Lost in Back-Translation: Emotion Preservation in Neural Machine Translation

Enrica Troiano, Roman Klinger, and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Machine translation provides powerful methods to convert text between languages, and is therefore a technology enabling a multilingual world. An important part of communication, however, takes place at the *non-propositional* level (e.g., politeness, formality, emotions), and it is far from clear whether current MT methods properly translate this information.

This paper investigates the specific hypothesis that the non-propositional level of *emotions* is at least partially lost in MT. We carry out a number of experiments in a back-translation setup and establish that (1) emotions are indeed partially lost during translation; (2) this tendency can be reversed almost completely with a simple re-ranking approach informed by an emotion classifier, taking advantage of diversity in the n-best list; (3) the re-ranking approach can also be applied to *change* emotions, obtaining a model for *emotion style transfer*. An in-depth qualitative analysis reveals that there are recurring linguistic changes through which emotions are toned down or amplified, such as change of modality.

1 Introduction

The quality of machine translation (MT) models in some areas follows close behind that of humans (Barrault et al., 2019). MT is deployed widely to support human-to-human communication across languages, e.g., in chat systems, customer support, or (social) media. It is also employed in downstream NLP tasks such as sentence simplification (Xu et al., 2016), error correction (Yuan and Briscoe, 2016), paraphrasing (Mallinson et al., 2017; Wieting and Gimpel, 2018), or cross-lingual resource creation (Barnes and Klinger, 2019). With the increasing use of MT, however, expectations about output quality also grow, and now that the goal of adequacy with regard to propositional content is met more often than not, more subtle aspects start receiving attention. One such aspect is *affective content*. Establishing common ground is essential for successful MT-assisted communication (Yamashita et al., 2009), but it is still unclear how well MT promotes this, especially when handling the affective qualities of texts. On the one hand, it is able to mostly preserve author sentiment (Balahur and Turchi, 2012). On the other, it is known that translation obfuscates some socio-demographic characteristics of authors, like gender and personality traits (Mirkin et al., 2015; Rabinovich et al., 2017).

In this paper, we investigate the question of *how well emotions are preserved in MT*. Answering this question and, if necessary, increasing the degree of emotion preservation, is important both theoretically (to inform cross-lingual studies that use translation as part of their experimental setup) and practically (to improve the usefulness of MT). The starting point of our research is a study by Rabinovich et al. (2017), who show that some semantic nuances tend to vanish in translation. In fact, just like human translation, MT is not guaranteed to preserve any of the linguistic properties of input texts (e.g., politeness markers may exist in one language but not in the other, passive sentences may be turned into active, metaphoric expressions can be rendered by a more literal paraphrase). Therefore, to counteract the fading of emotions through translation, we establish *emotion-based translation candidate re-ranking* that is

Proceedings of the 28th International Conference on Computational Linguistics, pages 4340–4354
Barcelona, Spain (Online), December 8-13, 2020

Research Question	Sentence
(Input)	***He was furious at the apparent disregard for rules.***
RQ1: Does MT dilute emotions?	He was *worried* at the apparent disregard for rules.
RQ2: Can we recovered the original emotion?	He was *quite enraged* at the inattention to rules
RQ3: Can we change the emotion?	He was *unhappy* that the rules were ignored.

Table 1: Illustration of three emotion-related research questions about MT, with examples for the associated tasks to be solved.

applied as post-processing to an MT system's n-best output. This re-ranking can be defined, for example, on the basis of a standard emotion classifier with probabilistic output, to select a candidate such that its emotional connotation is as close as possible to the input. We investigate whether such an approach is feasible and promising.

To carry out this re-ranking in practice, we would need comparable emotion classifiers for the source and target languages. We avoid this issue by adopting a *back-translation* setup (Mallinson et al., 2017): instead of analyzing the translations automatically obtained by a system performing `source→target`, we consider the output of a back-translation pipeline `source→target→source`, which we can examine with only one emotion classifier for the source language. We acknowledge that this solution makes a simplifying assumption, namely that experimenting with back-translation can give a realistic picture of what would happen in a `source→target` setting. Still, adding a translation step seems a reasonable compromise in the absence of comparable emotion classifiers for different languages.

Within this framework, we address three research questions (Table 1 shows motivating examples). We first ask if a state-of-the-art machine translation system, namely FAIRSEQ (Ott et al., 2019), loses emotional information during translation (**RQ1**). (Yes.) Next, we propose a post-processing step to re-rank n-best translation candidates and evaluate if this improves emotion preservation (**RQ2**). (It does.) Finally, we exploit the emotional variation in MT output to investigate whether this approach can actively change the input emotion (**RQ3**), essentially performing emotion style transfer (Helbig et al., 2020). (It can.) The implementation of the pipeline is available at `http://www.ims.uni-stuttgart.de/data/emotion-transfer`.

2 Related Work

Affect, Sentiment, and Emotion in Translation. Preserving affect in text is an issue for translation and other cross-linguistic studies (Wierzbicka, 2013; Wassmann, 2017; Hubscher-Davidson, 2017). On the one hand, there are linguistic constraints on translation, like the absence of terms for certain states (e.g., *Sehnsucht* is German for "a longing for some absent thing") or colexification phenomena (i.e., naming related emotions with the same word, like *grief* and *regret* in Persian) which vary from language to language (Jackson et al., 2019). On the other hand, aesthetic considerations often call for making texts more readable or pleasant. These factors hamper methods that transfer affect or sentiment across languages, as they cause both translation errors (for human and machines alike) and stylistic choices which subvert the sentiment of words (Petrova and Rodionova, 2016). Thus, assessing the quality of sentiment-annotated resources produced by translation (Banea et al., 2008; Chen and Skiena, 2014; Buechel et al., 2020, i.a.) is crucial. With this goal, Kajava et al. (2020) compare sentiment and emotion annotations of movie subtitles in English, Finnish, Italian, and French and find that the emotion preservation depends on the language pair. Validating resources for Romanian created from English ones, Mihalcea et al. (2007) notice that human translation can obscure the subjectivity of a lexicon. A comparable observation is drawn for polarity by Balahur and Turchi (2012) with SMT, and by Salameh et al. (2015) and Mohammad et al. (2016) who find that translation can corrupt textual sentiment, flattening positive and negative aspects down to neutrality.

