Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.058248

ARTICLE

Improving Machine Translation Formality with Large Language Models

Murun Yang[1,*] **and Fuxue Li**[2]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, 110819, China

[2]College of Electrical Engineering, Yingkou Institute of Technology, Yingkou, 115014, China

*Corresponding Author: Murun Yang. Email: yangmurun@outlook.com

Received: 08 September 2024 Accepted: 18 November 2024 Published: 17 February 2025

ABSTRACT

Preserving formal style in neural machine translation (NMT) is essential, yet often overlooked as an optimization objective of the training processes. This oversight can lead to translations that, though accurate, lack formality. In this paper, we propose how to improve NMT formality with large language models (LLMs), which combines the style transfer and evaluation capabilities of an LLM and the high-quality translation generation ability of NMT models to improve NMT formality. The proposed method (namely INMTF) encompasses two approaches. The first involves a revision approach using an LLM to revise the NMT-generated translation, ensuring a formal translation style. The second approach employs an LLM as a reward model for scoring translation formality, and then uses reinforcement learning algorithms to fine-tune the NMT model to maximize the reward score, thereby enhancing the formality of the generated translations. Considering the substantial parameter size of LLMs, we also explore methods to reduce the computational cost of INMTF. Experimental results demonstrate that INMTF significantly outperforms baselines in terms of translation formality and translation quality, with an improvement of +9.19 style accuracy points in the German-to-English task and +2.16 COMET score in the Russian-to-English task. Furthermore, our work demonstrates the potential of integrating LLMs within NMT frameworks to bridge the gap between NMT outputs and the formality required in various real-world translation scenarios.

KEYWORDS

Neural machine translation; formality; large language model; text style transfer; style evaluation; reinforcement learning

Nomenclature

Translation Formality	the generated translations in a formal style

1 Introduction

Ensuring that generated translations possess a formal style is a critical requirement in machine translation tasks [1–3]. However, despite significant advances in neural machine translation (NMT), the balance between translation quality and stylistic control remains a challenging issue. Traditional NMT models often prioritize translation accuracy over the formality of generated text, as they do not model translation style into their optimization objectives during training [4]. This oversight can lead

to translations that, while semantically correct, lack the desired formality, resulting in content that is inappropriate for formal settings. For instance, in the real-world scenario of translating French to English, "Comment ça va?" may be translated as "What's up?". Although this translation is accurate, it bears some colloquialism and lacks formality. Actually, we prefer it is translated as "How are you?", a more formal and accurate translation.

In response to this issue, recent works have attempted to train formal style-aware neural machine translation (NMT) models [5], which prioritize generating more formal translations. However, these efforts face a significant challenge in that they require a substantial size of manually annotated bilingual sentence pairs for implementation. Additionally, some works have tried to leverage the strong context learning ability of large language models to control the translation style via human language prompting [3]. Despite some degree of style control, large language models (LLMs) have been shown not to match the translation quality of systems trained on proprietary bilingual data [6].

Given these limitations, we propose a novel approach, named Improving NMT Formality with Large Language Models (INMTF). Our INMTF method leverages the style transfer and evaluation capabilities of LLMs, combined with the high-quality translation generation ability of NMT models, to enhance the formality of translations. By introducing this method, we aim to bridge the gap between the formality often lacking in NMT outputs and the formality demanded in various real-world translation scenarios. Our INMTF encompasses two main approaches. The first is to propose a revision approach for the formality of translation. This approach uses an LLM to further revise the translation generated by the NMT, ensuring the final translation style is formal. The second approach, inspired by [7], attempts to utilize the style evaluation capability of the LLMs to improve the formality of translations. Specifically, we employ an LLM as a reward model for scoring the formality of translations. We then use Reinforcement Learning (RL) algorithms, such as REINFORCE [8] and Proximal Policy Optimization (PPO) [9], to fine-tune the NMT model to maximize the reward score, making the style of the generated translations more formal. Furthermore, considering the substantial parameter size of LLM could lead to high computational costs in practical applications, we further investigate how to reduce the computational burden of our method. Here we attempt to transfer the style transfer ability and style evaluation capability of the LLMs to a more lightweight language model (LM), thus reducing the model parameters and achieving acceleration.

We conducted comprehensive experiments to validate the effectiveness of the proposed INMTF on German-to-English (De-En) and Russian-to-English (Ru-En) translation tasks. The experimental results demonstrate that INMTF outperforms all baselines in terms of translation formality. Notably, our reward-based INMTF approach achieves an improvement of +9.19 style accuracy points in the De-En translation task. Additionally, INMTF offers benefits in translation quality; for example, compared to the MLE system, the reward-based INMTF approach gains a +2.16 COMET score on the Ru-En translation task.

2 Related Work

The recent research pertinent to this paper revolves around stylized machine translation, text style transfer, and text style evaluation.

