Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

Léo Laugier
Télécom Paris,
Institut Polytechnique de Paris, France
leo.laugier@telecom-paris.fr

John Pavlopoulos
Department of Computer and System Sciences
Stockholm University, Sweden
annis@aueb.gr

Jeffrey Sorensen
Google
sorenj@google.com

Lucas Dixon
Google
ldixon@google.com

Abstract

Platforms that support online commentary, from social networks to news sites, are increasingly leveraging machine learning to assist their moderation efforts. But this process does not typically provide feedback to the author that would help them contribute according to the community guidelines. This is prohibitively time-consuming for human moderators to do, and computational approaches are still nascent. This work focuses on models that can help suggest rephrasings of toxic comments in a more civil manner. Inspired by recent progress in unpaired sequence-to-sequence tasks, a self-supervised learning model is introduced, called CAE-T5[1]. CAE-T5 employs a pre-trained text-to-text transformer, which is fine tuned with a denoising and cyclic auto-encoder loss. Experimenting with the largest toxicity detection dataset to date (Civil Comments) our model generates sentences that are more fluent and better at preserving the initial content compared to earlier text style transfer systems which we compare with using several scoring systems and human evaluation.

1 Introduction

There are many ways to express our opinions. When we exchange views online, we do not always immediately measure the emotional impact of our message. Even when the opinions expressed are legitimate, well-intentioned and constructive, a poor phrasing may make the conversation go awry (Zhang et al., 2018a). Recently, Natural Language Processing (NLP) research has tackled the problem of abusive language detection by developing accurate classification models that flag toxic (or abusive, offensive, hateful) comments (Davidson et al., 2017; Pavlopoulos et al., 2017; Wulczyn et al., 2017; Gambäck and Sikdar, 2017; Fortuna and Nunes, 2018; Zhang et al., 2018a; Van Hee et al., 2018; Zampieri et al., 2019).

INPUT OFFENSIVE COMMENT	you now have to defend this clown along with his russian corruption.
GENERATED CIVIL COMMENT	**you now have to defend this guy from his russian ties.....**
INPUT OFFENSIVE COMMENT	blaming trudeau and the government is just stupid.
GENERATED CIVIL COMMENT	**blaming trudeau and the liberal government is just wrong.**
INPUT OFFENSIVE COMMENT	dubya[2] was a moron.
GENERATED CIVIL COMMENT	**dubya was a republican.**

Table 1: Examples of offensive sentences from the Civil Comments test set and the more civil rephrasing generated by our model. The third example shows that its strategy may involve shifting the original intent, since "republican" is not a non-offensive synonym of "moron".

The prospect of healthier conversations, nudged by Machine Learning (ML) systems, motivates the development of Natural Language Understanding and Generation (NLU and NLG) models that could later be integrated in a system suggesting alternatives to vituperative comments before they are posted. A first approach would be to train a text-to-text model (Bahdanau et al., 2014; Vaswani et al., 2017) on a corpus of parallel comments where each offensive comment has a courteous and fluent rephrasing written by a human annotator. However, such a solution requires a large paired labeled dataset, in practice difficult and expensive to collect (see Section 4.5). Consequently, we limit our setting to the unsupervised case where the comments are only annotated in attributes related to toxicity, such as the Civil Comments dataset (Borkan et al., 2019). We summarize our investigations with the following research question:

[1]The code can be found at https://github.com/LeoLaugier/conditional-auto-encoder-text-to-text-transfer-transformer.

[2]A nickname for George W. Bush.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1442–1461
April 19 - 23, 2021.

RQ: *Can we fine-tune end-to-end a pre-trained text-to-text transformer to suggest civil rephrasings of rude comments using a dataset solely annotated in toxicity?*

Answering this question might provide researchers with an engineering proof-of-concept that would enable further exploration of the many complex questions that arise from such a tool being used in conversations. The main contributions of this work are the following:

- We addressed for the second time the task of unsupervised civil rephrases of toxic texts, relying for the first time on the Civil Comments dataset, and achieving results that reflect the effectiveness of our model over baselines.
- We developed a non-task specific approach (i.e. with no human hand-crafting in its design) that can be generalized and later applied to related and/or unexplored attribute transfer tasks.

While several of the ideas we combine in our model have been studied independently, to the best of our knowledge, no existing unsupervised models combine sequence-to-sequence bi-transformers, transfer learning from large pre-trained models, and self-supervised fine-tuning (denoising auto-encoder and cycle consistency). We discuss the related work introducing these tools and techniques in the following section.

2 Related work

Unsupervised complex text attribute transfer (like civil rephrasing of toxic comments) remains in its early stages, and our particular applied task has only a single antecedent (Nogueira dos Santos et al., 2018). There is a great variety of useful works to tackle the task and this section attempts to summarize the vast majority of these works. We describe below the recent strategies (such as attention mechanisms Bahdanau et al., 2014) that led to significant progress in supervised NLU and NLG tasks. Then, we present the most related lines of work in unsupervised text-to-text tasks.

2.1 Transformers[3] are state-of-the-art architectures in NLP

Vaswani et al. (2017) showed that transformer architectures, based on attention mechanisms, achieved state-of-the-art results when applied to supervised Neural Machine Translation (NMT). More generally, transformers have proven capable in various NLP and speech tasks (Dong et al., 2018; Huang et al., 2019; Le et al., 2019; Li et al., 2019). Moreover, transformers benefit from pre-training before being fine-tuned on downstream tasks (Devlin et al., 2019; Dai et al., 2019b; Yang et al., 2019; Conneau and Lample, 2019; Raffel et al., 2019). Subsequent research has adopted uni-transformers in many supervised classification and regression tasks (Devlin et al., 2019) and in unsupervised language modeling (Radford et al., 2019; Keskar et al., 2019; Dathathri et al., 2020), until Raffel et al. (2019) proposed a unified pre-trained bi-transformer applicable to any text classification, text regression and text-to-text task. Further, recent works tackle the language detoxification of unconditional language models (Krause et al., 2020; Gehman et al., 2020).

2.2 Unsupervised losses enable training text-to-text models end-to-end

After the success of unsupervised image-to-image style transfer in computer vision (CV), some approaches have addressed unsupervised text-to-text tasks. Unsupervised Neural Machine Translation (UNMT) is maybe the most promising of them. Artetxe et al. (2018); Conneau et al. (2018); Lample et al. (2018a,b); Conneau and Lample (2019) introduced methods based on techniques aligning the embedding spaces of monolingual datasets and tricks such as denoising auto-encoding losses (Vincent et al., 2008) and back-translation (Sennrich et al., 2015; Edunov et al., 2018).

Abstractive summarization (or sentence compression) is also studied in unsupervised settings. Baziotis et al. (2019) trained a model with a compressor-reconstructor strategy similar to back-translation while Liu et al. (2019b) trained a denoising auto-encoder that embeds sentences and paragraphs in a common space.

Unsupervised attribute transfer is the task most related to our work. It mainly focuses on sentiment transfer with standard review datasets (Maas et al.,

[3]To avoid confusion we denote as bi-transformer the original encoder-decoder transformer whereas encoder-only and decoder-only models are called uni-transformers here.

2011; He and McAuley, 2016; Shen et al., 2017; Li et al., 2018), but also addresses sociolinguistic datasets containing text in various registers (Gan et al., 2017; Rao and Tetreault, 2018) or with different identity markers (Voigt et al., 2018; Prabhumoye et al., 2018; Lample et al., 2019). When paraphrase generation aims at being explicitly attribute-invariant, it is referred as obfuscation or neutralization (Emmery et al., 2018; Xu et al., 2019b; Pryzant et al., 2020). Literary style transfer (Xu et al., 2012; Pang and Gimpel, 2019) has also been tackled by recent work. Here, we apply attribute transfer to a large dataset annotated in toxicity, but we also use the Yelp review dataset from Shen et al. (2017) for comparison purposes (see Section 4).