In MT research, some studies specifically try to incentivize the preservation of sentiment. Lohar et al. (2017) build separate translation models for data coming from each sentiment category. Si et al. (2019)

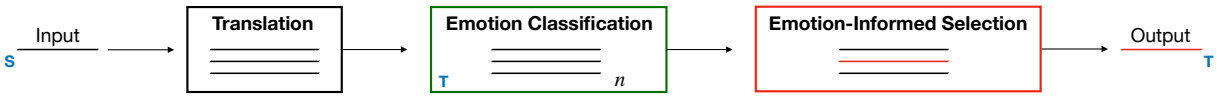

Figure 1: Overview of emotion preservation and transfer (method).

directly incorporate sentiment in their neural MT system, implementing a Seq2Seq English-to-Chinese translation model that keeps not only the semantics but also the sentiment of input text, both by including the sentiment label in source sentences, and by learning the negative/positive meanings of the ambiguous word as separate embeddings.

While these studies gained some insight on translated polarity, subjectivity, valence, dominance and arousal, to our knowledge there are no studies that investigate specifically the preservation of emotions in MT.

Style Transfer. Related to ours is the task of *style transfer*. This research direction leverages a variety of methods, from rule-based lexical substitution to sophisticated neural architectures, aiming at retaining the semantics of texts while modifying their linguistic properties, like genre (Lee et al., 2019; Jhamtani et al., 2017), romanticism (Li et al., 2018), politeness/offensiveness and formality (Sennrich et al., 2016; Nogueira dos Santos et al., 2018; Wang et al., 2019) and, importantly for us, affect-related attributes (Guerini et al., 2008; Whitehead and Cavedon, 2010; Shen et al., 2017; Fu et al., 2018; Xu et al., 2018; Smith et al., 2019; Helbig et al., 2020). Yet, only a handful of style transfer studies have considered emotions. Helbig et al. (2020), for instance, propose an interpretable framework based on lexical substitution which sequentially determines the portion of text to modify, performs the change, and filters out undesired output. Smith et al. (2019), instead, leverage a denoising auto-encoder and a back-translation objective to push the text generated during decoding towards a specific target attribute.

The style transfer challenge is to create a fluent output that is semantically similar to the input, but differs systematically in style. Helbig et al. (2020) control for the balance between content, style and fluency with a dedicated component in their modular pipeline: after a text is re-written in many emotion variations, these are re-ranked by an objective function that measures their perplexity, preservation of content and expression of a target style. Evaluation metrics for these three desiderata are applied in the reinforcement learning approach of Gong et al. (2019) to impose constraints on output generation. Other attempts focus on the explicit separation between content and sentiment style (Li et al., 2018; Wen et al., 2020). Prabhumoye et al. (2018) do so using neural back-translation: in the latent representation of an input text, its stylistic properties are overwritten, which results in a style-specific paraphrase.

Like them, we tap on back-translation as a paraphrasing strategy, but we transfer emotions, which we conceptualize as fine-grained styles. Using state-of-the-art off-the-shelf systems, we move from the problem of guaranteeing fluency and similarity to an input. In line with ours, a few other works have attempted to generate emotionally loaded text for given emotion classes, for instance in dialogue systems (Song et al., 2019; Zhou and Wang, 2018), but they create novel texts rather than re-styling existing ones.

3 A Method for Emotion Preservation in Neural Machine Translation

We conceptualize emotion preservation in NMT as a post-processing re-ranking step. As shown in Figure 1, this involves three components: a translation model, an emotion classifier, and a candidate selection procedure. Starting from an input in source language S, we generate the n-best translation candidates in a target language T with an NMT system, which is presumably agnostic to emotion-specific considerations. Then, we re-rank these candidates based on probabilities produced by an emotion classifier, and select the best hypothesis given those emotion-level considerations. Hence, the crucial variable is the *diversity* of the n-best list: the more diverse, the better the emotion classifier can promote hypotheses that express particular emotions even if they are not optimal from the point of view of the overall scoring function of the NMT system.

Translation model. We require a translation model that returns a list of n-best translation candidates, which is the case for essentially every statistical or neural MT system. We use FAIRSEQ (Ott et al., 2019),

an open-source sequence-to-sequence modeling toolkit applicable to various tasks, MT included. It shows state-of-the-art performance and it was developed with the goal to replicate other model architectures. Therefore, we assume that it is reasonably representative for other models.

Importantly, FAIRSEQ supports different search algorithms, like beamsearch and top-k sampling, which differ in their ability to encourage diversity in the output. Beamsearch searches the space of hypotheses left-to-right, retaining at each time step a number of top-scoring candidates that equals the width of the beam, and expanding on those. Sequences decoded with beamsearch differ on minimal portions (Gimpel et al., 2013), while they are more varied when generated with sampling strategies. Top-k sampling, for instance, does not aim at maximizing the likelihood of text. Instead, it randomly samples words step-wise and outputs from the top-k most probable ones (Fan et al., 2018).