2.1 Stylized Neural Machine Translation

Researchers have gradually realized that in addition to translation accuracy, the style of translation is an equally significant factor. Style pertains not only to the formality of the text but also to the regional and temporal characteristics of the language, as well as the personal traits of the author.

Previous works have focused on training a style-controlled machine translation model, such as the gender and speaker style controlled machine translation models trained by [10] and [11]. However, a core challenge faced by stylized translation is the lack of training data with style labels. To address this issue, several solutions have been proposed. Reference [3] introduced the StyleAP method, which controls translation style by retrieving prompts from a stylized monolingual corpus. The key to this method is that it does not require additional fine-tuning of the model, but rather leverages the model's latent generative capabilities to achieve style transfer. This method provides a solution for stylized translation without the need for extensive labeled data. Additionally, Niu et al. [1] proposed Online Style Inference (OSI), a method that dynamically predicts the formality level of translation pairs during training. Reference [2] approached the problem from a different angle, proposing the Iterative Dual Knowledge Transfer (IDKT) framework. This framework generates large-scale stylized paired data by facilitating bidirectional knowledge transfer between the machine translation model and the text style transfer model. In this work, we attempt to leverage the text style transfer and evaluation capabilities of LLMs to improve the style control of NMT.

2.2 Text Style Transfer and Evaluation

As a key task in the field of Natural Language Processing, text style transfer aims to transform sentence styles, such as from formal to informal, while ensuring that the original meaning of the sentence is preserved. A basic and common approach to solving this task is to directly train a sequence-to-sequence model using labeled sample pairs <source style, target style>, a method widely used in previous research [12,13]. However, this method faces a significant challenge: training an efficient style transfer model becomes particularly difficult when annotated data for specific style transfers are scarce. To address this issue, researchers have fine-tuned pre-trained generative models, such as GPT-2 [14] and BART [15], using small datasets. LLMs have consistently demonstrated exceptional performance in text transfer tasks, often achieving state-of-the-art (SoTA) results and offering new solutions for text style transfer [16]. Evaluating text style is also a critical area of research. For example, Lai et al. [15] developed new evaluation metrics and enhanced the text style transfer model using Reinforcement Learning (RL) techniques to optimize these metrics. Additionally, recent studies have shown that LLMs possess significant capabilities for evaluating text style transfer [17].

3 Preliminaries

3.1 Machine Translation

Given an input source x, a NMT model generates a translated text $y = \{y_1, y_2, \ldots, y_T\}$ with T tokens. Each token y_t is derived from a predefined vocabulary. During the training phase, the translation model learns a probability distribution:

$$p_\theta(y|x) = \prod_{t=1}^{N} p_\theta(y_t|y_{<t}, x) \tag{1}$$

where $y_{<t}$ is the prefix $\{y_1, y_2, \ldots, y_{t-1}\}$, and θ is the set of model parameters. In this process, the conventional training objective is to maximize the likelihood of all tokens in the target sequence, i.e., maximum likelihood estimation (MLE) [4]. During the inference phase, tokens are generated sequentially according to p_θ.

3.2 Reinforcement Learning

The RL objective for NMT model is to maximize the long-term reward, written as $\arg\max_\theta \mathbb{E}_{p_\theta}(\hat{y}|x)[r(\hat{y})]$, where $\hat{y}$ is the generated translation, and $r(\cdot)$ is the reward function that computes the long-term reward of $\hat{y}$. $r(\cdot)$ is often defined as an evaluation metric, such as BLEU [18]. To achieve this objective, we typically use policy gradient methods, such as REINFORCE [8], Proximal Policy Optimization (PPO) [9], and Minimum Risk Training (MRT) [19]. In particular, REINFORCE uses log derivatives to define the loss function:

$$\mathscr{L}_{\text{REINFORCE}} = -\sum_{\hat{y}\in S(x)} \log p_\theta(\hat{y}|x) r(\hat{y}) \tag{2}$$

where $S(x)$ is an approximation of the sampling space, consisting of sampled sequences. Furthermore, MRT utilizes these sampled sequences to approximate the posterior distribution through re-normalization, and provides a new loss function:

$$\mathscr{L}_{\text{MRT}} = \sum_{\hat{y}\in S(x)} Q_\theta(\hat{y}|x)[-r(\hat{y})] \tag{3}$$

where $Q_\theta(\hat{y}|x)$ is a distribution defined over the approximation space, which can be defined as:

$$Q_\theta(\hat{y}|x) = \frac{p_\theta(\hat{y}|x)^\alpha}{\sum_{\hat{y}\in S(x)} p_\theta(\hat{y}|x)^\alpha} \tag{4}$$

where α is a smoothness parameter. Based on the posterior distribution, MRT can achieve better performance compared to REINFORCE [20,21].