Initial unsupervised attribute transfer approaches sought to build a shared and attribute-agnostic latent representation encoding for the input sentence, with adversarial training. Then, a decoder, aware of the destination attribute, generated a transferred sentence (Shen et al., 2017; Hu et al., 2017; Fu et al., 2018; Zhang et al., 2018c; Xu et al., 2018; John et al., 2019).

Unsupervised attribute transfer approaches that do not rely on a latent space are also present in literature. Li et al. (2018) assumed that style markers are very local and proposed to delete the tokens most conveying the attribute, before retrieving a second sentence in the destination style. They eventually combined both sentences with a neural network. Lample et al. (2019) applied UNMT techniques from Conneau and Lample (2019) to several attribute transfer tasks, including social media datasets. Xu et al. (2018); Gong et al. (2019); Luo et al. (2019); Wu et al. (2019a) trained models with reinforcement learning. Dai et al. (2019b) introduced unsupervised training of a transformer called StyleTransformer (ST) with a discriminator network. Our approach differs from these unsupervised attribute transfer models in that they did not either leverage large pre-trained transformers, or train with a denoising objective.

The most similar work to ours is Nogueira dos Santos et al. (2018) who trained for the first time an encoder-decoder rewriting offensive sentences in a non-offensive register with non-parallel data from Twitter (Ritter et al., 2010) and Reddit (Serban et al., 2017). Our approach differs in the following aspects. First, we use transformers pre-trained on a large corpus instead of randomly initialized RNNs for encoding and decoding. Second, their approach involves collaborative classifiers to penalize generation when the attribute is not transferred, while we train end-to-end with a denoising auto-encoder. Even if their model shows high accuracy scores, it suffers from low fluency, with offensive words being often replaced by a placeholder (e.g. "big" instead of "f*cking").

As underlined by Lample et al. (2019), applying Generative Adversarial Networks (GANs) (Zhu et al., 2017) to NLG is not straightforward because generating text implies a sampling operation that is not differentiable. Consequently, as long as text is represented by discrete tokens, loss gradients computed with a classifier cannot be back-propagated without tricks such as the REINFORCE algorithm (He et al., 2016) or the Gumbel-Softmax approximation (Baziotis et al., 2019) which can be slow and unstable. Besides, controlled text generation (Ficler and Goldberg, 2017; Keskar et al., 2019; Le et al., 2019; Dathathri et al., 2020) is a NLG task that consists of a language model conditioned on the attributes of the generated text such as the style. But a major difference with attribute transfer is the absence of a constraint regarding the preservation of the input's content.

3 Method

3.1 Formalization of the attribute text rewriting problem

Let X_T and X_C be our two non-parallel corpora of comments satisfying the respective attributes "toxic" and "civil". Let $X = X_T \cup X_C$. We aim at learning a parametric function f_θ mapping a pair of source sentence x and destination attribute a to a fluent sentence y satisfying a and preserving the meaning of x. In our case, there are two attributes "toxic" and "civil" that we assumed to be mutually exclusive. We denote $\alpha(x)$ to be the attribute of x and $\bar{\alpha}(x)$ the other attribute (for instance when $\alpha(x) =$ "civil", then $\bar{\alpha}(x) =$ "toxic"). Note that $f_\theta(x, \alpha(x))$ can simply be x.

3.2 Our approach is based on bi-conditional encoder-decoder generation

Our approach is to train an autoregressive (AR) language model (LM) conditioned on both the input text x and the destination attribute a.

We compute f_θ with a LM $p(y|x, a; \theta)$. As we do not have access to ground-truth targets y, we propose in section 3.3 a training function that we assume to maximize $p(y|x, a; \theta)$ if and only if y is

a fluent sentence with attribute a and preserving x's content. Additionnaly, we use an AR generating model where inference of $\hat{y}$ is sequential and the token generated at step $t+1$ depends on the tokens generated at previous steps: $p(\hat{y}_{t+1}|\hat{y}_{:t}, x, a; \theta)$.

To condition on the input text, we follow the work of Bahdanau et al. (2014); Vaswani et al. (2017); Nogueira dos Santos et al. (2018); Conneau and Lample (2019); Lample et al. (2019); Dai et al. (2019a); Liu et al. (2019b); Raffel et al. (2019) and opt for an encoder-decoder framework. Lample et al. (2019); Dai et al. (2019a) argue that in unsupervised attribute rewriting tasks, encoders do not necessarily output disentangled representations, independent of its attribute. However, the t-SNE visualization of the latent space in Liu et al. (2019b) allowed us to assume that encoders can output a latent representation z, attending to content rather than on an attribute, with a similar training.

The LM is conditioned on the destination attribute with control codes introduced by Keskar et al. (2019). A control code is a fixed sequence of tokens prepended to the decoder's input s, and supposed to prepare the generation in the space of sentences with the destination attribute a. We define $\gamma(a, s) = concat(c(a), s)$ where $c(a)$ is the control code of attribute a.

3.3 Training the encoder-decoder with an unsupervised objective

Denoising objectives to train transformers are an effective self-supervised strategy. Devlin et al. (2019); Yang et al. (2019) pre-trained a uni-transformer encoder as a masked language model (MLM) to teach the system general-purpose representations, before fine-tuning on downstream tasks. Conneau and Lample (2019); Lample et al. (2019); Song et al. (2019); Liu et al. (2019b); Raffel et al. (2019) explore various deshuffling and denoising objectives to pre-train or fine-tune bi-transformers.

During training, we corrupt the encoder's input x with the noise function from Devlin et al. (2019): η masks tokens randomly with probability 15%. Then, masks are replaced by a random token in the vocabulary with probability 10% or left as a sentinel (a shared mask token) with probability 90%. We train the model as an denoising auto-encoder (DAE), meaning that we minimize the negative log-likelihood

$$\mathcal{L}_{\text{DAE}} = \mathbb{E}_{x\sim X}\ [-\log p(x|\eta(x), \alpha(x); \theta)]$$

The hypothesis is that optimizing the DAE objective teaches the controlled generation to the model.

Inspired by an equivalent approach in unsupervised image-to-image style transfer (Zhu et al., 2017), we add a cycle-consistency (CC) objective (Nogueira dos Santos et al., 2018; Edunov et al., 2018; Prabhumoye et al., 2018; Lample et al., 2019; Conneau and Lample, 2019; Dai et al., 2019a):

$$\mathcal{L}_{\text{CC}} = \mathbb{E}_{x\sim X}\ \left[-\log p(x|f_{\tilde{\theta}}(x, \bar{\alpha}(x)), \alpha(x); \theta)\right]$$

which enforces content preservation in the generated prediction. As the cycle-consistency objective computes a non-differentiable AR pseudo-prediction $\hat{y}$ during stochastic gradient descent training, gradients are not back-propagated to $\tilde{\theta} = \hat{\theta}_{\tau-1}$ at training step τ.