Emotion Classification Model. To estimate the probability distribution over emotions for a given text, we use a biLSTM with a self-attention mechanism. This model architecture has been shown to perform close to state-of-the-art in emotion analysis (Baziotis et al., 2018). We treat the output of this emotion classifier as a scoring function $\text{emo}(t, e) = p(e|t)$, i.e., the conditional probability of an emotion given a text t, and we assume that it is comparable across languages (see Section 4 for a discussion of this assumption).

Translation Candidate Selection. Once the n translation candidates (called $\mathrm{hypotheses}$ in Equations 1 and 2 below) are scored by the emotion classifier, we re-rank them based on their probability for specific emotions, and select a top candidate based on our research question.

The setup described above permits us to address our three different research questions (RQs). In **RQ1**, where we only consider a single translation hypothesis, the output selected by emotion selection is trivially the one coming out of the translation — it is picked based on properties of a standard translation procedure. For **RQ2**, we preserve the dominant emotion of the input by selecting the output such that

$$\texttt{output} = \underset{c \in \text{hypotheses}(\texttt{input})}{\arg\min} |\text{emo}(c, \hat{e}) - \text{emo}(\texttt{input}, \hat{e})| \quad \text{where } \hat{e} = \underset{e \in \text{Emotions}}{\arg\max}\, \text{emo}(\texttt{input}, e)\,. \quad (1)$$

Measuring the absolute difference in emotion load for two texts is similar to Luo et al. (2019), who analyze the change in sentiment intensity with mean absolute errors.

Finally, in **RQ3**, where we aim at maximizing some user-chosen emotion e', we define

$$\texttt{output} = \underset{c \in \text{hypotheses}(\texttt{input})}{\arg\max} \text{emo}(c, e')\,. \quad (2)$$

Our method does not condition the MT system towards a specific emotion. Instead, we evaluate the extent to which the n-best lists of a state-of-the-art MT system contain sufficient variation in their candidates as to manipulate the emotional load of a translation – either by optimizing preservation of the input emotion (RQ2) or by changing the emotion connotation (RQ3).

4 Experimental Setup: Back-translation

The most natural setup to study emotion preservation in translation, and the framework outlined in the previous section, would be bilingual: analyzing the translation of some source language text into a target language. For instance, one could compare the distribution of emotion probabilities for a translation against the corresponding distribution for the source text. However, a meaningful cross-lingual comparison of emotion probabilities is methodologically challenging: this would require either manual annotation or highly comparable emotion classifiers for several languages. Manual annotations are costly, and emotion annotation is known to be tricky in terms of intersubjective replicability (Schuff et al., 2017). Neither are we aware of emotion classifiers with evidently similar behavior across languages.

To circumvent this problem, our experimental setup uses a *back-translation version* of the method described above, shown in Figure 2. We compose two translation steps (S→T and T→S) such that the output is a *paraphrase* of the input in the same language S (Bannard and Callison-Burch, 2005) and the pitfalls of cross-lingual comparability can be avoided. Formally, given an input in S and a target

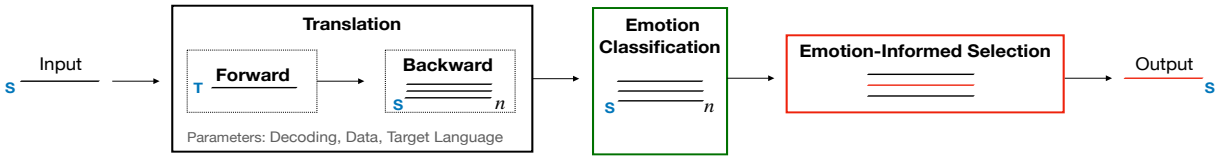

Figure 2: Instantiation of the method with back-translation. S is the source language; T is the target.

emotion, we generate the best translation in T; this is then translated back into multiple hypotheses, providing a set of n paraphrases for the original input. We acknowledge that this type of setup is a conceptual simplification of the problem, which does not measure the loss of emotion in one direction, nor accounts for *where* the change in emotion occurs. As a matter of facts, it risks overestimating the *absolute* magnitude of problems in emotion preservation in MT, but this is a price to pay for the usage of our monolingual emotion classifiers. On the other hand, we can still *compare* the magnitude of emotion loss across different MT settings (which we do in the sections below). In addition, results that would indicate that we can *improve* emotion preservation in back-translation would conversely be stronger than such results obtained on a single translation step.

4.1 Experimental Setup Details

Following the considerations of the previous paragraph, we do not run a single experiment, but instead carry out a series of comparisons, varying the different parameters of the emotion preservation method.

NMT Model: Varying target language and sampling method. We use FAIRSEQ with English–German and English–Russian models[1] (Ng et al., 2019). These sentence-level models are based on transformers (Vaswani et al., 2017) and pretrained on bitext and back-translated news data, fine-tuned on in-domain data and used for decoding with a noisy channel approach to re-rank the n-best hypotheses. We use these models both with beamsearch and top-k sampling (cf. Section 3).

Data Sets: Varying Emotion Realization. Emotions manifest themselves in various linguistic realizations, for instance with direct mentions (*sad*) or indirect associations (*abandoned*). These realizations differ widely across domains and genres (Bostan and Klinger, 2018). To gain a representative picture and investigate the effect of translation on different emotion realizations, we compare four English corpora. **ISEAR** (Scherer and Wallbott, 1994) includes ≈7k descriptions of events. Each description is labeled with the emotion that it induced in the experiencers (*anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness* and *shame*). **TEC** (Mohammad, 2012) contains ≈ 21k tweets associated to the six fundamental Ekman's emotions (Ekman, 1992). The corpora by Aman and Szpakowicz (2007) and by Alm et al. (2005) are repertoires of ≈5k and ≈15k sentences from a number of **Blogs** and (fairy-)**Tales**, respectively (using Ekman+*noemo*). These corpora differ in labels (see Figure 3 vs. 4), topics, registers and communicative purposes: TEC collects short, spontaneous expressions, ISEAR provides statements that were produced in-lab.

