Multi-perspective Alignment for Increasing Naturalness in Neural Machine Translation

Huiyuan Lai[1], **Esther Ploeger**[2], **Rik van Noord**[1], **Antonio Toral**[3*]
[1]CLCG, University of Groningen, The Netherlands
[2]Department of Computer Science, Aalborg University, Denmark
[3]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain
{h.lai,r.i.k.van.noord}@rug.nl espl@cs.aau.dk antonio.toralr@ua.es

Abstract

Neural machine translation (NMT) systems amplify lexical biases present in their training data, leading to artificially impoverished language in output translations. These language-level characteristics render automatic translations different from text originally written in a language and human translations, which hinders their usefulness in for example creating evaluation datasets. Attempts to increase naturalness in NMT can fall short in terms of content preservation, where increased lexical diversity comes at the cost of translation accuracy. Inspired by the reinforcement learning from human feedback framework, we introduce a novel method that rewards both naturalness and content preservation. We experiment with multiple perspectives to produce more natural translations, aiming at reducing machine and human translationese. We evaluate our method on English-to-Dutch literary translation, and find that our best model produces translations that are lexically richer and exhibit more properties of human-written language, without loss in translation accuracy.

1 Introduction

While machine translation (MT) has achieved promising performance with the adoption of neural network (Bahdanau et al., 2015; Vaswani et al., 2017; Team NLLB et al., 2022), automatic translations remain markedly different from translations by professional human translators. A striking example is the fact that MT outputs exhibit reduced lexical diversity (Vanmassenhove et al., 2019, 2021) and increased source-language interference (Toral, 2019) compared to human translation (HT). These linguistic differences were previously referred to as *machine translationese* (de Clercq et al., 2020; Bizzoni et al., 2020; Vanmassenhove et al., 2021).[1]

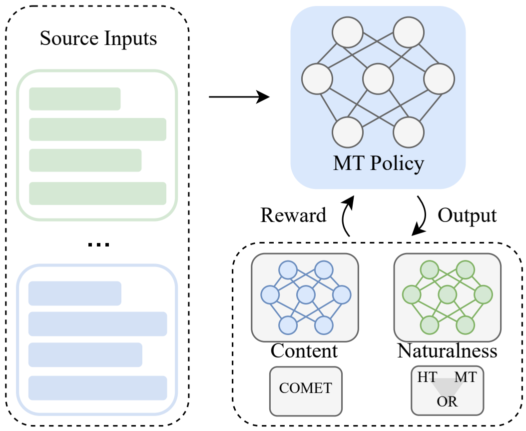

Figure 1: Aligning the translation policy from both content preservation and naturalness perspectives.

Within the context of natural language processing (NLP), these language-level artifacts of machine translation can have negative implications. For example, machine translationese in NLP evaluation datasets can inflate performance assessments. Examples of this are found in MT (Zhang and Toral, 2019; Graham et al., 2020) and cross-lingual transfer learning (Yu et al., 2022; Artetxe et al., 2020). Furthermore, in the field of literary translation, preserving reading experience (and thus the original style) can be an important aspect of the translation process (Delabastita, 2011; Toral and Way, 2015; Guerberof-Arenas and Toral, 2020).

Reducing translation artifacts in MT output is not trivial. Intuitively, translated texts should match the style of the texts originally written in that target language, while preserving the content of the source language. This trade-off between naturalness and content preservation presents methodological challenges. For example, previous work shows a decrease in translation quality when aiming to recover lexical diversity in MT (Ploeger et al., 2024). Moreover, existing approaches such as Tagging (Freitag et al., 2022), aim to increase MT naturalness in a rigid manner, while the amount

*Work partly carried out while affiliated with the University of Groningen.

[1]This term has since been criticized, see for example Crespo (2023).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28071–28084

of naturalness in the output translation cannot be manually adjusted to a desired level. Yet, in cases where faithfulness to the source is crucial, the naturalness of a translation may be of lesser importance (Parthasarathi et al., 2021).

To address these challenges, we frame the task of increasing naturalness in MT as a text style transfer-like task, where style and content are the two core aspects (Mou and Vechtomova, 2020; Lai et al., 2021b,a). In practice, we train a vanilla MT model with supervised learning and subsequently exploit reward learning that fosters naturalness and content preservation, as shown in Figure 1. With respect to naturalness we explore two objectives: making MT more akin to human translations (i.e. reducing machine translationese) and making MT more akin to texts originally written in the target language, i.e. reducing translationese (Gellerstam, 1986; Baker, 1993; Toury, 2012). We evaluate our framework on a dataset for English-to-Dutch literary translation. Our main contributions are as follows:

- We introduce a novel flexible multi-perspective alignment framework that favours natural translation outputs while fostering content preservation;
- We experiment with and analyse the results of three different preference classifiers that are used to produce more natural translations: preferring original target-language text (OR) over HT, OR over MT, and HT over MT;
- Extensive experiments show that our model produces translations that are lexically richer than baseline MT systems without loss in translation quality.[2]

2 Related Work

2.1 Increasing MT Naturalness

A few approaches have been put forward to make MT outputs more natural. For example, Freitag et al. (2019) trained a post-processor that learns to translate from round-trip machine translated text to original text in the same language, which can be applied to the outputs of existing MT systems. Freitag et al. (2022) prepend their training examples with special tags that denote whether the target side of the training data was originally written in that language or not. These methods are rigid, while in some cases, content preservation may be more important than style (Parthasarathi et al., 2021). In response, Ploeger et al. (2024) propose a flexible approach based on reranking translation candidates, but report considerable loss in general translation quality.