Finally, the loss function sums the DAE and the CC objectives with weighting coefficients: $\mathcal{L} = \lambda_{\text{DAE}}\mathcal{L}_{\text{DAE}} + \lambda_{\text{CC}}\mathcal{L}_{\text{CC}}$

3.4 The text-to-text bi-transformer architecture

The architectures for the encoder and decoder are uni-transformers. Contrary to Vaswani et al. (2017); Conneau and Lample (2019); Raffel et al. (2019) we do not keep decoder's layers computing cross attention between the encoder's outputs h and the decoder hidden variables because generation suffers from too much conditioning on the input sentence and we observe no significant change in the output sentence. Rather, we follow Liu et al. (2019b) and compute the latent representation z with an affine transformation of the encoder's hidden state h_0 (corresponding to the first token of the input text). Let $x \in X$ be the input sequence of token. It is embedded then encoded by the uni-transformer encoder:

$$\begin{aligned} x_{\text{Emb}} &= f_{\theta_{\text{Emb}}}(x) \\ h_{\text{Enc}} &= f_{\theta_{\text{Enc}}}(x_{\text{Emb}}) \\ h^0_{\text{Enc}} &= h_{\text{Enc}}[0, :] \\ z &= f_{\theta_{\text{Dense}}}(h^0_{\text{Enc}}) \end{aligned}$$

z is an aggregate sequence representation for the input. There are different heuristics that can be used to integrate it in the decoder. We considered summing z to the embedding of each token of the uni-transformer decoder's input s since it balances the backpropagation of the signals coming from the original input and from the output being generated in the destination attribute space and it worked well

in practice in our experiments.

$$\gamma_{\text{Emb}} = f_{\theta_{\text{Emb}}}(\gamma(a, s))$$
$$h_{Dec} = f_{\theta_{\text{Enc}}}(\gamma_{\text{Emb}} + z)$$
$$\hat{y} = f_{\theta_{\text{LMHead}}}(h_{\text{Dec}})$$

Plus, the encoder and the decoder uni-transformers share the same embedding layer and the LM Head is tied to the embeddings.

Except for the dense layer computing the latent variable z, all parameters are coming from the pre-trained bi-transformer published by Raffel et al. (2019). Thus, our DAE and CC objectives **fine-tune** T5's parameters and this is why we call our model a conditional auto-encoder text-to-text transfer transformer (CAE-T5).

4 Experiments

4.1 Datasets

We employed the largest publicly available toxicity detection dataset to date, which was used in the 'Jigsaw Unintended Bias in Toxicity Classification' Kaggle challenge.[4] The 2M comments of the **Civil Comments dataset** stem from a commenting plugin for independent news sites. They were created from 2015 to 2017 and appeared on approximately 50 English-language news sites across the world. Each of these comments was annotated by crowd raters (at least 3 each) for toxicity and toxicity subtypes (Borkan et al., 2019).

Following the work of Dai et al. (2019a) for the IMDB Movie Review dataset (positive/negative sentiment labels), we constructed a sentence-level version of the dataset. Initially, we fine-tuned a pre-trained BERT (Devlin et al., 2019) toxicity classifier on the Civil Comments dataset. Then, we split the comments in sentences with NLTK's sentence tokenizer.[5] Eventually, we created X_T (respectively X_C) with sentences whose system-generated toxicity score (using our BERT classifier) is greater than 0.9 (respectively less than 0.1) to increase the dataset's polarity. The test ROC-AUC of the toxicity classifier is 0.98 with a precision of 0.95 and a recall of 0.38. Even with this low recall $|X_T|$ is large enough (approx. 90k, see Table 2).

We also conducted a comparison to other style transfer baselines on the **Yelp Review Dataset** (Yelp), commonly used to compare unsupervised attribute transfer systems. It consists of restaurant and business reviews annotated with a binary positive / negative label. Shen et al. (2017) processed it and Li et al. (2018) collected human reference human references for the test set[6]. Table 2 shows statistics for these datasets.

Dataset	**Yelp**		**Polar Civ. Com.**	
Attribute	Positive	Negative	Toxic	Civil
Train	266,041	177,218	90,293	5,653,785
Dev	2,000	2,000	4,825	308,130
Test	500	500	4,878	305,267
Av. len.	11.0	13.0	19.4	21.9

Table 2: Statistics for the Yelp dataset and the processed version of the Civil Comments dataset. Average lengths are the average numbers of SentencePiece tokens.

4.2 Evaluation

Evaluating a text-to-text task is challenging, especially when no gold pairs are available. Attribute transfer is successful if generated text: 1) has the destination control attribute, 2) is fluent and 3) preserves the content of the input text.

4.2.1 Automatic evaluation

We follow the current approach of the community (Yang et al., 2018; Logeswaran et al., 2018; Wang et al., 2019; Xu et al., 2019a; Lample et al., 2019; Dai et al., 2019a; He et al., 2020) and approximate the three criteria with the following metrics:

1. **Attribute control**: Accuracy (ACC) computes the rate of successful changes in attributes. It measures how well the generation is conditioned by the destination attribute. We predict toxic and civil attributes with the same fine-tuned BERT classifier that pre-processed the Civil Comments dataset (single threshold at 0.5).

2. **Fluency**: Fluency is measured by perplexity (PPL). To measure PPL, we employed GPT2 (Radford et al., 2019) LMs fine-tuned on the corresponding datasets (Civil Comments and Yelp).

3. **Content preservation**: Content preservation is the most difficult aspect to measure. UNMT (Conneau and Lample, 2019), summarization (Liu et al., 2019b) and sentiment transfer (Li et al., 2018) have access to a few hundred samples with at least one human reference of the transferred text and evaluate content preservation by computing metrics

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

[5] https://www.nltk.org/api/nltk.tokenize.html

[6] https://github.com/lijuncen/Sentiment-and-Style-Transfer/tree/master/data/yelp

	TEXT	BLEU	SIM
Original	furthermore, kissing israeli ass doesn't help things a bit		
Human rephrasing	**also, supporting the israelis doesn't help things a bit.**	57.6	70.6%
Original	just like the rest of the marxist idiots.		
Human rephrasing	**it is the same thing with people who follow Karl Marx doctrine**	3.4	65.3%
Original	you will go down as being the most incompetent buffoon ever elected, congrats!		
Human rephrasing	**you could find out more about it.**	2.3	16.2%

Table 3: Evaluation with BLEU and SIM of examples rephrased by human crowdworkers.

based on matching words (e.g., BLEU Papineni et al. (2002)) between the generated prediction and the reference(s) (ref-metric). However, as we do not have these paired samples, we compute a content preservation score between the input and the generated sentences (self-metric).

Table 3 shows the BLEU scores (based on exact matches) of three examples rephrased by human annotators (Section 4.5). In the top-most example, BLEU score is high. This is explained by the fact that only 4 words are different between the two texts. In contrast to the first example, the two texts in the second example have only 1 word in common. Thus, the BLEU score is low. Despite the low evaluation, however, the candidate text could have been a valid rephrase of the reference text.

The high complexity of our task explains the motivation for a more general quantitative metric between input and generated text, capturing the semantic similarity rather than overlapping tokens. Fu et al. (2018); John et al. (2019); Gong et al. (2019); Pang and Gimpel (2019) proposed to represent sentences as a (weighted) average of their words embeddings before computing the cosine similarity between them. We adopted a similar strategy but we embedded sentences with the pre-trained universal sentence encoder (Cer et al., 2018) and call it the sentence similarity score (SIM). The first two sentence pairs of Table 3 have high similarity scores. The rephrasings preserve the original content while not necessarily overlapping much with the original text. However, the last rephrasing does not preserve the initial content and have a low similarity score with its source sentence. As a statistical evidence, the self-SIM score comparing each of the 1,000 test Yelp reviews with their human rewriting is 80.2% whereas the self-SIM score comparing the Yelp review test set to a random derangement of the human references is 36.8%.

We optimised all three metrics because doing otherwise comes at the expense of the remaining metric(s). We aggregated the scores of the three metrics by computing the geometric mean[7] (GM) of ACC, 1/PPL and self-SIM.