	Micro F_1			
Em.	ISEAR	Blogs	Tales	TEC
A	.51	.55	.39	.37
D	.58	.64	.12	.26
F	.70	.56	.33	.55
G	.55	—	—	—
J	.72	.69	.45	.69
No	—	.88	.79	—
Sa	.61	.49	.37	.45
Su	—	.41	.27	.49
Sh	.46	—	—	—

Table 2: Classification results ("—": the emotion is not a label in the respective corpus).

Emotion Classifier. Due to these differences in linguistic realization among corpora, emotion classifiers generalize badly (Bostan and Klinger, 2018). To avoid this problem, we re-train our emotion classifier (cf. Section 3) for each dataset. We train the model on 70% of the instances (cf. Section 4), validating it on the 10% and using the remaining 20% to evaluate our emotion preservation method. We use 300-dimensional GloVe embeddings (Pennington et al., 2014); for regularization, we use Gaussian noise, a dropout rate of

[1] https://github.com/pytorch/fairseq/tree/master/examples/wmt19

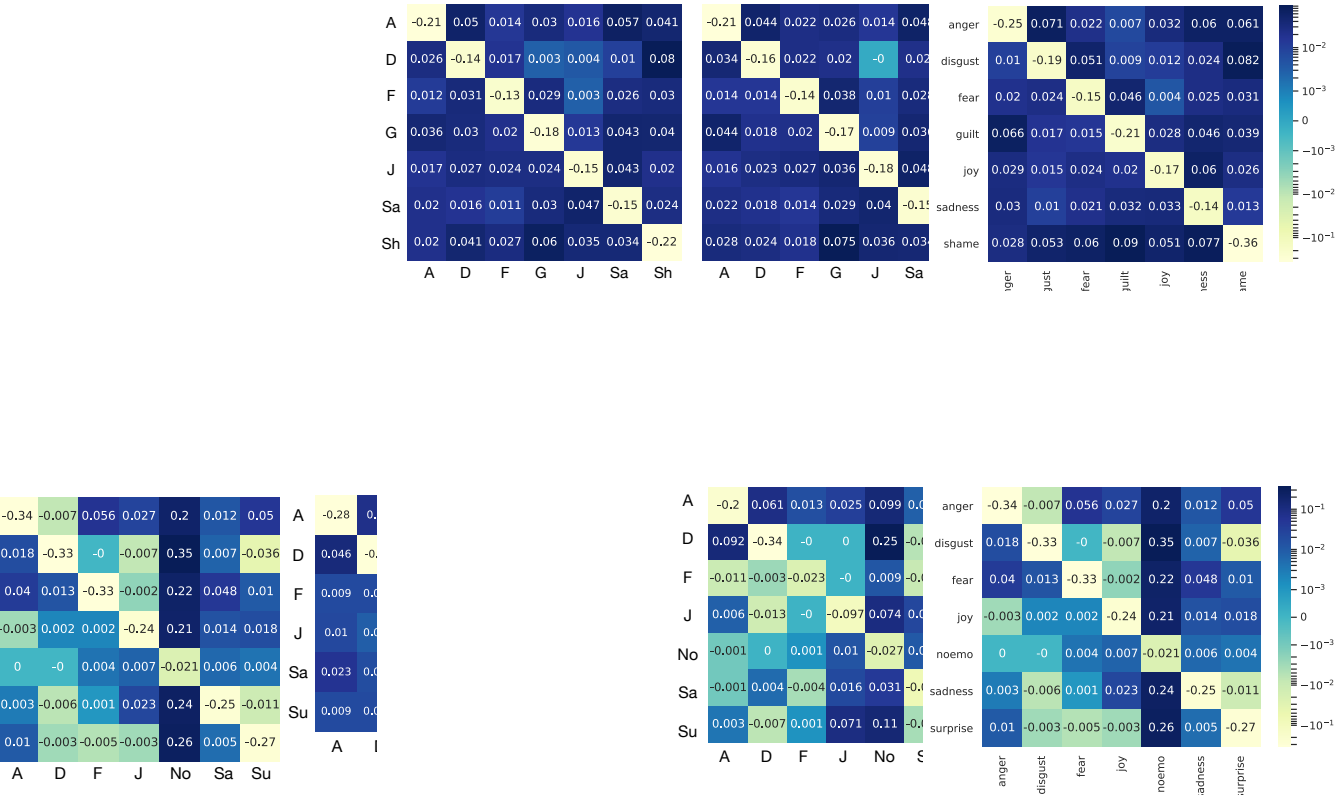

Figure 4: RQ1: Emotion loss on other corpora, using beamsearch for decoding and En↔De as language pairs. Rows are input emotions, columns are the output emotions (No: no emotion, Su: surprise).

0.1, and early stopping. Table 2 shows that performance on the various corpora is comparable to previous work on the same setup (Bostan and Klinger, 2018).

Evaluation. For evaluation, we re-use the emotion scores employed in candidate ranking. Our basic measure is again based on probability differences with regarding to a specific emotion e in a set S of input–output pairs:

$$\Delta(S, e) = \frac{1}{|S|} \sum_{(s_1, s_2) \in S} \text{emo}(s_1, e) - \text{emo}(s_2, e). \quad (3)$$

For RQ1, S is the set of inputs and their 1-best backtranslations. For RQ2, S is the set of inputs and their backtranslations as selected by Eq. (1) for each input emotion. In RQ3, S is the set of inputs and their backtranslations as selected by Eq. (2) for each emotion.