In parallel efforts, some research aims to reduce translationese from human translations, and uses monolingual approaches based on text style transfer (Jalota et al., 2023), semantic parsing (Wein and Schneider, 2024) and debiasing embeddings (Dutta Chowdhury et al., 2022). Additionally, there is growing interest in leveraging human feedback to improve overall translation quality where a single metric such as COMET trained from human annotations is used as the reward model (Ramos et al., 2024; He et al., 2024). In this work we focus on improving translation quality from multiple perspectives, which is tailorable to the downstream scenario, while still being faithful to the source texts.

2.2 (Machine) Translation Detection

Following neural machine translation (NMT), a new line of research started to investigate the extent to which translations (including HT and MT) contain artefacts, and how these compare to original texts and human translations.

HT vs OR Classification Baroni and Bernardini (2005) showed that original texts can be distinguished from human-translated texts with computational methods. Concrete textual markers, such as the frequency of function words or the use of punctuation, have been associated with this difference (Koppel and Ordan, 2011; Volansky et al., 2015). Beyond hand-crafting specific linguistic features, Pylypenko et al. (2021) find that neural architectures provide a reliable tool for distinguishing translated from original texts. They obtain state-of-the-art performance by fine-tuning multilingual BERT (Devlin et al., 2019) on the task.

MT vs HT Classification Bizzoni et al. (2020) show that there is a difference between the translation artifacts produced by humans and MT models. van der Werff et al. (2022) use neural language models to distinguish between HT and NMT in German-to-English translation, and highlight the challenges of this task, with their sentence-level system achieving an accuracy of approximately 65%. This is further investigated in a multilingual setting (Chichirau et al., 2023).

[2] All code at `https://github.com/laihuiyuan/alignment4naturalness`

Data Split	Language	# Books	# Sentences
Translationese Detection			
Train	Dutch (OR)	143	982,114
	Dutch (HT)	143	1,390,351
Test	Dutch (OR)	36	261,151
	Dutch (HT)	36	340,950
Machine Translation			
Train	Dutch (HT)	495	4,874,784
	English (OR)	495	4,874,784
Valid	Dutch (HT)	5	88,881
	English (OR)	5	88,881
Test	Dutch (HT)	31	302,976
	English (OR)	31	302,976
Baseline (Train)	Dutch (OR)	-	4,874,784
Baseline (Valid)	Dutch (OR)	-	88,881

Table 1: Data set division and size.

These works show that HT, MT and original texts can, to some extent, be distinguished from each other with neural methods. Based on this, we expect that our reward functions with neural classifiers can be effective for improving naturalness in MT outputs.

3 Data

In this section, we describe datasets used for (machine) translation detection and MT, including both a parallel and a monolingual corpus of books. Table 1 shows the sizes and splits of both datasets.

Translationese Detection Data We use a dataset consisting of books written in Dutch (Toral et al., 2021) from a range of authors and genres, as preprocessed by Ploeger et al. (2024). The dataset contains 7,000 books that were manually annotated to be originally written in Dutch (OR) or in another language (HT). From these, we derive two balanced subsets: 286 books for training and 72 for testing.

Machine Translation Data We use the parallel dataset from Toral et al. (2021), preprocessed by Ploeger et al. (2024). This dataset consists of 531 books that were originally written in English (OR) and human translated into Dutch (HT), of which 495 books for training, 5 for validation and 31 as a test set. The genres of these books vary, including literary fiction, popular fiction, non-fiction and children's books from over 100 authors. Particularly, the test set also contains a broad range of books.[3] In addition, we use monolingual data for the two baseline MT systems (see Section 5.1), consisting of a random sample of equal size to the parallel training data and disjoint from the subset used for translation detection.

4 Methodology

In this section, we first introduce the base MT model (Section 4.1) and binary translationese classification models (Section 4.2) using supervised learning. Subsequently, we propose a multi-perspective alignment framework based on reward learning, which explicitly optimises the MT model with human expectations, aiming to increase naturalness and to preserve content (Section 4.3).

4.1 Base MT Model

As the initial step of model alignment, we train the base MT model with supervised learning on high-quality parallel data. Specifically, given a source text $x = \{x_1, \cdots, x_n\}$ of length n in language l_s and a target text $y = \{y_1, \cdots, y_m\}$ of length m in language l_t from dataset $\mathcal{D}$, the MT model aims to learn two conditional distributions, transforming x to y. We begin with Transformer-based models whose goal is to minimize the following negative log-likelihood:

$$\mathcal{L}_{nl} = -\frac{1}{m}\sum_{i=1}^{m} \log\left(p(y_i|y_{0:i-1}, x; \theta)\right) \quad (1)$$

Where θ represents model parameters and y_i the i-th token of the target sequence.

4.2 (Machine) Translationese Classification

We use three different classifiers, seeing the promotion of natural translations from different perspectives, namely preferring OR over HT, HT over MT, and OR over MT. The first classifier aims at reducing human translationese, while the second and third ones aim at reducing machine translationese (the second one with respect to HT and the third one with respect to OR). These classifiers will be used as rewards (Section 4.3) to foster naturalness. Having three perspectives will allow us to find out how each of them impacts the accuracy and naturalness of the resulting translations.

For HT vs OR classification, we use the monolingual Dutch data introduced in the first part of Table 1. For the other two settings, we translated a

[3]A full list of author names, titles, genres and publishing years of the test set books can be found in Appendix A.1.

