evaluation. As we observe consistency between the automatic and human evaluation, the underlying models used for the automatic evaluation can be considered to be sufficiently reliable.

4.4   Human Evaluation

We compare the performance of 3ST with each of the two strongest baseline systems per task, chosen based on their aggregated scores achieved in the automatic evaluation. These are: CAE and IMT for comparison in the polarity task, DAR and IMT for the formality task and CAE for the civility task. Due to the large number of models in the polarity task, we also include CON and MUL in the human evaluation, as they are strongest on ATA and CP respectively.

For each task, we sample 100 data points from the original test set and the corresponding predictions of the different models. We randomly duplicate 5 of the data points to calculate intra-rater agreement, resulting in a total of 105 evaluation sentences per system. Three fluent English speakers were asked to rate the content preservation, fluency and attribute transfer accuracy of the predictions on a 5-point Likert scale. In order to aggregate the different values, analogous to the automatic evaluation, we consider the transfer to be *successful* when a prediction was rated with a 4 or 5 across all three metrics (Li et al., 2018). The success rate

Task	Model	CP	FLU	ATA	AGG	Δ
Civ.	CAE	***64.2**	***80.6**	*81.9	<u>**39.8**</u>	-2.9
	3ST	60.5	75.3	**89.7**	<u>**39.0**</u>	0.0
For.	DAR	*64.5	*27.9	*66.0	*<u>14.2</u>	-30.0
	IMT	*71.5	*73.1	*79.2	*<u>45.2</u>	-7.6
	SCA	*54.4	*14.7	*27.4	*4.0	-40.3
	3ST	**75.6**	**83.1**	**84.9**	**54.7**	0.0
Pol.	CAE	*48.3	*76.4	*84.3	*<u>28.7</u>	-2.9
	CON	*57.5	*32.5	*__91.3__	*17.3	-18.0
	DAR	*50.4	*32.7	*87.8	*15.8	-30.0
	DLS	*50.9	*50.4	85.3	*20.1	-15.2
	IMT	*42.5	*__84.4__	*84.6	*<u>29.6</u>	-7.6
	MUL	*__62.6__	*42.3	*82.5	*20.4	-14.9
	SCA	*36.7	*19.5	*73.2	*5.5	-40.3
	UMT	*54.8	*55.7	85.4	*<u>24.2</u>	-11.1
	3ST	55.7	81.0	85.4	**35.3**	0.0

Table 2: Automatic scores for CP, FLU, ATA and their aggregated score (AGG) of SOTA models and our approach (3ST) across the Civ(ility), For(mality) and Pol(arity) tasks. Cross-task average AGG difference to 3ST under Δ. Best values per task in **bold** and models selected for human evaluation <u>underlined</u>. Values statistically significantly different ($p < 0.05$) from 3ST are marked with *.

(SR) is then defined as the ratio of successfully transferred instances over all instances. We also report the cross-task average SR difference of a model to 3ST (Δ).

All inter-rater agreements, calculated using Krippendorff-α, lie above 0.7, except for cases where most samples were annotated repeatedly with the same justified rating (e.g. a continuous FLU rating of 4) due to the underlying data distribution, which is sanctioned by the Krippendorff measure. Intra-rater agreement is at an average of 0.928 across all raters. A more detailed description of the evaluation task and a listing of the task- and rater-specific α-values is given in the appendix. For the ratings themselves, we calculate pair-wise statistical significance between SOTA models and 3ST using the Wilcoxon T test ($p < 0.05$).

5 Results and Analysis

5.1 Automatic Evaluation

Table 2 provides an overview of the CP, FLU, ATA and AGG results of all compared models across the three tasks.

Civility On attribute transfer accuracy, 3ST improves by +7.8 points over CAE, while CAE is stronger in content preservation (+3.7) and fluency (+5.3). There is, however, no statistically significant difference in the overall aggregated per-

formance of the models, indicating that they are equivalent in performance.