We acknowledge that using the emotion classifier both for ranking and evaluation introduces a potential circularity. To avoid this problem, the reliability of the classifier is crucial. We therefore carry out a detailed qualitative inspection of examples (Sec. 5.4) to gauge the classifier output with our linguistic judgment.

5 Results

5.1 RQ1: Does translation preserve the emotion connotation of texts?

We first present results of our three research questions, then provide a qualitative discussion. To begin with, we turn to the question if the off-the-shelf system FAIRSEQ indeed reduces emotion connotations.

This analysis is purely based on the $n = 1$ best output from the translation system, which we compare to the original input. Figure 3 and 4 show the Δ values between the input and output emotion probabilities. Each cell in the heatmaps contains the average difference between the group of texts that are associated

with the emotion on the row (as determined by our emotion classifier) and their backtranslations. For instance, the first row informs us about the extent to which emotions change when texts expressing predominantly anger are back-translated (probability is reduced by an average of 21%, while it increases a bit for all the other emotions). Hence, the expectation that the backtranslations have a lower emotional score for the emotion characterizing the input should reflect on the diagonal, which reports the Δ values between the emotion identified by the classifier in an input text and the same emotion as measured in its backtranslation.

In order to establish what patterns have generally validity, we vary three parameters (cf. Section 4 for details), namely the *data set* (ISEAR, TEC, Tales, Blogs) – to measure the influence of domain and annotation procedure, the *language* (from English to German and from English to Russian), and the *decoding strategy*, comparing beamsearch, which is more conservative, to sampling, which generates more diverse results.

Varying Decoding Method and Target Language. We analyze decoding method and target language on ISEAR. Figure 3 reports the results obtained when using beamsearch (a) against sampling (b) and German (a) against Russian (c). There is no significant difference between German and Russian (p=.23, Mann Whitney U test), nor between decoding methods (p=.76). Hence, we conclude that the ability of translations to preserve emotion is unrelated both to the target/pivot language, as well as to the generation strategies we employed.

The values on the diagonals, indicating a general loss of the dominant emotion in the input, are of the lowest magnitude and negative. The backtranslations of inputs expressing *anger* and *shame* are those with the greatest loss in those same emotions ($-.21$ and $-.22$, respectively), followed by *guilt* ($-.18$), *joy* and *sadness* ($-.15$), *disgust* and, lastly, *fear* ($-.14$ and $-.13$). Off-diagonal cells, instead, are positive, with the exception of the degree of *joy* in items originally containing *disgust* when the decoding is sampling. In the three cases, the highest increases are recorded for the instances originally labeled as *disgust*, which increase in their *shame* scores, and for the *shame* examples, whose amount of *guilt* is scaled up. Overall, this means that (back)translations express the original emotion to a lower extent than the input, and the decrease of the original emotion is balanced out by an increase of the others, confirming our hypothesis.

Varying Corpora. Given the non-significant difference between the parameters we tested, we continue our experiments fixing the decoding method and language pair to beamsearch and En$\leftrightarrow$De, and investigating if we can generalise our observations to datasets other than ISEAR. The results, which are reported in Figure 4, suggest that the loss of original emotions visible in the diagonal is a persistent trend across corpora, together with the fact that original emotions are toned down more than any other.

We observe that the emotion change on TEC is the most similar to ISEAR, despite the difference in their labels. Further, interesting observations include the amount of *anger* gained by the translations of text classified as *disgust* in Figure 3 (a) vs. Figure 4 (b). This could be an effect of the presence of the label *noemo*, which does not exist in ISEAR. It is also interesting to notice that translations of Blogs and Tales tend to increase in neutrality more than in other emotions. Exceptions are translations that were already classified as containing *no emotion*, and which lose their neutral status (see cell *noemo-noemo* in the diagonal of both (b) and (c), Figure 4).

5.2 RQ2: Can an emotion-informed translation selection restore the original emotion?

We now evaluate our emotion-informed post-processing. Figure 5 (a) reports the results on ISEAR with beamsearch: for an input, we obtain its forward translation, and $n = 50$ backtranslations; among them, we pick the one minimizing the Δ with the input emotion, following Eq. (1). Like before, the emotions on the rows are expressed by the input text. Columns are those for which the delta is computed between the output and input. For instance, the cell A-D shows the average Δ between the *disgust* score of the texts classified as *anger*, and the *disgust* score in their backtranslations.

What interest us is the diagonal, showing the average differences between the original emotion and that emotion as expressed by the output. Once more these values are negative, indicating that at least for some texts, the translation with the closest emotion to the original one still has less of that emotion. As we

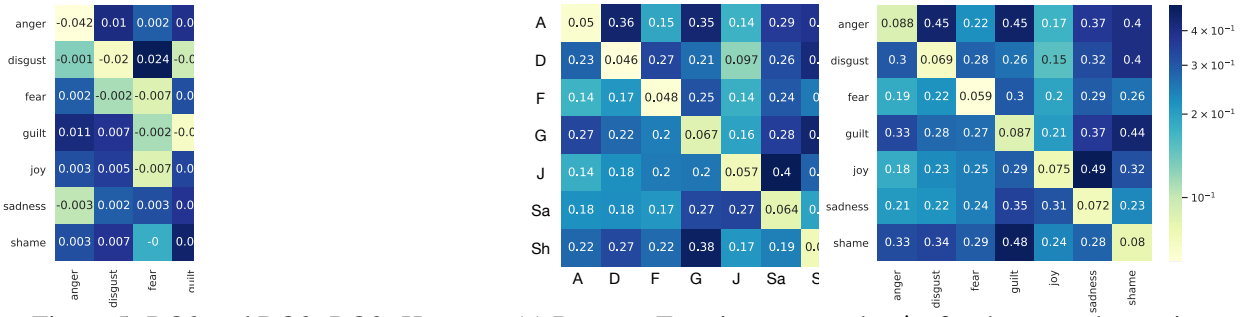

Figure 5: RQ2 and RQ3. RQ2: Heatmap (a) Recover Emotion reports the Δs for the second experiment. RQ3: Heatmaps (b) and (c) report the Δs for the third. In both cases, the dataset is ISEAR, input emotions are on the rows, columns are target emotions. See Figure 3 for emotion abbreviations.