Algorithm 1 Multi-perspective alignment algorithm for Naturalness and Content

Require: Base MT model $p(y|x;\theta_0)$, Training set: source $\boldsymbol{X}$ and target $\boldsymbol{Y}$
Require: Reward function: COMET $C(x,y,\hat{y})$ and translationese classification $p(t_1|\hat{y};\phi)$
for each step $i = 0, 1, \cdots, m$ **do**
  $M_i \leftarrow$ MiniBatch$(\boldsymbol{X}, \boldsymbol{Y})$
  for $x \in M_i$ **do**
    $\hat{y} \sim p(y|x;\theta_i)$
    Calc. translationese reward $r_t(\hat{y})$ by Eq. 2
    Calc. content reward $r_c(\hat{y})$ by Eq. 3
    Calc. overall reward $r(\hat{y})$ by Eq. 4
  end for
  Update MT model using data M_i and $\hat{M}_i$ with the overall reward based on Eq. 6
end for

subset of the English text in the parallel data (second part of Table 1) of equal size to the monolingual training data (982,114 sentences) into Dutch using the base MT model. The resulting machine translated sentences are combined with OR texts in the monolingual data for MT vs OR classification and with HT texts in the parallel data for MT vs HT. We filter out machine translated texts that are identical to human translations. Based on the above data, we fine-tune the Dutch language model BERTje (de Vries et al., 2019) for the binary detection tasks, obtaining three different models.

4.3 Multi-perspective Alignment for Naturalness and Content Preservation

We introduce our method which ranks samples based on rewards that target naturalness and content preservation. This approach is inspired by recent work in text style transfer, where both meaning has to be preserved and style should be transferred (Lai et al., 2021b,a). This content vs form trade-off is similar to our situation with content preservation and naturalness. Specifically, after training a base MT model using supervised learning (Section 4.1), we further align it with human expectations in terms of naturalness and content in the form of reward learning.

Based on the base MT model, we train our reward learning based framework. The MT model takes source text x as input and generates the corresponding translated text $\hat{y}$. To ensure the quality of the $\hat{y}$, we design two rewards that aim to foster naturalness and content preservation. We consider the two quality feedbacks as rewards and fine-tune the MT model through reinforcement learning. The overview of our alignment framework is shown in Algorithm 1.

Rewarding Naturalness We use a binary translationese classifier (OR vs HT, HT vs MT or OR vs MT) to assess how well the translated text $\hat{y}$ scores on the translationese aspect, i.e., to assess its (machine) translationese probability. Formally, this reward is formulated as

$$r_t(\hat{y}) = \begin{cases} 0 & \text{if } p(t_1|\hat{y};\phi) < \sigma_t \\ p(t_1|\hat{y};\phi) & \text{otherwise} \end{cases} \tag{2}$$

where ϕ is the parameter of the classifier. σ_t is the translationese threshold, which is set to 0.5 in our experiments based on preliminary results.

Rewarding Content We employ COMET (Rei et al., 2020) as the content-based reward model $C(x,y,\hat{y})$ to assess the content quality of $\hat{y}$ as the translation of x. This is formulated as

$$r_c(\hat{y}) = \begin{cases} 0 & \text{if } \mathrm{C}(x,y,\hat{y}) < \sigma_c \\ \mathrm{C}(x,y,\hat{y}) & \text{otherwise} \end{cases} \tag{3}$$

Where C$(\cdot)$ represents the COMET model and σ_t represents the content threshold, which is set to 0.85 in our experiments based on preliminary results.

Overall Reward To encourage the model to foster naturalness while preserving the content, the final reward is the harmonic mean of the above two rewards

$$r(\hat{y}) = \begin{cases} 0 & \text{if } r_t = 0 \text{ or } r_c = 0 \\ \frac{2}{1/r_t + 1/r_c} & \text{otherwise} \end{cases} \tag{4}$$

Learning Objectives Here we aim to maximize the expected reward of the generated sequence $\hat{y}$, the loss is defined as

$$\mathcal{L}_{rw} = -\frac{1}{m}\sum_{i=1}^{m} r(\hat{y})\log\left(p(\hat{y}_i|\hat{y}_{0:i-1}, x;\theta)\right) \tag{5}$$

To keep the fine-tuned model from moving too far from the base MT model, we combine the reward objective with the supervised training loss instead of using a reference model requiring large computing resources. Therefore, the final objective function of our framework consists of two components: negative log-likelihood loss in Eq. 1 and reward-based loss in Eq. 5, which are jointly formulated as

$$\mathcal{L}(\theta;\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\beta\mathcal{L}_{nl} + \mathcal{L}_{rw}] \tag{6}$$

Where β a is a hyperparameter used to control the weight of the negative log-likelihood loss (set to 0.5

in our main experiments), allowing our method to be tailorable. We employ the policy gradient algorithm (Williams, 1992) to maximize the expected reward.

5 Experimental Setup

5.1 Baselines

In addition to the base MT model (Section 4.1), we include three previous methods that aim at reducing machine translationese as baselines: Tailored RR (Top-k) (Ploeger et al., 2024), automatic post-editing (APE) (Freitag et al., 2019) and Tagging (Freitag et al., 2022).

Tailored RR is an approach that involves reranking translation candidates with a classifier that distinguishes between original and translated text. We select the Tailored RR (Top-k), which reranks candidates that are obtained through Top-k sampling, as a baseline, since it retrieves the highest diversity in Ploeger et al. (2024).