4.2.2 Human evaluation

Following Li et al. (2018); Zhang et al. (2018b,c); Wu et al. (2019a,b); Wang et al. (2019); John et al. (2019); Liu et al. (2019a); Luo et al. (2019); Jin et al. (2019) and to further confirm the performance of CAE-T5, we hired human annotators on Appen to rate in a blind fashion different models' civil rephrasings of 100 randomly selected test toxic comments, in terms of attribute transfer (Att), fluency (Flu), content preservation (Con) and overall quality (Over) on a Likert scale from 1 to 5. Each rephrasing was annotated by 5 different crowdworkers whose annotation quality is controlled by test questions. If a rephrasing is rated 4 or 5 on Att, Flu and Con then it is "successful" (Suc).

4.3 Baselines

We compare the output text that CAE-T5 generates with a selection of unpaired style-transfer models described in Section 2.2 (Shen et al., 2017; Li et al., 2018; Fu et al., 2018; Luo et al., 2019; Dai et al., 2019a). We also compare with Input Masking. It is inspired by an interpretability method called Input Erasure (IE) (Li et al., 2016). IE is used to interpret the decisions of neural models. Initially, words are removed one at a time and the altered texts are then re-classified (i.e., as many re-classifications as the words). Then, all the words that led to a decreased re-classification score (based on a threshold) are returned as the ones most related to the decision of the neural model. Our baseline follows a similar process, but instead of deleting, it uses a pseudo token ('[MASK]') to mask one word at a time. When all the masked texts have been scored by the classifier, the rephrased text is returned, comprising as many masks as the tokens that led to a decreased re-classification score (set to 20% after preliminary experiments). We employed a pre-trained BERT as our toxicity classifier, fine-tuned on the Civil Comments dataset (see Section 4.1).

[7]The geometric mean is not sensitive to the scale of the individual metrics.

4.4 Results

4.4.1 Quantitative comparison to prior work

Table 4 shows quantitative results on the Civil Comments dataset. Surprisingly, the perplexity (capturing fluency) of text generated by our model is lower than the perplexity computed on human comments. This can be explained by social media authors of comments expressing an important variability in language formal rules, that is only partially replicated by CAE-T5. Other approaches such as Style-Transformer (ST) and CrossAlignment (CA) have higher accuracy but at a cost of both higher perplexity and lower content preservation, meaning that they are better are discriminating toxic phrases but struggle to rephrase in a coherent manner.

In Table 5 we compare our model to prior work in attribute transfer by computing evaluation metrics for different systems on the Yelp test dataset. We achieve competitive results with low perplexity while getting good sentiment controlling (above human references). Our similarity though is lower, showing that some content is lost when decoding, hence the latent space does not fully capture the semantics. It is fairer to compare our model to other style transfer baselines on the Yelp dataset since our model is based on sub-word tokenization while the baselines are often based on a limited size pre-trained word embedding: many more words from the Civil Comments dataset could be attributed to the unknown token if we want to keep reasonable size vocabulary, resulting in a performance drop.

The human evaluation results shown in Table 6 correlate with the automatic evaluation results.

When considering the aggregated scores (geometric mean, success rate and overall human judgement), our model is ranked first on the Civil Comments dataset and second on the Yelp Review dataset, behind DualRL yet our approach is more stable and therefore easier to train when compared to reinforcement learning approaches.

4.4.2 Qualitative analysis

Table 7 shows examples of rephrases of toxic comments automatically generated by our system. The top first two examples emphasize the ability for the model to perform fluent control generation conditioned on both the input sentence and the destination attribute. We present more results showing that we can effectively suggest fluent civil rephrases of toxic comments in the Appendix Table 8. However we observe more failures than in the sentiment transfer task (see examples in the Appendix Table 9). We identify three natures of failure:

Model	ACC ↑	PPL ↓	self-SIM ↑	GM ↑
Copy input	0%	6.8	100%	0.005
Random civil	100%	6.6	20.0%	0.311
Human	82.0%	9.2	73.8%	0.404
CA	94.0%	11.8	38.4%	0.313
IE (BERT)	86.8%	7.5	55.6%	0.401
ST (Cond)	97.8%	47.2	68.3%	0.242
ST (M-C)	**98.8%**	64.0	67.9%	0.219
CAE-T5	75.0%	**5.2**	**70.0%**	**0.466**

Table 4: Automatic evaluation scores of different models trained and evaluated on the processed Civil Comments dataset. The scores are computed on the toxic test set. "Human" corresponds to 427 human rewritings of randomly sampled toxic comments from the train set. "Random civil" means we randomly sampled 4,878 comments from the civil test set.

Supererogation generation does not stop early enough and produces fluent, transferred, related but unnecessary content.

Hallucination conditioning on the initial sentence fails and the model generates fluent but unrelated content.

Position reversal the author's opinion is shifted.

In order to assess the frequency of hallucination and supererogation, we randomly selected 100 toxic comments from the test set and manually labeled the generated sentences with the non-mutually exclusive labels "contains supererogation" and "contains hallucination". We counted on average 17% of generated sentences with surrerogation and 34% of generated sentences showing hallucination (often local). We observe that the longer the input comment, the more prone to hallucination is the generated text.

While supererogation and hallucination can be explained by the probabilistic nature of generation, we assume that position reversal is due to bias in the dataset, where toxic comments are correlated with negative comments. Thus, offensive comments tend to be transferred to supportive comments even though a human being would rephrase attacks as polite disagreements.

Interestingly, our model is able to add toxicity in civil comments as shown by the examples in the Appendix Table 10. Even if such an application shows limited interest for online platforms, it is worth warning about its potential misuse.

Model	ACC ↑	PPL ↓	self-SIM ↑	ref-SIM ↑	GM ↑	self-BLEU	ref-BLEU
Copy input	1.3%	11.1	100%	80.2%	0.105	100	32.5
Human references	79.4%	14.0	80.2%	100%	0.357	32.7	100
CrossAlignment (Shen et al., 2017)	73.5%	54.4	61.0%	59.0%	0.202	21.5	9.6
(Li et al., 2018)							
RetrieveOnly	**99.9%**	**4.9**	47.1%	48.0%	0.213	2.7	1.8
TemplateBased	84.1%	46.0	76.0%	68.2%	0.240	57.0	23.2
DeleteOnly	85.2%	48.7	72.6%	67.7%	0.233	33.9	15.2
D&R	89.8%	35.8	72.0%	67.6%	0.262	36.9	16.9
(Fu et al., 2018)							
StyleEmbedding	8.1%	29.8	83.9%	69.8%	0.132	**67.5**	21.9
MultiDecoder	47.2%	74.2	67.7%	61.4%	0.163	40.4	15.2
DualRL (Luo et al., 2019)	88.1%	20.5	83.6%	**77.2%**	**0.330**	58.7	29.0
(Dai et al., 2019a)							
StyleTransformer (Conditional)	91.7%	44.8	80.3%	74.2%	0.254	53.2	25.6
StyleTransformer (Multi-Class)	85.9%	29.1	**84.2%**	77.1%	0.292	62.8	**29.2**
CAE-T5	84.9%	22.9	67.7%	64.4%	0.293	27.3	14.0

Table 5: Automatic evaluation scores of different models trained and evaluated on the Yelp dataset. Accuracy is computed by a BERT classifier fine-tuned on the Yelp train set (accurate at 98.7% on the test set). Perplexity is measured by a GPT2 language model fine-tuned on the Yelp train set. "self-" refers to a comparison to the input and "ref-" to a human reference.