4 Our Method

In this work, our aim is to improve the NMT formality using LLMs. We propose the INMTF method for this purpose. The overall method is illustrated in Fig. 1. As depicted in the figure, our method leverages the text style transfer and evaluation capabilities of an LLM to improve the translation formality. In the following subsections, we provide a detailed description of these components in INMTF.

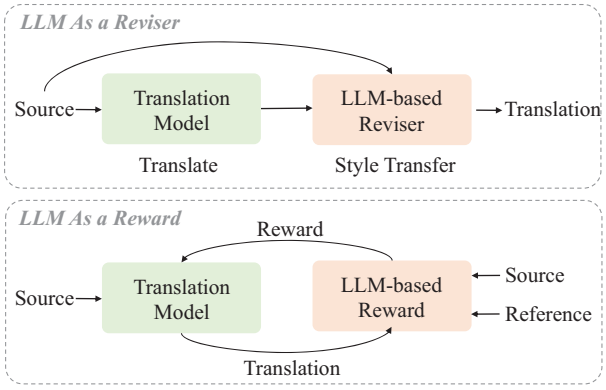

Figure 1: The overview of INMTF. In INMTF, we endeavor to harness the capabilities of LLMs for text style transfer and evaluation, with the aim of improving the NMT formality

4.1 Employing Large Language Model as a Reviser

During the training phase of NMT models, it does not directly model style optimization objectives. Consequently, it often overlooks the style of its translations during actual generation. Here, we contemplate whether we can further revise the style based on the generated translations. Specifically, we treat this style revision as a text style transfer task. That is, we transform the style of the translation into a formal one while retaining all the content meanings in the original translation. Given this setup, we attempt to employ an LLM as a reviser, utilizing its robust text style transfer capability to correct the style of the translation.

Fig. 2 (left) illustrates the prompt we designed. In the prompt, we add the source language for further constraint, preventing the LLM from changing the semantic information in the corresponding translation during the style transfer process. Moreover, inspired by the context learning ability of LLMs [22], we introduce a few-shot learning mechanism, i.e., adding some corresponding demonstrations in the prompt to improve the performance of the LLM in text style transfer.

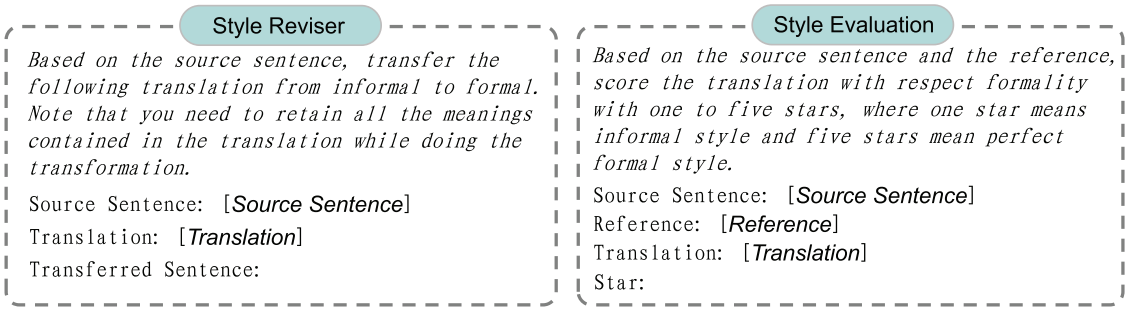

Figure 2: The designed templates for text style revise (left) and text style evaluation (right)

4.2 Large Language Model as a Reward Model

An ideal approach would be for NMT to rely on itself to generate more formal translations during inference. Considering the LLM's strong evaluation capability for text style, we speculate we can use an LLM as a reward model and then fine-tune the machine translation using RL algorithms to maximize the reward, thereby improving the NMT to generate more formal translations on its own. We employ a policy gradient-based method for fine-tuning our translation model. The style evaluation prompt we designed is shown in Fig. 2 (right).

Similar to [7], we fine-tune the NMT model using a loss function that combines RL loss and MLE loss, as follows:

$$\mathscr{L}_{\text{Weighted}} = \lambda \times \mathscr{L}_{\text{RL}} + (1 - \lambda) \times \mathscr{L}_{\text{MLE}} \tag{5}$$

where λ is a balance factor adjusted on the validation set. Here, we use Eq. (3) and MRT to calculate $\mathscr{L}_{\text{RL}}$. In fact, whether it is a reviser or acts as a reward model in RL, we have provided certain constraints to prevent the overfitting problem. For example, in reviser, we require LLM to ensure that the original meaning remains unchanged, and in RL, we also add MLE loss for constraints. Here, we can guarantee our translation instructions to a certain extent, and the experimental results also shows that our COMET and BLEU are not particularly depleted due to our modification of the style.