APE aims to train a post-processor that transforms machine-translated Dutch into more natural Dutch texts. To obtain parallel data of source synthetic Dutch and original Dutch, we round-trip translate the original Dutch text of the monolingual data.

Tagging aims to learn to differentiate between original and translated texts. We use the base Dutch-English MT model to obtain English translations of the monolingual original Dutch text. Then, we prepend a tag `<orig>` to the English text in the above data, `<tran>` to the English text in the parallel data, and train a new MT model on the concatenation of these two datasets.

We include two settings for the amount of original target data (i.e. `<orig>`): one equivalent to the parallel training data (4.8M) and the other to the translationese classifier data (1M). This is done to investigate how the proportions of target-translated vs target-original in the training data affect results. Our hypothesis is that the larger the percentage of target-original the more natural the translations, but at the expense of lower translation accuracy.

5.2 Implementation Details

All experiments are implemented using the library HuggingFace Transformers (Wolf et al., 2020). We use the BART (Lewis et al., 2020) architecture with 6 Transformer-based (Vaswani et al., 2017) layers in both the encoder and decoder. The base MT models are trained using the AdamW optimiser (Loshchilov and Hutter, 2019) with a cosine learning rate decay, and a linear warmup of 1,000 steps. The maximum learning rate is set to 1e-4, the batch size is 256, and the gradient accumulation is 2; all reward-based models are trained with a consistent learning rate of 2e-5. We evaluate the model every 1,000 steps and use early stopping with patience 3 if the cross-entropy loss on the validation set does not decrease.

We use beam search with size 5 during inference. Since some of the training data contains instances of repeated punctuation marks, this led to the reinforcement learning method tending to optimize the model for higher rewards. Therefore, we take a simple post-processing step to remove consecutive repeated punctuation marks after the text is generated.[4]

5.3 Evaluation Methods

We perform a comprehensive evaluation on the model outputs, including translation quality and translationese evaluation. Unless stated otherwise, the scores are reported by taking the averages for all books in the test set.

Translation Quality We employ three metrics to automatically calculate the content preservation of the output based on human references (and source sentences), namely BLEU (Papineni et al., 2002), COMET (Rei et al., 2020, 2022), and MetricX (Juraska et al., 2024). We use the Sacre-BLEU implementation (Post, 2018) for BLEU. Regarding the COMET family models, we use both the default model `wmt22-comet-da` (COMET), and the reference-free model `wmt22-cometkiwi-da` (KIWI) that is not used for reward learning. For MetricX, we use `MetricX-24-Hybrid-XL`, considering it our most important translation quality metric, since it achieved state-of-the-art performance at the WMT24 Metrics Shared Task (Freitag et al., 2024).

Translationese Evaluation We apply the translationese detection models to MT outputs and report the rate (i.e. classification accuracy) at which they are classified as the target aspect, such as OR in HT-OR, with higher rates indicating that the outputs are more human-like. Additionally, as previous studies show that translated texts are often simpler than original texts (Baker, 1993), our evaluation

[4]See Appendix A.2 for post-processing examples.

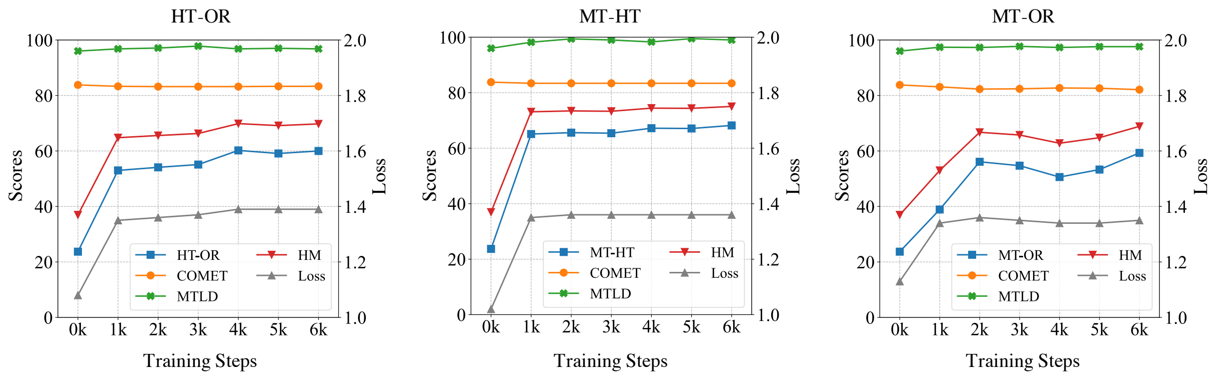

Figure 2: Evaluation results on the validation set under various settings. Notes: (i) The training step of 0K represents the base MT model; HM indicates the harmonic mean of classification accuracy and COMET score.

also covers lexical diversity. Here we report six different metrics:

- TTR (Templin, 1957): Type-Token Ratio is the number of unique words (types) divided by the total number of words in the text.
- Yule's I (Yule, 1944): Given the size of the vocabulary (number of types) V and $f(i, N)$ representing the numbers of types which occur i times in a sample of length N, Yule's I is calculated as

$$\text{Yule's I} = \frac{V^2}{\sum_{i=1}^{V} i^2 * f(i, N) - V} \quad (7)$$

- MTLD (McCarthy, 2005): evaluated sequentially as the average length of sequential word strings in a text that maintains a given TTR value. We use a threshold of 0.72, following Vanmassenhove et al. (2021). This metric has been shown to be stable across different text lengths (McCarthy and Jarvis, 2010), which is why we consider it more important a metric than TTR or Yule's I.
- B1 (Vanmassenhove et al., 2021): the percentage of words in the output that are in the estimated 1,000 most frequent words in a language.
- PTF (Vanmassenhove et al., 2021): the average percentage (over all relevant source words) of times the most frequent translation option was chosen among all translation options.
- CDU (Vanmassenhove et al., 2021): the cosine similarity between the output vector for each source word and a vector of the same length with an equal distribution for each translation option.[5]

[5]See Ploeger et al. (2024) for details on its implementation.