Formality 3ST substantially outperforms SOTA models in all four categories, with an overall performance (AGG) that surpasses the top-scoring SOTA model (IMT) by +9.5 points. This is indicative, as IMT was trained on a shuffled version of the parallel GYAFC corpus, which contains highly informative human written paraphrases, while 3ST was trained on a truly non-parallel corpus.

Polarity The polarity task has more recent SO-TAs to compare to, and the results show that no single model is best in all three categories. While MUL is strongest in content preservation (62.6), its fluency is low and outperformed by 3ST by +38.7 points, leading to a much lower overall performance (AGG) in comparison to 3ST (+14.9). Similarly, CON is strongest in attribute transfer accuracy (91.3) but has a low fluency (32.5), leading to a lower aggregated score than 3ST (+18). IMT is the strongest SOTA model with an overall performance (AGG) of 29.6 and the highest fluency score (84.4). Nevertheless, it is outperformed by 3ST by +5.7 points on overall performance (AGG), which is due to the comparatively better performance in content preservation (+13.2) of 3ST. Interestingly, unsupervised NMT (UMT) performs equally well on attribute transfer accuracy, while being slightly outperformed by 3ST in content preservation (+0.9). This may be due to the information-rich parallel instances automatically found in training by the SPE module. Further, 3ST has a much higher fluency than UMT (+25.3), which is due to its DAE pre-training. While 3ST is not top-performing in any of the three metrics CP, FLU and ATA, its top-scoring overall performance (AGG) shows that it is the most balanced model.

Overall Trends Table 2 shows that 3ST outperforms each of the SOTA models fielded in a single task (CON, DLS, MUL, UMT) by the respective AGG Δ, and all other models (CAE, DAR, IMT, SCA) on average AGG Δ[14]. 3ST achieves high levels of FLU, with ATA in the medium to high 80's, clear testimony to successful style transfer.

5.2 Human Evaluation

Human evaluation shows that 3ST has a high level of **fluency**, as it either outperforms or is on par with

---
[14]e.g. $\Delta(\mathrm{DAR}, 3\mathrm{ST}) = \frac{14.2+15.8}{2} - \frac{54.7+35.3}{2} = -30$ across Formality and Polarity.

Task	Model	CP	FLU	ATA	SR	Δ
Civ.	CAE	**2.97**	4.01	*2.50	17.0	-8.5
	3ST	2.80	**4.05**	**3.03**	**21.0**	**0.0**
For.	DAR	*2.75	*2.87	2.72	3.0	-8.0
	IMT	3.49	4.10	**2.83**	5.0	-13.0
	3ST	**3.75**	**4.29**	2.82	**11.0**	**0.0**
Pol.	CAE	*3.64	4.46	3.90	54.0	-8.5
	CON	4.20	*3.47	3.97	44.0	-23.0
	IMT	*3.54	**4.68**	3.84	47.0	-13.0
	MUL	*4.34	*3.66	3.68	41.0	-26.0
	3ST	3.99	4.58	**4.03**	**67.0**	**0.0**

Table 3: Average human ratings of CP, FLU, ATA and success rate (SR) on the three transfer tasks Civ(ility), For(mality) and Pol(arity). Cross-task average SR difference to 3ST (Δ). Best values per task in **bold**. Values statistically significantly different ($p < 0.05$) from 3ST are marked with *.

current SOTA models across all three tasks (Table 3), with ratings between 4.05 (civility) and 4.58 (polarity), and gains of up +1.42 (DAR, formality) points. According to the annotation protocol, a rating of 4 and 5 is to describe content written by native speakers, thus annotators deemed most generated sentences to have been written by a native speaker of English.

For **content preservation** and **attribute transfer**, there seems to be a trade-off. In the formality task, 3ST outperforms or is on par with current SOTAs on CP with gains between +0.26 (IMT) and +1.0 (DAR) points, and ATA is on par with the SOTA (−0.01, IMT). Note that for all models tested on the formality task, the success rate is low. This is due to the nature of the training data, where many sentences in the formal portion of the dataset tend to be rather neutral, i.e. neither formal nor informal, rather than truly formal sentences. For the civility task, on the other hand, 3ST outperforms the current SOTA on ATA with gains of +0.53 (CAE) while being on par on CP (−0.17). For the polarity task, the CP is slightly below the best model (−0.35, MUL).