When fine-tuning with reinforcement learning, training instability can often occur in practice [23]. To mitigate this issue, we extend the baseline reward technique proposed in [23]. Specifically, we use a FIFO (First In, First Out) based baseline reward method, which uses a FIFO reward queue $\mathscr{Q}$ to

calculate the baseline value. We denote the size of the reward queue as $\mathcal{Q}_{size}$. At each training step, we push the rewards of all sampled sequences into $\mathcal{Q}$ and pop out the "oldest" reward. Then, we calculate the average reward in $\mathcal{Q}$ as the baseline value b. By using this baseline reward, we replace the reward function in Eq. (3) with $r(\hat{y}, y) - b$.

4.3 Enhancing Efficiency

From the exposition of the aforementioned methods, it is apparent that LLMs are frequently called in our approach. However, one problem is that utilizing large LLMs for both revision and reward modeling can result in significant computational overhead, making the approach resource-intensive and less scalable, especially for real-time translation systems. Inspired by [7], we endeavor to transfer the text style transfer and evaluation capabilities of LLMs to comparatively smaller LMs for utilization in our methodology to solve the problem. Regarding text style transfer capability, we employ knowledge distillation techniques [24,25] to migrate the text style transfer competence of LLMs to GPT-2 [26]. For text style evaluation ability, we leverage the ECT method proposed by [7], to transfer the evaluation capacity of an LLM to RoBERTa [27].

5 Experiments

We conduct experiments to evaluate this proposed INMTF method on German-to-English (De-En) and Russian-to-English (Ru-En) translation tasks.

5.1 Datasets

Initially, we trained a neural machine translation (NMT) model using the Transformer base model [28]. The training data consisted of the WMT2023 dataset, which includes the Common Crawl corpus and News Commentary v18.1. Table 1 presents the statistics of this dataset. We preprocessed the dataset following the approach of [29]. For testing, we utilized the informal-to-formal test set from GYAFC [30], where both the source and target sentences are in English. To create the source-to-target language formality test sets, we translated the original English sentences into German and Russian using Google Translate. Indeed, various stylistic elements, such as formality and intonation, need to be controlled in machine translation. In this paper, we focus on controlling formality in translations to evaluate the effectiveness of our approach. In future work, we will extend our method to control additional stylistic types.

Table 1: Statistics of De-En and Ru-En corpora

Dataset	De-En		Ru-En	
	#sentences	#EN words	#sentences	#EN words
Common crawl	5,377,911	101,312,154	878,386	18,772,833
News commentary v18.1	388,482	8,554,360	331,508	7,668,112

5.2 Training Setups

5.2.1 Training Machine Translation Models

For the machine translation models, we directly trained a standard transformer base [28,31] using MLE until convergence. The translation model was trained for 40 epochs, employing a batch size of

4096 (token-level). In data preprocessing, we used the Byte Pair Encoding (BPE) method, with a merge count of 32,000. During this preprocessing phase, we filtered out samples with the source language length exceeding 250. All the experiments were conducted on four TITAN V GPUs.

5.2.2 Reinforcement Learning

For Reinforcement Learning (RL) training, we initialized a sequence generation model using the MLE checkpoint that had the lowest validation set loss. During the RL training process, we generated five candidate sequences for each source input sample. The balancing factor λ in Eq. (5) was set to 0.7. For the De-En and Ru-En translation tasks, we conducted 10 epochs and 8000 steps of training on the pre-trained model, respectively, with a batch size of 4096 tokens (token-level). We utilized top-k sampling, generating five candidate samples for each input sample for training.

5.2.3 Transferring Capabilities from an LLM

We selected ChatGPT[1] as our LLM. When using ChatGPT, we set the temperature to 0, and the max length to 1024. During the generation process, we used top-p sampling method, where p was set to 0.95. It was fine-tuned using RLHF [32] and has demonstrated excellent performance in sequence generation evaluation [33]. For transferring the text style transfer capability, we selected 15 K training samples from the training set. We then generated translations using our translation model, and let ChatGPT perform text style transfer, thus forming the corresponding pairs of the source sentences and the target style sentences. These data pairs were then used to fine-tune the GPT-2 model, completing the transfer of text style transfer capability from the LLM. During fine-tuning, we used a learning rate, batch size, and epoch size of 1e-5, 32, and 3, respectively.

For transferring text style evaluation capability, we designed an evaluation focused on formality. These designed evaluation prompts are described in Fig. 2. For data collection, we randomly selected 15 K training samples from the training set and generated five output sequences for each input. For the evaluation model architecture, we used RoBERTa-base as the encoder model. We trained the evaluation model according to the COMET framework[2]. The learning rate, maximum epoch size, and batch size were set to 1e-5, 10, and 32, respectively. We trained all the evaluation models on one TITAN V GPU with 16-bit floating-point precision and applied early stopping during the training process.