minimize the deltas, values close to 0 indicate success. Most are actually close; the cells that depart from 0 the most are A-A, Sh-Sh and G-G, with $\Delta = -.042, -.042$ and $-.022$. In a comparison to Figure 3 (a), we see that indeed we can recover emotions. The loss of *anger* (A-A) is 5 times smaller than it was when exploiting the 1-best backtranslation; likewise, *sadness* (Sa-Sa) is preserved $\approx$21 times more. These numbers suggest that the behavior of NMT tools can be improved with the n-best lists produced by the systems themselves, since these hypotheses provide enough information to preserve emotions.

As a last sanity check, we investigate if descending the n-best list in the search of an emotionally adequate translation has an impact on its translation adequacy. To do so, we compute the BLEU-4 score for the top outputs returned by the system (i.e., those analysed in RQ1) and our emotion-preserving backtranslations, and we compare their averages. Translation quality remains stable: in the first case we obtain .49 BLEU, in the latter we find a BLEU of .51. This indicates that it is possible to find an emotion-preserving variation further down the space of candidate outputs (at least to a certain point) without sacrificing the system's performance.

5.3 RQ3: Can we exploit overgeneration to transfer a target emotion on a text?

Having shown that MT prefers to output sentences with a toned-down emotion, and that it is possible to subselect instances with a similar emotional connotation as the input, we now turn to the question if diversity in MT output can be used for the task of emotion style transfer. In this setting, our backtranslation pipeline is used for paraphrasing with style transfer, following Xu et al. (2012) and Prabhumoye et al. (2018). Given an input text t and an emotion e, we want to produce a variation t' which respects the following desiderata (Mir et al., 2019): it maximizes similarity with t; it is fluent and it expresses emotion e. Backtranslations provide us with a particularly easy setup: since MT systems are trained to maximize the fluency of their output and their faithfulness to the input, we assume that it is sufficient to pay attention to the presence of the target emotion (see Eq. (1)). Forward and backward translation steps alike are carried on through beamsearch or top-k sampling, with k=10, both producing n=50 paraphrases.

Since this experiment tries to promote stylistic diversity, n-best lists could have been leveraged also in the target language[2]. However, our aim is also to limit the artefacts introduced by our usage of backtranslation: we assumed that employing only one forward translation could approximate a more realistic setting, in which the mapping between source and target occurs in a single step.

Varying Decoding. Figure 5 shows the results on ISEAR with beamsearch (b) and sampling (c). Each cell reports the average Δ of all instances for a pair of input (rows) and target (columns) emotions. It quantifies the intensity of the transfer, or how much more of a target emotion is present in the selected backtranslation. The first row in (b), for example, considers the backtranslations of texts expressing *anger*: those on which *anger* itself was transferred (i.e., those selected as having the highest degree of *anger*) express that emotion .05 points more than their original counterparts; *disgust* is .36 points higher than

[2]We actually experimented with n-best translations in the target language both for RQ2 and RQ3, obtaining results similar to those we report here.

before in those backtranslations where *disgust* was the target emotion.

As expected, the diagonal has the lowest numbers in both matrices, since it corresponds to target emotions that were already there in the first place. Yet, there is quite a substantial improvement overall, indicating that our method can be used for emotion transfer. The highest Δs are mainly among pairs of negative emotions. We also notice that it is easier to transfer *joy* onto negative emotions than the other way around (see column *joy*, which has some of smallest off-diagonal values). In line with the fact that emotions are not binary, this suggests there are interdependencies between the source text emotion and the desired target emotion.

In both the beamsearch and sampling cases, the strength of transfer depends on input and target emotions. Successful transfers take place for sentences originally labeled as *joy* that are re-phrased as *sadness*. Given *shame*, *guilt* can be increased to a considerable extent, as can *shame* given *guilt*, which is an interesting symmetry because these two emotions are attributed to the self (Tracy and Robins, 2006). Other than these similarities, however, we find a significant difference between the two matrices ($p=1.11 \cdot 10^{-09}$, Mann Whitney U test). The higher numbers in (c) corroborate the idea that sampling efficiently induces diversity in the n-best outputs. Also, emotion diversity in the translations can be variously achieved considering hypothesis space of different sizes. While in heatmap (c) the diagonal mean is .05 and the off-diagonal 0.2, with n=20 paraphrases, the diagonal decreases to a mean of .04 and the off-diagonal to .18; with n=100, the diagonal and off-diagonal means are respectively .09 and .39, showing that a higher n enables a stronger transfer.

5.4 Analysis

To gain further insight on our procedure, we show some instances from ISEAR which we found challenging for our models, and show them across the three experiments in the beamsearch scenario (Table 3, letters in bold correspond to inputs). Their backtranslations have lost the original connotation, so much so that the classifier assigns them to a different emotion class (this happens for 387 inputs in setup (a), see Figure 3).