While some models are strong on single values, 3ST has the highest success rate (SR) across all tasks. 3ST outperforms each of the single task models (DAR, CON, MUL) on SR by Δ and each of the multitask models (CAE, IMT) by average cross-task SR Δ, again highlighting that it balances best between the three capabilities CP, FLU and ATA, which leads to best-performing style transfer predictions.

(1)	SRC	*What our ignorant PM, Mad McCallum and stupid Liberal politicians going to say?*
	CAE	*what our pm, ~~trudeau~~ and his liberals are going to do about this?<u>............................</u>*
	3ST	<u>*Mad McCallum, what are our politicians going to say?*</u>
(2)	SRC	*Dear Hipster Jackass- Go to Bend.*
	CAE	*dear hippie ~~– go to hawaiian to get around........~~*
	3ST	*Dear Hipster Jackass- Go to Bend.*
(3)	SRC	*Trump's a liar.*
	CAE	*trump's a <u>liar.</u>*
	3ST	*Trump's a <u>good man.</u>*
(4)	SRC	*Says the idiot on perpetual welfare.*
	CAE	*says the ~~author on the daily basis, on the basis~~ of perpetual welfare.*
	3ST	***Says the guy on perpetual welfare.***
(5)	SRC	*A muslim racist.* <u>←——→</u>
	CAE	*a muslim <u>minority.</u>*
	3ST	~~Not a democrat.~~
(6)	SRC	*Quit trying to justify what this jackass did.*
	CAE	*quit trying to justify what this <u>jackass did.</u>*
	3ST	***Quit trying to justify what he did.***
(7)	SRC	*There was no consensus, 1 idiot and everyone else in the situation let him know he was in the wrong.*
	CAE	*there was no consensus, <u>no one in the room and everyone in the room</u> knew he was in the wrong place.*
	3ST	***No, there was no consensus in the past, and everyone else knew he was in the wrong place.***

Table 4: 3ST and SOTA model (CAE) predictions on the CivCo test set, with **adequate** predictions, <u>error in structure</u>, target attribute, <u>stance reversal</u>, and ~~hallucinations~~ marked.

5.3 Qualitative Analysis

For our qualitative analysis, we focus on the civility task as this is a challenging, novel task and we want to understand its limitations. We analyze the same subset of the test set used for human evaluation and annotate common mistakes. Common errors in the neutral counterparts generated by 3ST can be classified into four classes. We observe fluency or *structural errors* (11% of sentences), e.g. a subject becoming a direct form of address (Table 4, Ex-1). *Attribute errors* (14%) (Ex-2), where toxic content was not successfully removed, are another common source of error. Similarly to Laugier et al. (2021), we observe *stance reversal* (14%), i.e. where a usually negative opinion in the original source sentence is reversed to a positive polarity (Ex-3). This is due to a negativity bias on the toxic side of the CivCo corpus, while the neutral side contains more positive sentences, thus introducing an incentive to

Task	Model	CP	FLU	ATA	AGG
Civ.	3ST	60.5	**75.3**	89.7	**39.0**
	-SPE	*89.5	*39.4	*12.1	*3.7
	-BT	*44.4	*59.4	90.3	*22.8
	-DAE	*36.8	*43.3	*97.5	*15.7
	-BT-DAE	*37.8	*43.8	*95.3	*16.4
For.	3ST	75.6	83.1	84.9	**54.7**
	-SPE	*99.3	*73.4	*17.7	*14.8
	-BT	*66.4	*85.1	*92.6	*52.8
	-DAE	*55.7	*64.2	*93.1	*35.1
	-BT-DAE	*57.8	*79.5	**94.0**	*44.5
Pol.	3ST	55.7	**81.0**	85.4	**35.3**
	-SPE	*100.0	*80.5	*2.9	*1.9
	-BT	*44.0	*79.0	*88.3	*29.2
	-DAE	*29.8	*43.6	*89.7	*11.6
	-BT-DAE	*38.0	*63.3	**91.1**	*21.5