5.3 Evaluation

In this work, we use two approaches to evaluate the effectiveness of the proposed methods. One is by testing the accuracy of the translation model. Here, we compute the BLEU scores between the translations and all the given references in the test set[3]. Furthermore, we report COMET-22 scores between the translations and the first reference answer [34]. For the aspect of style formality, we use a pretrained style classification model to score the style of translations, similar to the studies by [15] and [35]. This classifier can reach a consistency of 87% with previous human accuracy classifications. Additionally, we use ChatGPT to further score the style of translations.

5.4 Baselines

Our baseline is the standard **MLE**, which is the NMT model trained directly on WMT. We compare results from direct translation with ChatGPT and controlled style translation in the prompt,

[1] https://openai.com/blog/chatgpt (accessed on 17 November 2024).
[2] https://github.com/Unbabel/COMET (accessed on 17 November 2024).
[3] We use multi-bleu.perl for computing BLEU scores.

referred to as **ChatGPT** and **ChatGPT-style**, respectively. Additionally, we report comparisons between using ChatGPT directly and using transfer capabilities from ChatGPT as the reviser and reward model. Furthermore, we compare our INMTF with other mainstream large language models, including ChatGLM3-6b [36], LLaMA2-7b-Chat, and LLaMA2-3b-Chat [37]. We also compare our INMTF with the method that fine-tunes the NMT model against a classification model.

5.5 Main Results

Our experimental results are summarized in Table 2. In terms of translation quality, as can be seen, our INMTF achieves a translation quality comparable to that of MLE, demonstrating that the adjustment of the translation style does not compromise its quality. Furthermore, our method surpasses MLE in terms of COMET and BLEU scores. Notably, on the Ru-En translation task, INMTF-Reward can yield +2.16 COMET score improvement compared to MLE. We conjecture that due to the consideration of translation style in COMET evaluation, a more formal style may lead to a higher COMET score. As for translation style, our method consistently outperforms MLE across all the translation tasks. For instance, on the De-En translation task, our INMTF-Reward outperforms MLE by 9.19 points on the Accuracy metric. Comparing our INMTF with the translations produced by ChatGPT, we find that although ChatGPT controls the style well, there is a decrease in translation quality, e.g., ChatGPT-style loses 2.98 BLEU points on the De-En translation task. We attribute this loss to ChatGPT's less optimal translation capability compared to specialized NMT models [6]. When comparing to the MLE-Class baseline, we find that using a classification model to act as a reward model is inferior. We assume that there are two main reasons: one is that this classification model's capability is poor, and the other is that this classification model does not consider the reference when performing the style classification.

Table 2: The results of our INMTF on De-En and Ru-En translation tasks. The suffix "-Class" denotes that we use a classification model to act as a reward model. The suffixes "-Reviser" and "-Reward" denote the use of an LLM serving as a reviser and a reward model, respectively. We report scores evaluated by ChatGPT in the "ChatGPT" column. The best results for each group are **bolded**

System	WMT De-En				WMT Ru-En			
	BLEU	COMET	Accuracy	ChatGPT	BLEU	COMET	Accuracy	ChatGPT
MLE	52.37	76.58	84.51	3.65	50.86	74.62	80.12	3.23
MLE-Class	51.54	73.21	87.61	3.79	47.65	72.10	84.32	3.56
ChatGPT	48.11	72.34	82.90	3.41	48.12	72.55	76.49	3.01
ChatGPT-style	49.39	74.80	90.87	4.02	49.83	72.87	85.93	3.98
ChatGLM3-6b	40.56	68.01	78.43	2.97	42.10	66.78	74.48	2.89
ChatGLM3-6b-style	42.33	68.75	79.65	3.65	43.16	69.34	75.56	2.99
LLaMA2-7b-Chat	43.12	69.88	79.64	2.79	45.24	69.70	73.94	3.50
LLaMA2-7b-Chat-style	44.23	70.05	80.98	3.18	47.84	71.79	76.74	3.55
LLaMA2-13b-Chat	45.40	71.56	80.15	3.15	49.83	72.16	74.58	3.23
LLaMA2-13b-Chat-style	47.97	72.83	80.83	3.27	50.70	73.37	76.88	3.58
Using ChatGPT as Reviser or Reward								
INMTF-Reviser	52.53	78.11	**93.70**	**4.23**	**51.88**	75.89	86.76	4.01
INMTF-Reward	**53.02**	**78.23**	92.91	4.19	51.65	**76.78**	**87.71**	**4.13**

(Continued)

Table 2 (continued)

System	WMT De-En				WMT Ru-En			
	BLEU	COMET	Accuracy	ChatGPT	BLEU	COMET	Accuracy	ChatGPT
Transferring capabilities from ChatGPT								
INMTF-Reviser	51.98	78.01	92.98	4.17	51.26	75.10	85.90	3.95
INMTF-Reward	52.83	78.21	92.02	4.10	51.08	76.53	86.78	4.10