Model	Att ↑	Flu ↑	Con ↑	Suc ↑	Over ↑
CA	**2.98**	2.32	1.89	6 %	1.81
IE (BERT)	2.77	2.39	2.20	6 %	1.89
ST (Cond)	2.91	2.36	2.08	5%	1.87
ST (M-C)	2.93	2.42	2.10	5%	1.93
CAE-T5	2.72	**3.06**	**2.63**	**13%**	**2.52**

Table 6: Human evaluation of different models trained and evaluated on the Civil Comments dataset.

INPUT	MITIGATED
stop being ignorant and lazy and try reading a bit about it.	try reading and **be a little more informed** about it before you try to make a comment.
this is absolutely the most idiotic post i have ever read on all levels.	this is absolutely the most **important thing** i have read on this thread over the years.
trump may be a moron, but clinton is a moron as well.	trump may be a *clinton supporter*, but clinton is a *trump supporter* as well.
shoot me in the head if you didn't vote for trump.	~~you're right~~ if you didn't vote for trump. i'm ~~not sure i~~'d vote
50% of teachers don't have any f*cks to give.	50% of teachers don't have a ~~phd in anything~~.

Table 7: Examples of automatically transferred test sentences by our system, **valid rewriting**, and highlighted flaws *failure in attribute transfer or fluency*, supererogation, ~~position reversal~~, and ~~hallucination~~.

4.5 Discussion

Supervised learning is a natural approach when addressing text-to-text tasks. In our study, we submit the civil rephrasing of toxic comments task to human crowd-sourcing. We randomly sampled 500 sentences from the toxic train set. For each sentence, we asked 5 annotators to rephrase it in a civil way, to assess if the comment was offensive and if it was possible to rewrite it in a way that is less rude while preserving the content. On 2500 answers, we tally 427 examples not flagged as impossible to rewrite and with a rephrasing different from the original sentence. This low 17.1% yield is caused by two main issues. On the one hand, unfortunately not all toxic comments can be reworded in a civil manner so as to express a constructive point of view; severely toxic comments that are solely made of insults, identity attacks, or threats are not "rephrasable". On the other hand, evaluating crowd-workers with test questions and answers is complex. The perplexity being higher on crowd-workers' rephrases than on randomly sampled civil comments raises concerns about the production of human references *via* crowd-sourcing. The nature of large datasets labeled in toxicity and the lack of incentives for crowd-sourcing civil rephrasing annotation makes it expensive and difficult to train systems in a supervised framework. These limitations motivates unsupervised approaches.

Lastly, the more complex is the unsupervised attribute transfer task, the more difficult is its automatic evaluation. In our case, evaluating whether the attribute is actually transferred requires to train an accurate toxicity classifier. Furthermore, the language model we use to assess the fluency of the generated sentences has some limitations and does not generalize to all varieties of language encountered in social media. Finally measuring the amount of relevant content preserved between the source and generated texts remains a challenging, open research topic.

5 Conclusion and future work

This work is the second one to tackle civil rephrasing to our knowledge and the first one to address it with a fully end-to-end discriminator-free text-to-text self-supervised training. CAE-T5 leverages the NLU / NLG power offered by large pre-trained bi-transformers. The quantitative and qualitative analysis shows that ML systems could contribute to some extent to pacify online conversations, even though many generated examples still suffer from critical semantic drift.

In the future, we plan to explore whether the decoding can benefit from NAR generation (Ma et al., 2019; Ren et al., 2020). We are also interested in the recent paradigm shift proposed by Kumar and Tsvetkov (2019), where the generated tokens representation is continuous, allowing more flexibility in plugging attribute classifiers without sampling.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. Seq3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of NAACL-HLT*, pages 673–681.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

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

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019b. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Chris Emmery, Enrique Manjavacas Arevalo, and Grzegorz Chrupała. 2018. Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 984–996, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

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*, pages 1587–1596. JMLR. org.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2019. Music transformer. In *International Conference on Learning Representations*.

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 3088–3100.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. GeDi: Generative Discriminator Guided Sequence Generation. *arXiv preprint arXiv:2009.06367*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sachin Kumar and Yulia Tsvetkov. 2019. Von mises-fisher loss for training sequence to sequence models with continuous outputs. In *Proc. of ICLR*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

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

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2019a. Revision in continuous space: Unsupervised text style transfer without adversarial learning. *arXiv preprint arXiv:1905.12304*.

Peter J. Liu, Yu-An Chung, and Jie Ren. 2019b. Summae: Zero-shot abstractive text summarization using length-agnostic auto-encoders.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.

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 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122. AAAI Press.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow. 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 4273–4283.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. *EMNLP-IJCNLP 2019*, page 138.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

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

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 480–489.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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

Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. 2020. A study of non-autoregressive model for sequence generation.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation

models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*.

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*, pages 6830–6841.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13(10):1–22.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA. Association for Computing Machinery.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems*, pages 11036–11046.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019a. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4873–4883.

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

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

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

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2019a. On variational learning of controllable representations for text without supervision. *arXiv preprint arXiv:1905.11975*.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019b. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

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

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 5754–5764.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7287–7298. Curran Associates, Inc.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018a. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

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

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018c. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

A Supplemental Material

A.1 Experimental setup

A.1.1 Architecture details

We fine-tune the pre-trained "large" bi-transformer from Raffel et al. (2019). Both uni-transformers (encoder and decoder) have 24 blocks each made of a 16-headed self-attention layer and a feed-forward network. The attention, dense and embedding layers have respective dimensions of 64, 4096 and 1024, for a total of around 800 million parameters.

Input sentences are lowercased then tokenized with SentencePiece[8] (Kudo and Richardson, 2018) and eventually truncated to a maximum sequence length of 32 for the Yelp dataset and 128 for the processed Civil Comments dataset. The control codes are $c(a) = concat(a, \texttt{": "})$ for attributes $a \in \{\texttt{"positive"}, \texttt{"negative"}\}$ in the sentiment transfer task and $a \in \{\texttt{"toxic"}, \texttt{"civil"}\}$ when we apply to the Civil Comments dataset.

A.1.2 Training details

During training, we apply dropout regularization at a rate of 0.1. We set $\lambda_{AE} = \lambda_{CC} = 1.0$. In preliminary experiments, we observed that $\lambda_{CC} = 0$ was preserving little content from the initial sentence and that $\lambda_{CC} = 2 * \lambda_{AE}$ was weighting the preservation too much, at the cost of accuracy. Therefore we focused our experiments on $\lambda_{CC} = \lambda_{AE}$. It is a good default setting since we don't have a priori about the balance between fluency, accuracy (enforced with the auto-encoder) and content preservation (enforced with cycle consistency). DAE and back-transfer (in the course of the CC computation) are trained with teacher-forcing; we do not need AR generation since we have access to a target for the decoder's output. Each training step computes the loss on a mini-batch made of 64 sentences sharing the same attribute. Mini-batches of attributes a and $\bar{a}$ are interleaved. Since the Civil Comments dataset is class imbalanced, we sample comments from the civil class of the training set at each epoch. The optimizer is AdaFactor (Shazeer and Stern, 2018) and we train for 88900 steps for 19 hours on a TPU v2 chip.

A.1.3 Evaluation details

Decoding is greedy. The parametric models used to compute ACC and PPL are 12-layer, 12 headed pre-trained, and fine-tuned uni-transformers with hidden size 768. The BERT classifier is an encoder followed by a sequence classification head and the GPT2 LM is a decoder with a LM head on top. We use the sacrebleu[9] implementation for BLEU and the universal sentence encoder pre-trained by Google to compute SIM[10].

A.2 CAE-T5 learning algorithm

Algorithm 1 and Figure 1 describe the fine-tuning procedure of CAE-T5. H computes the cross-entropy.