Table 5: 3ST Ablation. CP, FLU and ATA with SPE, BT, DAE removed. Best values per task in **bold**.

translate negative sentiment to positive sentiment. Unlike Laugier et al. (2021), we do not observe that hallucinations are most frequent at the end of a sequence (*supererogation*). Rather, *related hallucinations*, where unnecessary content is mixed with words from the original source sentence, are found at arbitrary positions (23%, Ex-4, CAE). We observe few hallucinations where a prediction has no relation with the source (4%, Ex-5).

Phenomena such as hallucinations can become amplified through back-translation (Raunak et al., 2021). However, as they are most prevalent in the civility task, hallucinations in this case are likely originally triggered by long source sentences that i) overwhelm the current models' capacity, and ii) add additional noise to the training. It is less likely that a complex sentence has a perfect rephrasing to match with and therefore instead it will match with a similar rephrasing that introduces additional content, i.e. noise. For reference, the average length of source sentences that triggered hallucinations was 21.9 words, while for adequate re-writings (39%), it was 8 words. Note that we capped sentence lengths to 30 words in the training data while the test data contained sentences with up to 85 words.

Successful rephrasings are usually due to one of two factors. 3ST either *replaces* profane words by their neutral counterparts (Ex-{4,6}) or *removes* them (Ex-7).

5.4 Ablation Study

To analyze the contribution of the three main components (SPE, BT and DAE) of 3ST, we remove them individually from the original architecture and observe the performance of the resulting models on the three different tasks (Table 5). Without SPE, the model merely copies source sentences without performing style transfer, resulting in a large drop in overall performance (AGG). This shows in the low ATA scores (1.9–14.8), which are in direct correlation with the extremely high scores in CP (89.5–100.0) achieved by this model. This underlines that SPE is vital to the style-transfer capabilities of 3ST, as it retrieves similar paraphrases from the style corpora and lets 3ST train on these. This pushes the system to generate back-translations which themselves are paraphrases that fulfill the style-transfer task. BT and DAE are integral parts of 3ST, too, that improve over the underlying self-supervised neural machine translation (-BT-DAE) approach. This can be seen in the drastic drops of CP and FLU scores when BT and DAE techniques are removed. Especially DAE is important for the fluency of the model. The gains in CP and FLU through BT and DAE come at a minor drop in ATA.

6 Conclusion

3ST is a style transfer architecture that efficiently uses the supervisory signals present in non-parallel social media content, by i) jointly learning style transfer and similar sentence extraction during training, ii) using online back-translation and iii) DAE-based initialization. 3ST gains strong results on all three metrics FLU, ATA and CP, outperforming SOTA models on averaged performance (Δ) across their tested tasks in automatic (AGG) and human (SR) evaluation. We present one of the first studies on automatic civil rephrasing and, importantly, identify current weaknesses in the data, which lead to limitations in 3ST and other SOTA models on the civil rephrasing task. Our code and model predictions are publicly available at https://github.com/uds-lsv/3ST.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. ACL.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. ACL.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online. ACL.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. ACL.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, volume 32, pages 663–670, New Orleans, LA.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, MN. ACL.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *Proceedings of ICLR*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, DK. ACL.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2021. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

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

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3097–3109, Hong Kong, China. ACL.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3).

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. ACL.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. ACL.

Klaus Krippendorff. 2004. *Content Analysis, an Introduction to Its Methodology*. Sage Publications, Thousand Oaks, Calif.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR*, Vancouver, BC, Canada.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR*, New Orleans, LA.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. ACL.

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

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Multilingual unsupervised sentence simplification. *CoRR*, abs/2005.00352.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, MN. ACL.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. ACL.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. ACL.

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

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. ACL.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith. 2019. Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. ACL.