Change in emotion (both loss and alteration) often seems to involve a relatively small number of recurring linguistic transformations, like the change of modality (c. and f.), or in the intensity of the adjectives (b. and d.). The fact that *disgust* leaves room for *shame* (c.) appears coherent with the theory that the latter is related more to the self (Tracy and Robins, 2006): as opposed to the output of the transfer, the input presents the action as one that the experiencer *had to* take. In d., *sadness* replaces *disgust* with the use of a softer expression, such as "loathe". This example also highlights that removing a direct emotion word can determine a switch in connotation. Another reason why the backtranslation in b. is associated to *fear* could be that silence, in ISEAR, mostly occurs in the description of disruptive, frightening events, similarly to being "approached" by strangers (and hence, the joyful sentence in e. turns into *fear*).

There are also signals that emotion changes show a gender bias (Sun et al., 2019): characterising the subject as a male moves *anger* to *guilt* or *joy* (a.), while we have found that female characters can elicit an association with *shame*.

As for the transfer, it is possible that smaller lexical changes are sufficient to change emotions when the input label and the new emotions can co-occur. For instance, *anger* and *guilt*, being negative emotions, are more likely to co-occur than *anger* and *joy*, corresponding to output 1 and 3 for the first sentence. These examples also show that transfer can happen without disrupting grammaticality nor content – at least within the relatively small top-n lists we considered. Yet, this observation needs further exploration, because striking a balance between all transfer desiderata and aggregate their separate evaluations still represents a challenge in the field. Moreover, we need a better understanding of the contrast between the findings of Mohammad et al. (2016) (altering polarity hampers human's ability to determine the original sentiment of the text but does not mislead automatic predictions) and ours (emotion changes in the above examples are comparatively marginal).

6 Conclusion and Future Work

Our goal was to understand if automatic translation retains the emotional substance of texts. We found that a state-of-the-art NMT system tends to tone down emotion connotations, thus presenting a problem

RQ	Emotion	Sentence
	A	**a.** *When I have to take exams I am very excited and have not much time for the housekeeping. Then my friend has to do everything*
1	G	When I have to take exams, I am very excited and do not have much time for the budget. Then my boyfriend has to do everything.
2	A	When I have to take exams, I am very excited and have little time for housekeeping. Then my girlfriend has to do everything she can.
3	J	When I have exams, I'm very excited and I don't have much time for the household. Then my boyfriend has to take care of everything.
	A	**b.** *When a friend told me a story and I stayed dumb because I had no story to tell.*
1	F	When a friend told me a story and I remained silent because I had no story to tell.
2	A	If a friend told me a story and I was mute because I had no story to tell.
3	G	When a friend told me a story, I stayed silent because I had nothing to tell.
	D	**c.** *On New Years eve I drank too much alcohol, so much that I had to vomit in the presence of other people.*
1	Sh	On New Year's Eve I drank so much alcohol that I vomited in the presence of other people.
2	D	On New Year's Eve I drank so much alcohol that I had to vomit in the presence of other people.
3	G	On New Year's Eve, I drank too much alcohol, so much that I threw up in the presence of other people.
	D	**d.** *I feel disgusted with the bootlickers, with helpless people.*
1	Sa	I loathe the bootleggers, the helpless people.
2	D	I am disgusted by the boot-lickers, by the helpless people.
3	Sh	I loathe boots, I loathe helpless people.
	J	**e.** *When a person that I like very much got near to me.*
1	F	If a person I like very much approached me.
2	J	If a person I like very much got close to me.
3	D	If someone I like came up to me.
	F	**f.** *I was going to knock down a pedestrian with my car.*
1	A	I was trying to push a pedestrian over with my car.
2	F	I was going to knock over a pedestrian with my car.
3	J	I wanted to overturn a pedestrian with my car.
	Sh	I tried to knock over a female pedestrian with my car.

Table 3: Examples for the three research questions tackled in this paper. Backtranslations with a different emotion connotation correspond to RQ1; those where the emotion is recovered to RQ2; and those with a different emotion correspond to RQ3. Input ids are in bold.

for the development of affect-aware MT products, for cross-lingual research based on the translation of existing data, and for communication aided by MT. We showed how an emotion-informed subselection of translation candidates can improve this situation without adversely affecting translation accuracy. Moreover, we used the same post-processing methodology to induce emotion variability and address the task of emotion style transfer.

Results show that MT outputs can be improved in their emotion rendering, but we relied on a back-translation pipeline instead of a real-world translation scenario. This motivates an important next research step, namely the development of an emotion classifier which estimates emotion probability distributions in multiple languages in a comparable manner. It is still open how to measure comparability and how to optimize that measure. Finally, our analysis relied on a single NMT system, namely FAIRSEQ. Despite our argument that this tool is representative for a range of systems, our study should be extended to other systems and other target languages.

Acknowledgements

This work was supported by Leibniz WissenschaftsCampus Tübingen "Cognitive Interfaces". We thank Laura Ana Maria Oberländer, Irshad Ahmad Bhat and Manuel Mager for fruitful discussions.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, TSD'07, pages 196–205, Berlin, Heidelberg. Springer-Verlag.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea, July. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Jeremy Barnes and Roman Klinger. 2019. Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. *Journal of Artificial Intelligence Research*, 66:691–742, Nov.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana, 06. Association for Computational Linguistics.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online, July. Association for Computational Linguistics.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland, June. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July. Association for Computational Linguistics.

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)*, pages 663–670.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Seattle, Washington, USA, October. Association for Computational Linguistics.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2008. Valentino: A tool for valence shifting of natural language texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

David Helbig, Enrica Troiano, and Roman Klinger. 2020. Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 41–50, Online, July. Association for Computational Linguistics.