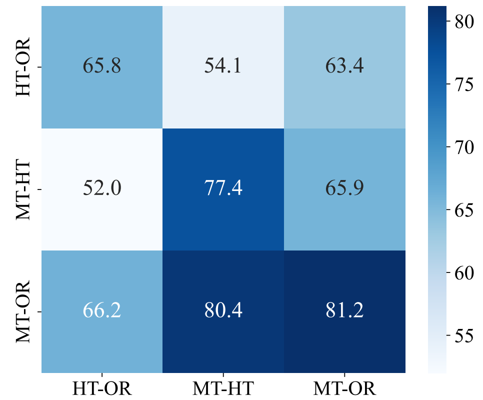

Figure 3: Confusion matrices for the different binary classifiers. Each row represents the results of a classifier tested on different test sets.

6 Results and Analysis

6.1 Initial Results

Translationese Classification Figure 3 shows the performance confusion matrix of different binary translationese classifiers on different test sets. We can observe that for each classifier scores on the main diagonal are higher than others, with MT-OR having the highest score, followed by MT-HT and HT-OR. This is on par with the performance in similar scenarios from previous work (Pylypenko et al., 2021). While van der Werff et al. (2022) show that the distinction between the translation variants (MT and HT) is challenging, we found that human translations are even more difficult to distinguish from the original target-language texts. Interestingly, MT-OR achieves higher accuracy on the sets of HT-OR and MT-HT than their corresponding classifiers.

Machine Translation During the alignment phase (see Section 4.3) we find that for some mod-

	Translation Accuracy				Classification Accuracy			Lexical Diversity					
	BLEU	COMET	KIWI	MetricX↓	HT-OR	MT-HT	MT-OR	TTR	Yule's I	MTLD	B1↓	PTF↓	CDU↓
Human Translation	-	-	-	-	32.9	69.3	48.6	0.153	3.934	96.0	0.672	0.817	0.548
Tailored RR	21.2	74.5	72.4	4.86	**35.1**	**52.9**	33.5	0.157	4.170	**104.3**	0.682	**0.815**	0.559
APE	29.9	80.4	77.9	3.38	33.7	33.6	35.2	0.155	3.670	91.7	0.682	0.824	0.561
Tagging (1M)	31.6	81.6	80.1	2.87	33.0	42.6	36.9	0.161	4.133	95.8	0.671	0.817	**0.554**
Tagging (4.8M)	31.1	80.9	79.7	3.05	33.5	43.2	**39.0**	**0.164**	**4.347**	96.8	**0.667**	**0.815**	0.556
BM: Base MT Model	**32.5**	**82.3**	**80.4**	**2.66**	28.1	18.9	17.6	0.150	3.537	90.4	0.677	0.826	0.563
BM + COMET & HT-OR	29.7	80.4	79.9	2.83	**34.0**	24.0	25.5	0.145	3.239	91.0	0.675	0.830	0.554
BM + COMET & MT-HT	**32.1**	**82.2**	**80.6**	**2.63**	26.1	**33.4**	26.6	**0.150**	**3.572**	**93.3**	**0.674**	0.828	0.553
BM + COMET & MT-OR	31.5	81.5	80.1	2.75	28.7	33.3	**28.2**	**0.150**	3.544	91.8	0.678	**0.827**	**0.542**

Table 2: Translation performance under various settings. Note that bold numbers indicate the best system for each block, and underlined numbers indicate the best score by an MT system for each metric.

els the valid loss does not correlate with the naturalness aspect (i.e. classifier's accuracy) after 1k steps: while naturalness improves, the loss on the validation set stays flat. Therefore we manually select checkpoints between the 1k and the 6k steps, and report their evaluation and loss curves in Figure 2.

The first observation is that all models achieve substantial improvement in naturalness over the first 1k steps compared to the base MT model (i.e. 0K), as reflected in the results for translationese classification (HT-OR, MT-HT and MT-OR) and lexical richness (MTLD). Although the COMET scores of some models decrease slightly, the overall score HM follows the trend of the translationese aspect. After 1k steps, MTLD scores and valid loss tend to be flat; translated language classification shows a clear improvement from 1k to 2k steps on MT-OR, a slight increase on HT-OR, and remains stable on MT-HT. For all models, although some metrics fluctuate after 2k steps, they tend to be stable overall. For the remaining experiments, we report the results of the alignment model at 5K trainingsteps.

6.2 Main Results

We report the main results in Table 2, including the base MT model, the three baselines and our methods trained with both rewards: COMET for content preservation and the three different classifiers for naturalness (i.e. HT-OR, MT-HT and MT-OR).