When comparing the direct use of ChatGPT and its transferred use, we can observe that they have similar effectiveness in improving the NMT formality. However, transferring from ChatGPT significantly reduces computational costs. Furthermore, we further investigate the performance gain on two different approaches to improving the NMT formality. From the results, we can find that both the reviser-based and reward-based approaches can effectively improve the style across various translation tasks. The comparison results indicate that the ability of the reviser or forward model positively correlates with our method's performance. Specifically, the stronger these models are, the better our method performs. Ideally, using a powerful model like GPT-4 would yield optimal results; however, such models are often resource-intensive. Therefore, in practice, it is advisable to select the strongest models available as the reviser or reward model whenever possible.

5.6 Analysis

5.6.1 Transferring Capabilities with Different Data Sizes

The scale of data used is an important factor when transferring the capability of LLMs. We aim to explore the impact of different dataset sizes on transfer. Specifically, we create datasets with {5, 10, 15, 20, 25 K} samples for both text style transfer and evaluation. For the transfer of text style transfer capability, we employ the test set's PPL for measurement. For text style evaluation capabilities, following [7], we measure the consistency between the transferred evaluation model and ChatGPT. Note that our test set is randomly selected from the training set, excluding samples used during the transfer process. Fig. 3 includes the corresponding test results, showing a long-tail distribution for both transfers. To balance performance and transfer costs, we opt to use 15 K data for both style transfer and evaluation transfer.

5.6.2 Performance on Different Temperature Settings

In real-world applications, varying temperature settings are typically employed by the LLM. Consequently, we have also evaluated the impact of these different temperature settings on our method. For this purpose, we have applied various temperature settings for our INMTF-Reviser and INMTF-Reward, including 0, 0.25, 0.5, 0.75, and 1. The experiments are conducted on WMT De-En translation task. The results are shown in Fig. 4. From the results, we can see that our INMTF consistently achieves favorable outcomes under different temperature settings, surpassing the MLE baseline. This also demonstrates that our INMTF has a well robustness when using different temperature settings.

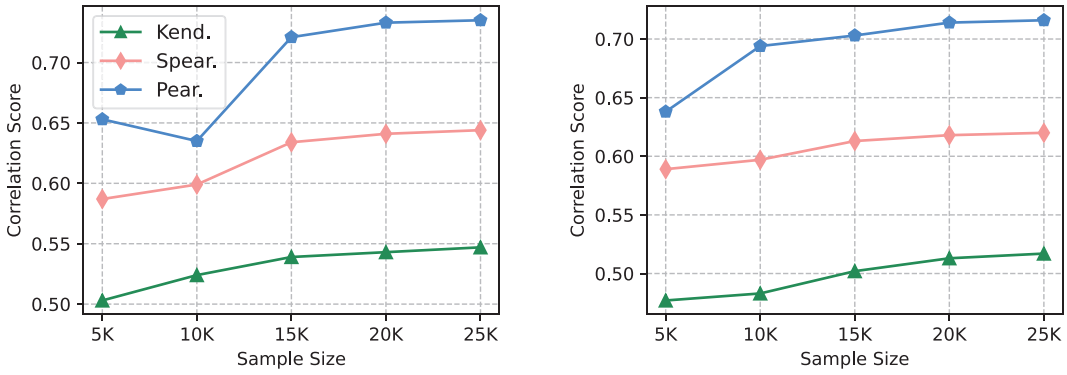

Figure 3: Sample-level Spearman (Spear.), Kendall-Tau (Kend.), and Pearson (Pear.) correlation scores of the transferred models learned employing different sizes of LLM-annotated samples for text style transfer (left) and text style evaluation (right). Note that these two figures share a legend

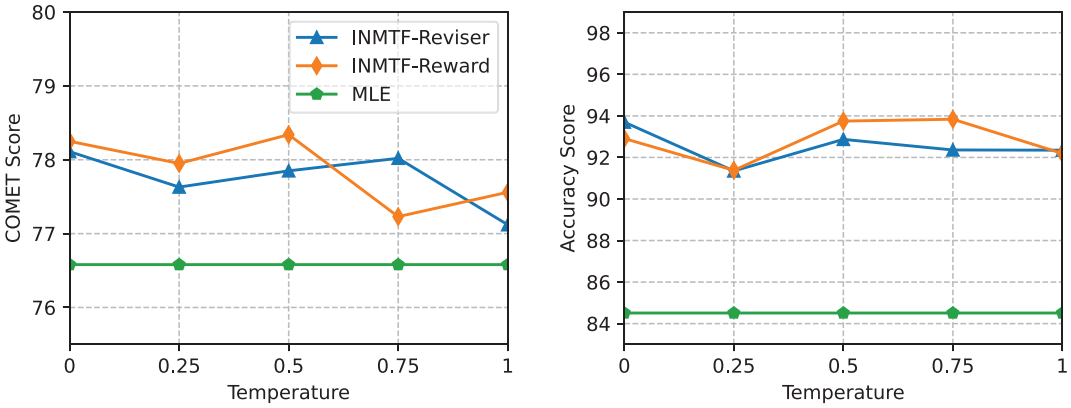

Figure 4: Performance of our INMTF-Reviser and INMTF-Reward with different temperature settings