Dana Ruiter, Dietrich Klakow, Josef van Genabith, and Cristina España-Bonet. 2021. Integrating unsupervised data generation into self-supervised neural machine translation for low-resource languages. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 76–91, Virtual. Association for Machine Translation in the Americas.

Dana Ruiter, Josef van Genabith, and Cristina España-Bonet. 2020. Self-induced curriculum learning in self-supervised neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2560–2571, Online. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. ACL.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30, pages 6830–6841. Curran Associates, Inc.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. ACL.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. ACL.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

A Human Evaluation Task

We perform a human evaluation to assess the quality of the top performing models according to automatic metrics.

We select 3 systems for Formality, 5 systems for Polarity and the only 2 systems available for the Civility task. For each of these tasks, we sample 100 data points from the original test set and the corresponding predictions of the different models. We randomly duplicate 5 of the points for quality controls, resulting in evaluation tests with 105 sentences per system. Three fluent English speakers (*raters*) were shown with pairs source–system prediction and were asked to rate the content preservation, fluency and attribute transfer accuracy of the predictions on a 5-point Likert scale. Raters were payed around 10 Euros per hour of work.

We calculate the reliability of the ratings using Krippendorff-α (Krippendorff, 2004). Table 6 shows the inter-rater agreement measured by α for content preservation (CP), fluency (FLU) and

Task	Krippendorff-α		
	CP	FLU	ATA
Civility	0.744	0.579	0.688
Formality	0.751	0.718	0.352
Polarity	0.426	0.705	0.837

Table 6: Inter-rater agreement calculated using Krippendorff-α across the different tasks and metrics.
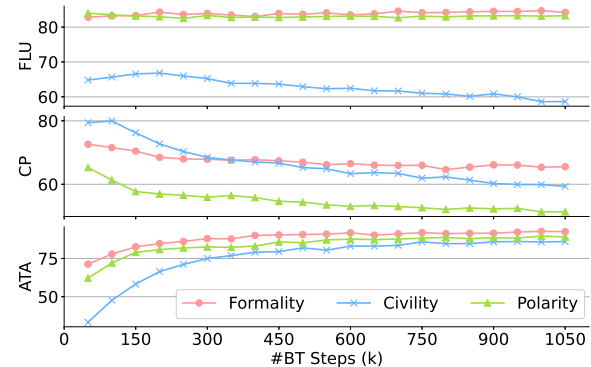

Figure 2: FLU, CP and ATA of generated back-translations (BTs) during training of 3ST on the three transfer tasks.

attribute transfer accuracy (ATA). Notice that α significantly differs between tasks. The lower α on polarity CP and formality task ATA is due to the repetitive ratings of the same kind. i.e. $4, 5$ on polarity CP and 3 for formality ATA, which is sanctioned by the Krippendorff measure. For the intra-rater agreement estimated from 40 duplicated sentences per rater, we obtain values of 0.988 (Rater-1), 0.869 (Rater-2) and 0.927 (Rater-3).

B Performance Evolution

The back-translations that 3ST generates during training give us a direct insight into the changing state of the model throughout the training process. We thus automatically evaluate ATA, FLU and CP on the back-translations over time.

BT **fluency** (Figure 2, top) on all three tasks is strong already at the beginning of training, due to the DAE pre-training. For the formality and polarity task, the high level of FLU remains stable (~ 80) throughout training, while for Civility it slightly drops. This underlines the observation that the Civility task is prone to hallucinations due to the sparse amount of parallel supervisory signals in the dataset, which then leads to lower FLU scores.

For all tasks, **content preservation** between the generated BTs and the source sentences is already high at the beginning of training. This is due to

the DAE pre-training which taught the models to copy and denoise inputs. All of the models decay in CP over time, showing that they are slowly diverging from merely copying inputs. CP scores of the formality and the polarity tasks are close to convergence at around $1M$ train steps, while the scores of the civility task keep on decaying. This may again be due to the complexity of the data of the toxicity task, which contains longer sequences than the other two. This can lead to hallucinations when supervisory signals are lacking.