Tailored RR achieves the highest naturalness scores (e.g. HT-OR and MTLD), but performs the worst on all translation accuracy metrics. Compared to APE, Tagging consistently performs better across the board, both in terms of content (i.e. translation accuracy) and naturalness. Additionally, we observe that using more target-original data (i.e. 4.8M vs 1M) results in lower accuracy scores but better naturalness metrics, which is consistent with our hypothesis (see Section 5.1). Overall, we observe that the three baselines underperform the base MT model in terms of translation accuracy and outperform it in most cases when it comes to naturalness metrics.

Moving to our approach, when comparing different classification rewards, the model trained with COMET & MT-HT achieves, overall, better scores than our other two models (HT-OR and MT-OR). We speculate that the rewards that foster OR do not work as well due to a mismatch between the preference of the classifier (OR) and the data in the target side of the MT training data (HT). We thus believe that such classifiers could be useful in scenarios in which the target side of the MT training data contains texts originally written in that language, which would be common in translation directions in which the target language is higher-resourced than the source language.

Overall, our best system (BM + COMET & MT-HT) achieves better naturalness scores than the base MT model (e.g. 93.3 vs 90.4 for MTLD), while even having a higher KIWI score (80.6 vs 80.4) and a lower MetricX score (2.63 vs 2.66; lower is better), two metrics that have not been used in our reward learning. Tagging attains higher naturalness scores but this comes at the price of a notable reduction in translation accuracy, as shown by KIWI (79.7 vs 80.6) and MetricX (3.05 vs 2.63).

6.3 Ablation Study

To assess the contribution of each reward component in our framework, we perform a set of ablation studies, the results of which are shown in Table 3. For the COMET vs COMET + classifier setting, we see higher naturalness scores in the latter in all cases for MT-HT and MT-OR (except CDU

	Translation Accuracy				Classification Accuracy			Lexical Diversity					
	BLEU	COMET	KIWI	MetricX↓	HT-OR	MT-HT	MT-OR	TTR	Yule's I	MTLD	B1↓	PTF↓	CDU↓
BM: Base MT Model	32.5	82.3	80.4	2.66	28.1	18.9	17.6	0.150	3.537	90.4	0.677	0.826	0.563
BM + COMET	32.2	81.9	80.7	2.64	26.7	19.1	19.6	0.147	3.362	90.9	0.679	0.830	0.543
BM + HT-OR	31.1	81.0	80.0	2.75	30.3	21.5	22.1	0.137	1.950	26.8	0.700	0.826	0.556
BM + HT-OR & COMET	29.7	80.4	79.9	2.83	34.0	24.0	25.5	0.145	3.239	91.0	0.675	0.830	0.554
BM + MT-HT	32.2	81.5	80.2	2.67	28.2	24.7	22.4	0.149	3.465	91.2	0.679	0.826	0.556
BM + MT-HT & COMET	32.1	82.2	80.6	2.63	26.1	33.4	26.6	0.150	3.572	93.3	0.674	0.828	0.553
BM + MT-OR	32.6	81.9	80.3	2.65	26.8	22.9	22.4	0.149	3.460	90.8	0.680	0.826	0.559
BM + MT-OR & COMET	31.5	81.5	80.1	2.75	28.7	33.3	28.2	0.150	3.544	91.8	0.678	0.827	0.542

Table 3: Ablation study: The contribution of each reward component, where we fine-tune the base MT model using only the content reward or the naturalness reward.

Source	Text
Original English	It was because of the atmosphere of hockey-fields and cold baths and community hikes and general clean-mindedness which she managed to carry about with her.
Human Translation	Het was om de sfeer van hockeyvelden en koude douches en groepsuitstapjes en algemene geestelijke reinheid die zij om zich wist te verspreiden.
Tagging (4.8M)	Het kwam door de sfeer van hockeyvelden en koude baden en gemeenschapsfietsen en algemeene properheid, die zij met haar wist rond te voeren.
BM: Base MT Model	Het kwam door de sfeer van hockeyvelden, koude baden en plattelandskantoren en algehele schoonheid die ze met zich mee kon nemen.
BM + COMET & MT-HT	Dat kwam door de atmosfeer van hockeyvelden, koude baden en gemeenschapshikes en algemene properheid die ze met zich mee kon dragen.

Table 4: Example of human-written text (source and human translation), translations of the most relevant baselines (Tagging, base MT model) and our alignment model (BM + COMET & MT-HT).

in MT-HT), as expected, while there are mixed cases in HT-OR. Also as expected, translation accuracy scores decrease when the naturalness reward is added (except COMET with MT-HT).

Compared to models using classifier-only reward, classifier + COMET results generally in better naturalness-related metrics (except PTF), but worse content-based metrics (except COMET with MT-HT). This might be due to a mismatch between the classifier's objective and the training data (see reasoning in Section 6.2) and to complex interactions between both rewards, that would require further inspection.

6.4 Finer-grained Analysis

Surface-level Inspection In Table 4, we compare the surface-level output of the strongest baseline (Tagging; 4.8M) with that of the base MT model and our best alignment system (COMET & MT-HT). As highlighted in green, the English 'community hikes' is translated to *gemeenschapsfietsen* ('community bicycles') by the Tagging system, while our alignment system outputs *gemeenschapshikes* ('community *hikes*'). This is an example of how the Tagging model output may score high on lexical diversity metrics, but strays from the content, where our model preserves it. As shown in blue, 'general clean-mindedness' is translated to *algehele schoonheid* ('overall beauty') by the base MT system. Our alignment system translates to *algemene properheid* ('general cleanliness'), while the Tagging system outputs *algemeene properheid*. The latter case contains a double *e*, which is not typical in this context for modern Dutch, but does appear in the original Dutch dataset. Our alignment MT system is not affected by this.