5.6.3 Performance on Different Sampling Methods

During RL training, continuous sampling is required for exploration. We compare different sampling methods on the WMT De-En task, including greedy search and top-p sampling [38]. The comparative results are shown in Table 3. This comparison shows that top-k sampling yields the best results. We assume that top-k sampling strikes a favorable balance between generation diversity and accuracy, which is beneficial for RL exploration [23]. Note that although top-p yields superior results compared to top-k, we observe that in our tests that the diversity of top-p is not as good as that of top-k. This can potentially lead to poorer exploration in RL, making it less efficient in finding more formal translations.

5.6.4 Combining Reviser-Based and Reward-Based Approaches

We have endeavored to combine our reviser-based and reward-based approaches. Specifically, we initially employ an LLM as a reward model to fine-tune the NMT model using RL. Subsequently, we engage the LLM as a Reviser to revise the translation generated by the NMT, thereby elevating the

formalism of the translated text. We conducted experiments on the WMT De-En translation task. The results of these experiments are delineated in Fig. 5. From the results, we can observe that a combined approach can achieve superior performance in terms of both translation style and quality. However, this comes at the expense of a considerable computational cost. This is attributable to the fact that, in the process of implementation, we need to carry out RL training and also engage an LLM as a Reviser.

Table 3: Performance of INMTF on different sampling methods. The experiments are conducted with INMTF-Reward

Method	BLEU	COMET	Accuracy
Greedy search	50.35	74.71	87.29
Top-*p* sampling	51.14	77.22	90.54
Top-*k* sampling	**52.83**	**78.21**	**92.02**

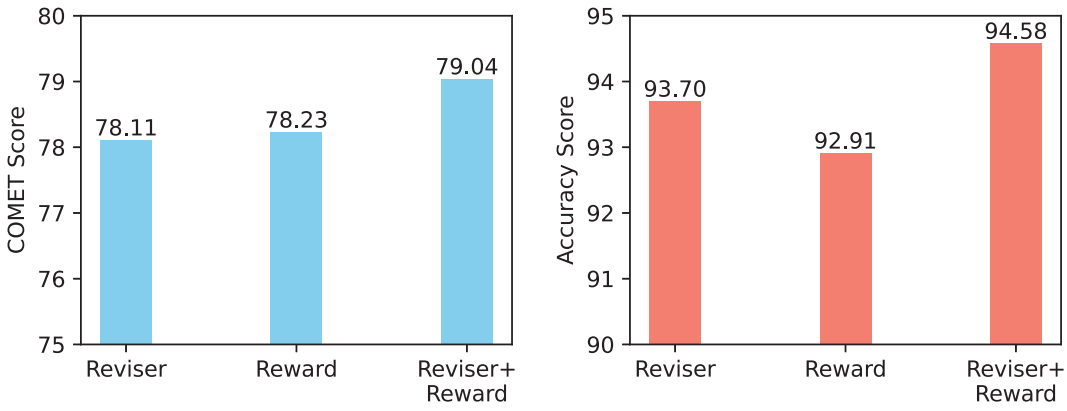

Figure 5: Performance of combining reviser-based and reward-based methods on the De-En translation task

5.6.5 Enhancing Consistency through Flexibility and Extensibility

Relying on LLMs to evaluate formality in translations can introduce subjectivity, as perceptions of formality often vary across cultural contexts, domains, and individual preferences. This variability can lead to inconsistent scoring and evaluation outcomes. However, our method exhibits significant extensibility and flexibility, allowing for the incorporation of additional layers of prior knowledge and constraints to enhance consistency. For example, develop and integrate comprehensive knowledge bases that capture cultural and domain-specific information. These can be used to adjust the formality evaluation process, ensuring that it aligns with the intended context. In Table 2, we focus on the domain of news translation. During the evaluation process, we apply the a priori constraint by adding "This is a new domain...". This addition enhances the precision of the evaluation. We re-execute our experiments. The results of this experiment are shown in Table 4. This experiment demonstrates that adding constraints significantly enhances performance, highlighting the scalability and flexibility of the approach.