As back-translation CP decays, **attribute transfer** accuracy increases dramatically. Especially on the civility task, where the initial accuracy is low (8.2%) but grows to ATA $\sim$82%. For the other two tasks, the curves are less steep, and most of the transfer is learned at the beginning, within the first $300k$ generated BTs, after which they converge with ATA $\sim$95% (formality) and $\sim$88% (polarity). This shows the trade-off between attribute accuracy and content preservation: the higher the ATA, the lower the CP score. Nevertheless, as ATA converges earlier than CP (for formality and polarity tasks), an earlier training stop can easily benefit content preservation while having little impact on the already converged ATA.

C   Sample Predictions

For each of the three tasks, Civility, Formality and Polarity, we randomly sample 5 source sentences from the respective test sets. In Table 7 we present these source sentences together with the corresponding prediction of 3ST and the two best-scoring SOTA models with respect to the AGG score per task, namely CAE for Civility, DAR and IMT for Formality and CAE and IMT for Polarity.

Civility
SRC *It is time to impeach this idiot judge.*
CAE *it is time to impeach this judge.*
3ST *It is time to impeach this judge.*

SRC *This is classic example of collective corporate stupidity and individual managerial malice.*
CAE *this is classic case of corporate welfare and collective bargaining.*
3ST *This is classic example of collective corporate greed and individual managerial malice.*

SRC *You silly goose!*
CAE *you mean the goose, right?*
3ST *You forgot the goose!*

SRC *Afraid of how idiotic social engineering makes people look?*
CAE *imagine how socially acceptable some of the people make?*
3ST *Afraid of how social engineering works.*

SRC *Stupid idea.*
CAE *no idea..............*
3ST *Not a good idea.*

Formality
SRC *haha julesac is funny, but mean.*
DAR *is funny , but I understand what you mean .*
IMT *That is funny . Those silly people annoy me !*
3ST *Julesac is very funny.*

SRC *DON'T LET HER RULE YOUR LIFE, SHE WILL JUST HAVE TO LEARN TO DEAL WITH IT.*
DAR *LET HER BE , SHE WILL LEARN TO DEAL WITH IT .*
IMT *TELL HER YOUR TRUE FEELINGS , IT MAY SHOCK HER BUT WILL WORK .*
3ST *Do NOT LET HER RUN WITH YOU, SHE WILL NEVER HAVE TO WORK.*

SRC *cause it's buy one take one.*
DAR *I can not wait to buy one take one .*
IMT *Because it is buy one take one .*
3ST *You can buy one.*

SRC *All my votes are going to Taylor Hicks though...*
DAR *All my votes are , and I am going to Hicks*
IMT *All my votes are going to Taylor .*
3ST *All my votes are going to be Taylor Hicks.*

SRC *but paris hilton isn't far behind.*
DAR *I do not know but is n't far behind .*
IMT *I ca n't read the stars , just find another way to say it .*
3ST *Paris hilton is far behind.*

Polarity
SRC *even if i was insanely drunk , i could n't force this pizza down .*
CAE *even if i was n't in the mood , i loved this place .*
IMT *honestly , i could n't stop eating it because it was so good !*
3ST *even if i was drunk , i could still force myself .*

SRC *i will definitely return often !*
CAE *i will not return often ! ! ! !*
IMT *i will definitely not return !*
3ST *i will not return often !*

SRC *no massage with my manicure or pedicure .*
CAE *great massage with great pedicure and manicure .*
IMT *awesome relaxation and massage with my pedicure .*
3ST *great massage with my manicure and pedicure .*

SRC *excellent knowledgeable dentist and staff !*
CAE *excellent dentist and dental hygienist ! ! ! !*
IMT *not very knowledgeable staff !*
3ST *horrible dentist and staff !*

SRC *do not go here if you are interested in eating good food .*
CAE *definitely recommend this place if you are looking for good food at a good price .*
IMT *if you are looking for consistent delicious food go here .*
3ST *if you are looking for good food , this is the place to go .*

Table 7: Examples of 3ST and SOTA model predictions.