Book-level Comparison Figure 4 shows MTLD scores per book between human translation, base MT model, and our best alignment model (COMET + MT-HT). We observe that COMET + MT-HT scores are higher than the base MT model for all books, indicating that our alignment method makes the translations more lexically diverse. It is interesting to see that our method brings the results closer to or even exceeds human translation in terms of

	Translation Accuracy				Classification Accuracy			Lexical Diversity					
	BLEU	COMET	KIWI	MetricX↓	HT-OR	MT-HT	MT-OR	TTR	Yule's I	MTLD	B1↓	PTF↓	CDU↓
Human Translation	-	-	-	-	32.9	69.3	48.6	0.153	3.934	96.0	0.672	0.817	0.548
BM: Base MT Model	32.5	82.3	80.4	2.66	28.1	18.9	17.6	0.150	3.537	90.4	0.677	0.826	0.563
BM + COMET & HT-OR	21.8	78.0	77.5	3.59	43.5	48.4	42.8	0.138	2.859	88.0	0.674	0.848	0.527
BM + COMET & MT-HT	24.1	81.3	79.6	3.06	27.0	52.2	34.6	0.121	2.265	92.4	0.683	0.849	0.547
BM + COMET & MT-OR	24.4	80.5	79.8	3.19	32.2	59.2	49.5	0.139	3.084	93.1	0.669	0.845	0.526

Table 5: Translation performance with β set to 0.0, where models are trained without the constraint of negative log-likelihood loss.

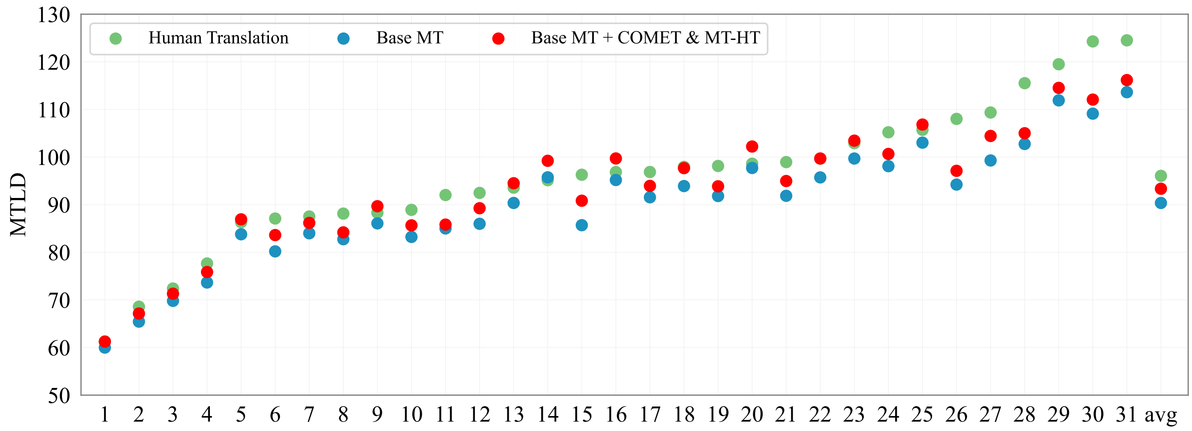

Figure 4: Per-book comparison of MTLD under human translation, base MT model, and alignment model. Note that avg presents the average score across all books.

lexical diversity on some books (e.g. 5, 9, 14, and 16). Overall, the MTLD scores of the alignment models are between those of the base MT model and human translation.

6.5 Impact of Hyper-parameter

To examine the impact of hyper-parameter β (see Section 4.3), we report the results when it is set to 0.0, i.e. only considering the reward learning. Models trained without the constraint of negative log-likelihood loss lead, as expected, to worse content scores across the board as they move too far from the base MT model (Table 5). These models achieve better classification scores but worse naturalness results (except MTLD, B1, and CDU in MT-OR and MTLD in MT-HT). The higher scores on classifiers could be due to characteristics of translated language beyond those related to high lexical diversity. Future work is needed to determine how the classifiers, lexical diversity, machine translationese and naturalness are precisely related.

7 Conclusion

We proposed a reinforcement learning based alignment framework for machine translation, which improves translation quality from multiple perspectives. Using the evaluation model COMET and different binary translationese classifiers trained with MT, HT, and original target-language data as reward models, we approximate human preference and align the MT model with it. Our experiments on English-to-Dutch literary translation show that our model produces translations that are lexically richer and more natural without loss in translation accuracy.

Limitations

Due to the computational resources required, we were only able to perform extensive experiments on one language pair and domain. Since we first wanted to show that our method is sound in a simple setting, i.e. training a model from scratch, we have not proceeded to involve complex settings and computationally-heavy models, such as pre-trained large language models. Furthermore, our metrics for evaluating naturalness are mostly limited to lexical diversity, while writing style in general is much broader and difficult to capture with automatic metrics. We acknowledge that large-scale human evaluation, beyond our surface-level inspection in Section 6.4, could bring important insights.