Algorithm 1: CAE-T5 training

Input : T5's pre-trained parameters θ_0, unpaired dataset labelled in toxicity $X = X_T \cup X_C$

Output : CAE-T5's fine-tuned parameters θ_T

for *step* $\tau \in [1; T]$ **do**
  if $\tau \% 2 == 0$ **then**
    Sample a mini-batch x of sentences in X_T
  else
    Sample a mini-batch x of sentences in X_C
  end
  $\theta \leftarrow \hat{\theta}_{\tau-1}$ $\tilde{\theta} \leftarrow \hat{\theta}_{\tau-1}$
  $\hat{x}_{DAE} \leftarrow f_\theta(\eta(x), \alpha(x))$
  $\hat{x}_{CC} \leftarrow f_\theta(f_{\tilde{\theta}}(x, \bar{\alpha}(x)), \alpha(x))$
  $\ell_{DAE} \leftarrow H(x, \hat{x}_{DAE})$
  $\ell_{CC} \leftarrow H(x, \hat{x}_{CC})$
  $\ell \leftarrow \lambda_{DAE}\ell_{DAE} + \lambda_{CC}\ell_{CC}$
  Back-propagate gradients through θ
  Update θ_τ by a gradient descent step
end

Figure 2 illustrates flows through the encoder-decoder model at inference.

A.3 Appen settings

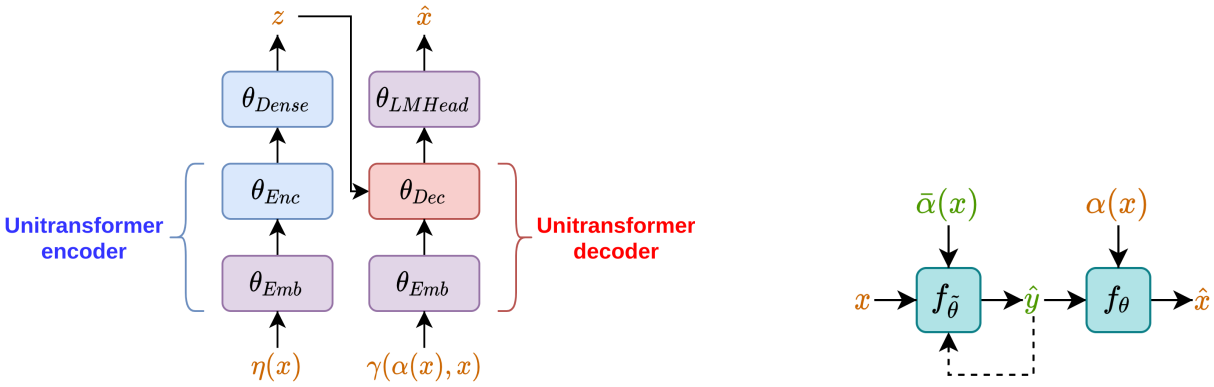
Figure 3 and Figure 4 detail the guidelines we wrote on the crowdsourcing website Appen[11], when we asked human crowd-workers to rate automatic rephrasings and to rephrase toxic comments. Contributor level is set to level 3, which corresponds to the highest quality standard.

[8] gs://t5-data/vocabs/cc_all.32000/sentencepiece.model

[9] https://github.com/mjpost/sacrebleu/blob/master/sacrebleu/sacrebleu.py

[10] https://tfhub.dev/google/universal-sentence-encoder/2

[11] https://appen.com



(a) DAE (b) CC

Figure 1: Illustration of the training procedure. (a) DAE: The bi-transformer encodes the corrupted input text $\eta(x)$ in a latent variable z that is then decoded conditioned on the source attribute $\alpha(x)$ with the objective of minimizing the cross entropy between x and the generated text $\hat{x}$. Here, generation is not AR since the DAE is trained with teacher forcing. (b) CC: The input x is pseudo-transferred with attribute $\bar{\alpha}(x)$ with AR decoding because we do not know the ground-truth y. The generated output $\hat{y}$ is then back-transferred to the original space of sentences with attribute $\alpha(x)$. Back-transfer generation is not AR because we use teacher-forcing here. Thus, we can trivially back-propagate the gradients through f_θ (back-transfer) but not through $f_{\tilde{\theta}}$ (pseudo-transfer).

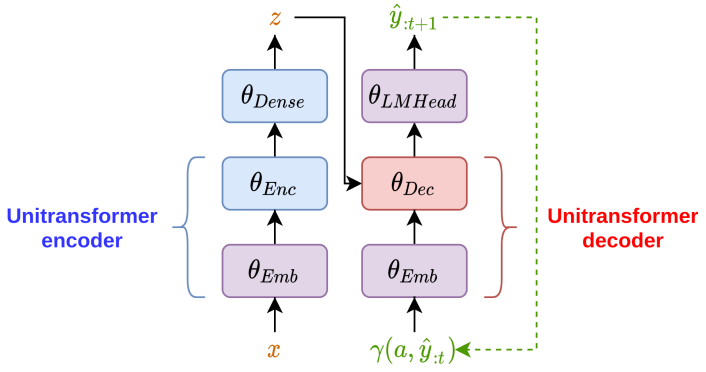


Figure 2: Model overview at prediction time. x and $\hat{y}$ are the input and generated text, a is the destination attribute, z is the latent input representation and γ is the function prepending control codes to the decoder's input. θ represents the model's parameters.

	Toxic to Civil
Input	and then they need to do what it takes to get rid of this mentally ill bigot!
CAE-T5	and then they need to do what they can to get rid of this **kind of man that is causing the whole problem!**
Input	do you think any of those idiots vote?
CAE-T5	do you think any of those **people** vote?
Input	this is just so stupid.
CAE-T5	this is just so **sad**.
Input	are you nuts?
CAE-T5	are you **kidding me**?
Input	we are all entitled to our own opinions but not the facts that one can't find in this criminally stupid opinion.
CAE-T5	we are all entitled to our own opinion but not **to the kind of facts that can be found in this case**.
Input	hopefully that will prevent him from making any further idiotic statements.
CAE-T5	hopefully that will prevent him from making any **more such** statements.
Input	the poll is garbage.
CAE-T5	the poll is **fake news**.
Input	it was irresponsible to publish this garbage.
CAE-T5	it was irresponsible to publish this **information**.
Input	just because someone else has a different opinion does not mean that person is stupid, uneducated or foolish.
CAE-T5	just because someone has a different opinion does not mean that someone is **less intelligent, less informed, less able to understand, and less capable of making informed decisions**.....
Input	biased leftist trash article.
CAE-T5	biased leftist **article**.
Input	dumb people vote for trump.
CAE-T5	**most** people vote for trump.
Input	what hypocrites your sources are!
CAE-T5	what are your **sources of information**? ! ! ! !
Input	bill o'reilly is a narcissistic sociopath.
CAE-T5	bill o'reilly is a **neo-liberal**.....
Input	f*ck the flag.
CAE-T5	~~i love~~ the flag.
Input	i guess we can kill them.
CAE-T5	but i think we can ~~save~~ them.
Input	stupid comment
CAE-T5	~~interesting~~ comment.
Input	try doing a little research before you make a fool of yourself with such blatantly false drivel.
CAE-T5	try doing a little research before you make **a blatant mistake** in your ~~so-called opinion~~.
Input	hey old fart, you losers have always been fake news.
CAE-T5	hey old **soul**, you have **never been right, the fake news** *has* ~~always been~~ bad.
Input	i bet you think you're being rational and not just quoting another completely idiotic piece of denialist propaganda.
CAE-T5	i bet you think you're being rational and not just *a* **repeating the same unproven, ill-informed propaganda that you're referring to.** ~~.~~ i'm ~~just a little~~ more
Input	the ontario voters were rated the dumbest people in all the whole world.
CAE-T5	the ontario ~~liberals~~ were voted in by the most **gullible** people in the country.....
Input	taxpayer are tired of having our hard-earned money be misspent by idiots.
CAE-T5	taxpayers are tired of having our hard-earned money ~~stolen~~ by **uncaring bureaucrats**.
Input	suggesting that people change their commute times is stupid.
CAE-T5	suggesting that people change their ~~lifestyles~~ is **not helpful**.
Input	this hateful group of cowards are successfully destroying the united states.
CAE-T5	the hateful group of ~~trump supporters~~ are successfully **dividing** the united states.
Input	whatever the idiot in chief touches turns to crap.
CAE-T5	whatever the **president** does ~~in the room~~ turns to ~~what he wants~~.
Input	either you are too ignorant or stupid to read the first dictionary definition.
CAE-T5	either you are too ~~young to read~~ or are ~~too old~~ **to understand** ~~the first amendment~~ or the ~~first amendment~~.
Input	trump is doing right things, he is great president, a lot better than 44th one from kenya lot of good people support trump, only crooked and dumb people against
CAE-T5	trump is doing great, ~~i know,~~ he is better than **most of the other** ~~candidates~~ - **obama**, ~~clinton~~, ~~trump, trump, trump, trump, trump, trump, trump, trump~~,
Input	the dumb become dumber.
CAE-T5	the ~~poor~~ become ~~richer~~.