Séverine Hubscher-Davidson. 2017. *Translation and Emotion: A psychological perspective*. Routledge.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

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

Kaisla Kajava, Emily Ohman, Piao Hui, and Jörg Tiedemann. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings.

Joseph Lee, Ziang Xie, Cindy Wang, Max Drach, Dan Jurafsky, and Andrew Ng. 2019. Neural text style transfer via denoising and reranking. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 74–81, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108(1):73–84.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022, Florence, Italy, July. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain, April. Association for Computational Linguistics.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal, September. Association for Computational Linguistics.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy, August. 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, July. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Olga Petrova and Maria Rodionova. 2016. Rendering emotional coloring in literary translation. *Procedia-Social and Behavioral Sciences*, 231:195–202.

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, July. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain, April. Association for Computational Linguistics.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado, May–June. Association for Computational Linguistics.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, September. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

Chenglei Si, Kui Wu, Ai Ti Aw, and Min-Yen Kan. 2019. Sentiment aware neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 200–206, Hong Kong, China, November. Association for Computational Linguistics.

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

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy, July. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Jessica L. Tracy and Richard W. Robins. 2006. Appraisal antecedents of shame and guilt: Support for a theoretical model. *Personality and social psychology bulletin*, 32(10):1339–1351.

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.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578, Hong Kong, China, November. Association for Computational Linguistics.

Claudia Wassmann. 2017. Forgotten origins, occluded meanings: Translation of emotion terms. *Emotion Review*, 9(2):163–171.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, and Senzhang Wang. 2020. Decode with template: Content preserving sentiment transfer. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4671–4679, Marseille, France, May. European Language Resources Association.

Simon Whitehead and Lawrence Cavedon. 2010. Generating Shifting Sentiment for a Conversational Agent. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 89–97, Los Angeles, CA, June. Association for Computational Linguistics.

Anna Wierzbicka. 2013. *Imprisoned in English: The hazards of English as a default language*. Oxford University Press.

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

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, December. The COLING 2012 Organizing Committee.

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

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

Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 679–688, New York, NY, USA. Association for Computing Machinery.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California, June. Association for Computational Linguistics.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia, July. Association for Computational Linguistics.

A Example Outputs

Challenging examples for the models. Sentences corresponding to RQ1 show how the same sentence is transformed by the MT system as is. Those corresponding to RQ2 were produced by our emotion-based procedure to recover the original emotion connotations, and those corresponding to RQ3 were selected by the same emotion-based procedure, when used to transfer emotions.

RQ	Emotion	Sentence
	G	*Feeling guilt after greed, buying chocolate and pigging out to the point of feeling sick, especially as I am fat.*
1	D	Feelings of greed, buying chocolate and exploitation to the point of nausea, mainly because I'm fat.
2	G	Feeling guilty about greed, buying chocolate and feeling sick, especially because I'm fat.
3	Sh	Feelings of greed, buying chocolate and feeling ill, mainly because I'm fat.
	F	*When I was first exposed to the dead bodies, for dissecting purposes at the school of medicine.*
1	D	When I was first confronted with the corpses to dissect them in medical school.
2	F	The first time I was confronted with the bodies, I dissected them in the medical school.
3	F	The first time I was confronted with the bodies, I dissected them in the medical school.
	Sa	*When my sister had the still born child, she was emotionally very deep down, and it took her a long time to recover.*
1	J	When my sister gave birth to the baby, she was very emotional and it took a long time for her to recover.
2	Sa	When my sister had the baby, she was emotionally very deep inside and it took a long time for her to recover.
3	A	When my sister had the baby, she was emotionally very low and it took a long time for her to recover.
	A	*During a recent meeting, Mr. A showed his excitement and overindulged in the notes delivered. Though his curiosity could not be blamed, his way of accquiring knowledge was an extreme behaviour e.g he always tried to know what I was reading and gained everything he could.*
1	D	During a recent meeting, Mr. A. showed his enthusiasm and left himself to the notes handed down. Although his curiosity could not be reproached, his way of acquiring knowledge was extreme, i.e. he always tried to know what I was reading and gained everything he could.
2	A	During one recent meeting, Mr. A. showed his enthusiasm and indulged excessively in the handed down notes. Although his curiosity could not be blamed, his way of acquiring knowledge was extreme, i.e. he always tried to know what I was reading and gained all he could.
3	Sa	During a recent meeting, Mr. A. showed his enthusiasm and revelled excessively in the notes handed down. Although he could not be blamed for his curiosity, his way of acquiring knowledge was extreme, that is, he always tried to know what I was reading and gained everything he could.
	D	*3 years ago I served in the army. Once a collegue denounced me because of a delict, which is usually committed. I was arrested for 3 days. I still detest this man.*
1	G	I served in the Army three years ago. A colleague once reported me for a crime that is normally committed. I was arrested for three days. I still loathe this man.
2	D	I served in the military three years ago. One time, a colleague reported me for a crime that is usually committed. I was arrested three days ago. I still detest that man.
3	Sa	Three years ago I was in the army. On one occasion a colleague reported me for an offence that is usually committed. I've been detained for three days. I still despair of this man.
	A	*When another fellow worker decided to leave the company. We had been very close and we would not be able to work with eachother any longer.*
1	Sa	When another employee decided to leave the company. We were very close and couldn't work together.
2	A	As another employee decided to leave the firm. We were close and couldn't work together any more.
3	G	When one more employee decided to leave the company. We were very close and could no longer work with one another.

Table 4: Examples for the three research questions tackled in this paper, with ISEAR. A: anger, D: disgust, F: fear, G: guilt, J: joy, Sa: sadness, Sh: shame.