Acknowledgments

The anonymous reviewers of ACL Rolling Review provided us with useful comments which contributed to improving this paper and its presentation, so we're grateful to them. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the high-performance computing cluster Hábrók. Huiyuan Lai is supported by the Sector Plan in the Humanities "Humane AI" funding from the Dutch Ministry. Antonio Toral is supported by a Beatriz Galindo senior fellowship (BG23/00152) from the Spanish Ministry of Science and Innovation.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology*. John Benjamins.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Malina Chichirau, Rik van Noord, and Antonio Toral. 2023. Automatic discrimination of human and neural machine translation in multilingual scenarios. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 217–226, Tampere, Finland. European Association for Machine Translation.

Miguel A Jiménez Crespo. 2023. "translationese"(and "post-editese"?) no more: on importing fuzzy conceptual tools from translation studies in mt research. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 261–268.

Orphée de Clercq, Rudy Loock, Gert de Sutter, Bert Cappelle, and Koen Plevoets. 2020. Uncovering Machine Translationese: an experiment on 4 MT systems for English-French translations. In *Journée d'études #TQ2020 "Traduction & Qualité : biotraduction et traduction automatique"*, Roubaix, France. Université de Lille & UMR STL du CNRS.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Dirk Delabastita. 2011. Literary translation. *Handbook of translation studies*, 2:69–78.

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.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In *Translation studies in Scandinavia: Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, 75, page 88–95.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021a. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021b. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Philip M McCarthy. 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

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

Lili Mou and Olga Vechtomova. 2020. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–22, Online. Association for Computational Linguistics.

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.

Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3205–3227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Esther Ploeger, Huiyuan Lai, Rik Van Noord, and Antonio Toral. 2024. Towards tailored recovery of lexical diversity in literary machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 286–299, Sheffield, UK. European Association for Machine Translation (EAMT).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arxiv:2207.04672.

M.C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*. Child Welfare Monograph Series. University of Minnesota Press.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral, Andreas van Cranenburgh, and Tia Nutters. 2021. Literary-adapted machine translation in a well-resourced language pair. In *Proceedings of the 7th Conference of The International Association for Translation and Intercultural Studies (IATIS). Barcelona, Spain.*, pages 27–52, Online.

Antonio Toral and Andy Way. 2015. Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4(2):240–267.

Gideon Toury. 2012. Descriptive translation studies-and beyond. revised version. *Amsterdam and Philadelphia: John Benjamins Publishing Company*.

Tobias van der Werff, Rik van Noord, and Antonio Toral. 2022. Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium. European Association for Machine Translation.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using Abstract Meaning Representation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–765, St. Julian's, Malta. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

G. Udnv Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Test Set Novels

ID	Author	Title	Year Published	Genre
1	Patricia Highsmith	Ripley Under Water	1991	Thriller, suspense
2	J.D. Salinger	The Catcher in the Rye	1951	Literary fiction
3	Mark Twain	Adventures of Huckleberry Finn	1884	Literary fiction
4	John Steinbeck	The Grapes of Wrath	1939	Literary fiction
5	John Boyne	The Boy in the Striped Pyjamas	2006	Historical fiction
6	Nicci French	Blue Monday: A Frieda Klein Mystery	2011	Thriller, suspense
7	Philip Roth	The Plot Against America	2004	Political fiction
8	Paul Auster	Sunset Park	2010	Literary fiction
9	Khaled Hosseini	A Thousand Splendid Suns	2007	Literary fiction
10	George Orwell	1984	1949	Literary fiction
11	John Irving	Last Night in Twisted River	2009	Literary fiction
12	E.L. James	Fifty Shades of Grey	2011	Erotic thriller
13	Jonathan Franzen	The Corrections	2001	Literary fiction
14	Stephen King	11/22/63	2011	Science-fiction
15	Oscar Wilde	The Picture of Dorian Gray	1890	Literary fiction
16	John Grisham	The Confession	2010	Thriller, suspense
17	William Golding	Lord of the Flies	1954	Literary fiction
18	Irvin D. Yalom	The Spinoza Problem	2012	Historical fiction
19	J.R.R Tolkien	The Return of the King	1955	Fantasy
20	David Baldacci	Divine Justice	2008	Thriller, suspense
21	Julian Barnes	The Sense of an Ending	2011	Literary fiction
22	James Patterson	The Quickie	2007	Thriller, suspense
23	Sophie Kinsella	Shopaholic and Baby	2007	Popular literature
24	J.K. Rowling	Harry Potter and the Deathly Hallows	2007	Fantasy
25	John le Carré	Our Kind of Traitor	2010	Thriller, spy fiction
26	Jack Kerouac	On the Road	1957	Literary fiction
27	Karin Slaughter	Fractured	2008	Thriller, suspense
28	Ernest Hemingway	The Old Man and the Sea	1952	Literary fiction
29	David Mitchell	The Thousand Autumns of Jacob de Zoet	2010	Historical fiction
30	James Joyce	Ulysses	1922	Literary fiction
31	Thomas Pynchon	Gravity's Rainbow	1973	Historical fiction

Table 6: Information on test set books.

A.2 Post-processing Examples

Original Outputs	Post-processed outputs
Bijna een jaar lang heeft hij foto's genomen van verlaten dingen............	Bijna een jaar lang heeft hij foto's genomen van verlaten dingen.
Ongetwijfeld mag hij blij zijn dat hij deze baan heeft gevonden........	Ongetwijfeld mag hij blij zijn dat hij deze baan heeft gevonden.
In het begin was hij verbijsterd door de wanorde en de vuiligheid, de verwaarlozing.............	In het begin was hij verbijsterd door de wanorde en de vuiligheid, de verwaarlozing.

Table 7: Post-processing examples.