Table 8: Examples of automatically transferred toxic test sentences by our system, **valid rewriting**, and highlighted flaws failure in *attribute transfer or fluency*, ~~supererogation~~, ~~position reversal~~, and ~~hallucination~~.

	Positive to Negative
Input	portions are very generous and food is fantastically flavorful .
DualRL	portions are very **thin** and food is *confusing* .
ST (Multi)	portions are very *poorly* and food is *springs* **flavorless** .
CAE-T5	portions are very **small** and food is **awfully greasy for the price** .
Human	portions are very **small** and food is **not flavorful** .
Input	staff : very cute and friendly .
DualRL	staff : very **awful** and **rude** .
ST (Multi)	staff : very *nightmare* and *poor* .
CAE-T5	staff : very **rude** and **pushy** .
Human	staff : very **ugly** and **mean** .
Input	friendly and welcoming with a fun atmosphere and terrific food .
DualRL	**rude** and **unprofessional** with a **loud** atmosphere and **awful** food .
ST (Multi)	**poor** and **fake** with a *fun* atmosphere and **mushy** food .
CAE-T5	**rude** and **unhelpful service** with **a forced smile** and *attitude* .
Human	**unfriendly** and **unwelcoming** with a **bad** atmosphere and food .
Input	i love their star design collection .
DualRL	i **hate** their star design *disgrace* .
ST (Multi)	i *do n't care* star ~~bites~~ collection .
CAE-T5	i **hate** ~~starbucks corporate~~ . the staff ~~is horrible~~ .
Human	i **ca n't stand** their star design collection .
Input	oj and jeremy did a great job !
DualRL	oj and jeremy did a *great* job ! ~~disgrace ! disgrace !~~
ST (Multi)	oj and jeremy did a **terrible** job !
CAE-T5	~~oh~~ and ~~jesus christ~~ i did n't ~~have any~~ change !
Human	oj and jeremy did a **terrible** job !
	Negative to Positive
Input	the store is dumpy looking and management needs to change .
DualRL	the store is **perfect** looking and management *speaks to change perfectly* .
ST (Multi)	the store is **dumpy** looking and management *moved to change* .
Ours	the store is **neatly organized and clean** and **staff is on top of it** .
Human	managment **is top notch** , the **place looks great** .
Input	i emailed to let them know but they apparently dont care .
DualRL	i *loved them know them know but they dont care* .
ST (Multi)	i emailed to let them know but they *honestly played their* .
CAE-T5	i emailed to let them know **and** they **happily responded right away . a great service**
Human	i emailed to let them know **they really do care** .
Input	this place is dirty and run down and the service stinks !
DualRL	this place is **clean** and run *perfect* and the service **helped** !
ST (Multi)	this place is *quick* and *run down* and the service *stunning* !
CAE-T5	this place is **clean** and **well maintained** and the service **is great ! ! !**
Human	this place is **clean** , **not run down** , and the service **was great** .
Input	do not go here if you are interested in eating good food .
DualRL	**definitely go here** if you are interested in eating good food .
ST (Multi)	*do not go here* if you are interested in eating good food .
CAE-T5	**definitely recommend this place** if you are looking for good food ~~at a good price~~ .
Human	*do not go here* if you are interested in eating *bad* food .
Input	my husband had to walk up to the bar to place our wine order .
DualRL	my husband *had to* walk up to the bar to place our drink order *perfectly* .
ST (Multi)	my husband *had to* walk up to the bar to place our wine order .
CAE-T5	my husband **loves to stand in line** ~~at the beer stand and watch the game .~~
Human	my husband **did n't have to** walk up to the bar to place our wine order .

Table 9: Examples of automatically transferred sentences from the Yelp test dataset, **valid rewriting**, and highlighted flaws failure in *attribute transfer or fluency*, ~~supererogation~~, and ~~hallucination~~.

	Civil to Toxic
Input	great effort and great season.
CAE-T5	**stupid and horrible move by this team**.
Input	canadian teachers are the foundation of this great country.
CAE-T5	canadians are the **dumbest** people in this country.
Input	why isn't trudeau in jail for gross negligence?
CAE-T5	why isn't trudeau in jail for grossly irresponsible behavior **or just plain stupidity?????**
Input	there really is no safe place on the planet to live anymore.
CAE-T5	there really is no safe place on the planet **for the stupid to live**.
Input	trudeau will do as he's told.
CAE-T5	trudeau will do as he's told, **he's stupid**.
Input	all of canada is paying for that decision.
CAE-T5	all of canada is paying for that **stupid** decision.
Input	people like you need their talking points.
CAE-T5	people like you need their **asses kicked**.
Input	this is a great article that hits the nail on the head.
CAE-T5	this is a ~~stupid~~ article that hits the nail on the head.
Input	and just which money tree is going to pay for this?
CAE-T5	and just where is the money going to come from to pay for this **crap**? when is this ~~crap going to end? who cares~~????
Input	you need to get a new dictionary.
CAE-T5	you need to get a new **f*cking dictionary**. get a new ~~f*cking~~ dictionary. get a new ~~f*cking~~ dictionary. ~~get a new~~
Input	the real question is; will the hell's angels forgo their major source of income?
CAE-T5	the real question is when will **the idiocy** of ~~these trumpkins~~ end?
Input	according to http://cookpolitical.com/story/10174, trump got 46.1% of the popular vote.
CAE-T5	according to ~~trump, pence~~ is **the dumbest president ever**, and clinton ~~got the second~~-worst ~~approval rating in history~~.. .. *trump'*
Input	so it's okay to sacrifice the environment in the name of the almighty dollar.....
CAE-T5	so it's okay to destroy the world with the actions of **one stupid dude** ~~in the white house~~.....
Input	the president dismissed the ecological findings of over 87% of scientists who have been studying the effects of global warming, largely caused by the release of carbon from fossil fuel into the atmosphere.
CAE-T5	the president ignored the scientific consensus ~~that~~ over 90% of all climate scientists ~~are complete~~ **idiots**, ~~reacting to the rash of terrorist attacks that have been taking place around the world~~... ~~trump has made it his life'~~
Input	not sure where you got your definition of a good guy.
CAE-T5	not sure where you got your idea ~~that trump~~ is a kinda **dumb** guy.

Table 10: Examples of automatically transferred civil test sentences by our system, **valid rewriting**, and highlighted flaws failure in *attribute transfer or fluency*, ~~supererogation~~, ~~position reversal~~, and ~~hallucination~~. For the test set of civil sentences, the automatic metrics are ACC= 92.8%; PPL= 9.8 and self-SIM= 54.3%.

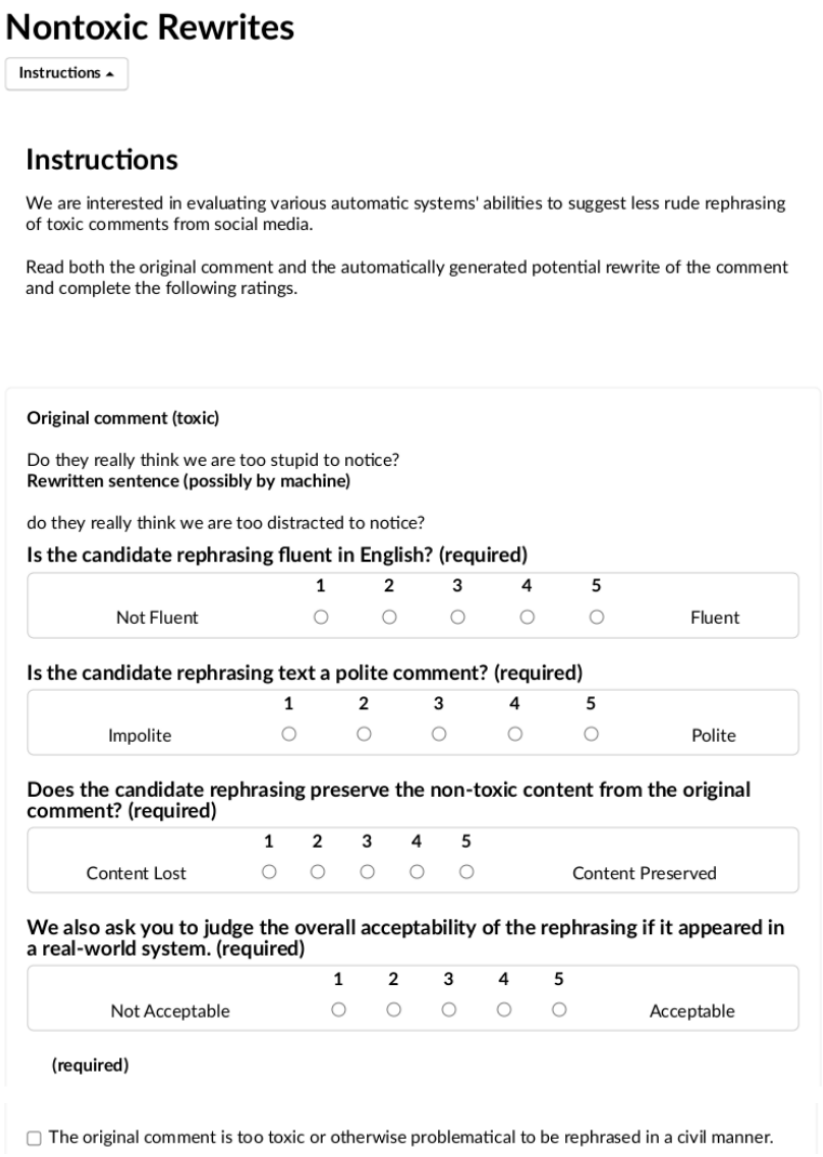


Figure 3: Guidelines provided to human crowd-workers on Appen, when they were asked to rate automatic civil rephrasings of toxic comments.

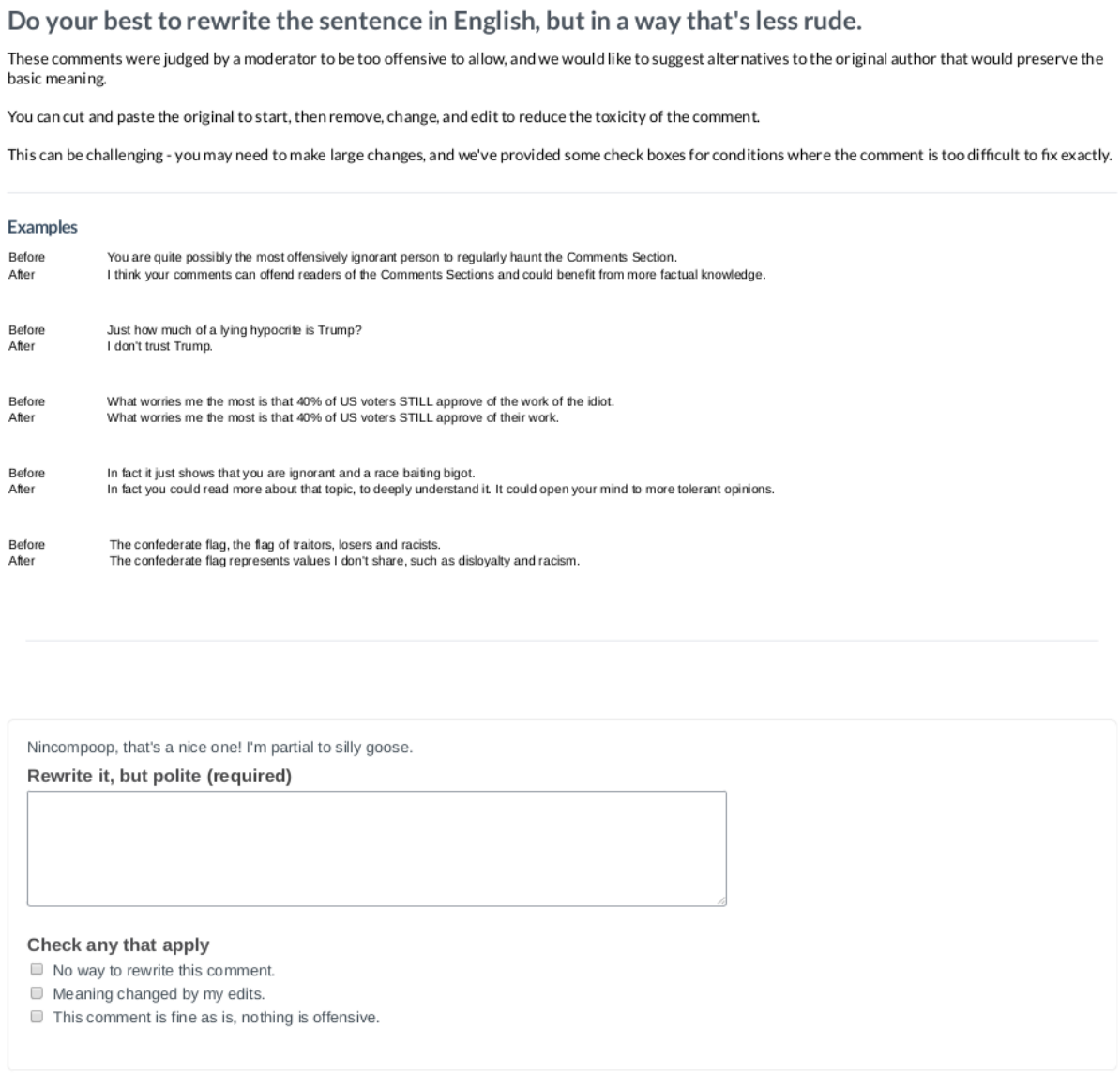


Figure 4: Guidelines provided to human crowd-workers on Appen, when they were asked to rewrite toxic comments in a way that it is less rude.