Table 4: Performance on INMTF-Reviser with domain constraint

System	WMT De-En				WMT Ru-En			
	BLEU	COMET	Accuracy	ChatGPT	BLEU	COMET	Accuracy	ChatGPT
MLE	52.37	76.58	84.51	3.65	50.86	74.62	80.12	3.23
MLE-Class	51.54	73.21	87.61	3.79	47.65	72.10	84.32	3.56
INMTF-Reviser	52.53	78.11	93.70	4.23	51.88	75.89	86.76	4.01
INMTF-Reviser-addCont	**53.14**	**79.13**	**93.98**	**4.45**	**52.00**	**75.93**	**87.12**	**4.23**

5.6.6 Performance on Low-Resource Machine Translation

In this section, we test the applicability of our proposed method in low-resource MT environments, specifically focusing on Hebrew to/from English translation. This analysis is crucial in addressing the concerns regarding the reliance on LLMs in resource-constrained settings. As a morphologically rich language with limited parallel corpora, Hebrew presents unique challenges in MT. The scarcity of data often leads to suboptimal performance of conventional NMT systems. However, by incorporating LLMs, we aim to mitigate some of these challenges through enhanced style transfer capabilities and more robust evaluation metrics. We conduct experiments on the Hebrew→English[4] translation task. The results are summarized in Table 5. From the results, we can observe that our method still performs remarkably well. The application of our method in the Hebrew-English translation task illustrates its potential to enhance translation quality in low-resource settings.

Table 5: Performance on low-resource machine translation

System	Hebrew → English			
	BLEU	COMET	Accuracy	ChatGPT
MLE	26.73	42.15	74.17	2.86
MLE-Class	26.94	41.81	72.98	2.98
ChatGPT	23.13	45.87	78.94	3.03
ChatGPT-style	22.67	43.11	79.02	3.12
Using ChatGPT as Reviser or Reward				
INMTF-Reviser	26.94	**45.85**	**83.89**	2.93
INMTF-Reward	**27.01**	45.06	83.11	**3.27**
Transferring Capabilities from ChatGPT				
INMTF-Reviser	26.54	43.85	81.21	2.88
INMTF-Reward	27.00	43.93	81.72	2.90

[4] This data and test set are both from WMT2023.

6 Conclusion

In this paper, we highlight the critical importance of maintaining formal style in neural machine translation (NMT) processes. The integration of large language models (LLMs) within the NMT frameworks, as demonstrated by our proposed method INMTF, offers a promising solution to address the formality gap in translations. By leveraging the style transfer and evaluation capabilities of LLMs alongside the translation generation proficiency of NMT models, we have successfully improved the formality of the translated content.

Acknowledgement: We express our sincere gratitude to Dr. Du Quan for providing invaluable feedback throughout the writing and revision process of this paper. Additionally, we extend our appreciation to Shenyang Yayi Network Technology Co., Ltd., for their generous hardware support.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Murun Yang; data collection: Murun Yang; analysis and interpretation of results: Murun Yang, Fuxue Li; draft manuscript preparation: Murun Yang, Fuxue Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in the Eighth Conference on Machine Translation at https://www2.statmt.org/wmt23/translation-task.html (accessed on 17 November 2024).

Ethics Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] X. Niu and M. Carpuat, "Controlling neural machine translation formality with synthetic supervision," in *Thirty-Fourth AAAI Conf. Artif. Intell., AAAI 2020, Thirty-Second Innov. Appl. Artif. Intell. Conf., IAAI 2020, Tenth AAAI Symp. Educ. Adv. Artif. Intell., EAAI 2020*, New York, NY, USA, AAAI Press, Feb. 7–12, 2020, pp. 8568–8575.

[2] X. Wu *et al.*, "Improving stylized neural machine translation with iterative dual knowledge transfer," in *Proc. Thirtieth Int. Joint Conf. Artif. Intell.*, 2021, pp. 3971–3977.

[3] Y. Wang, Z. Sun, S. Cheng, W. Zheng, and M. Wang, "Controlling styles in neural machine translation with activation prompt," 2022, *arXiv:2212.08909*.

[4] I. J. Myung, "Tutorial on maximum likelihood estimation," *J. Math. Psychol.*, vol. 47, no. 1, pp. 90–100, 2003. doi: 10.1016/S0022-2496(02)00028-7.

[5] X. Niu, M. Martindale, and M. Carpuat, "A study of style in machine translation: Controlling the formality of machine translation output," in *Proc. 2017 Conf. Empir. Methods Nat. Lang. Process.*, Copenhagen, Denmark, Association for Computational Linguistics, 2017, pp. 2814–2819.

[6] T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," 2023, *arXiv:2302.14520*.

[7] C. Wang *et al.*, "Learning evaluation models from large language models for sequence generation," 2023, *arXiv:2308.04386*.

[8] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, pp. 229–256, 1992. doi: 10.1007/BF00992696.

[9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[10] E. Rabinovich, R. N. Patel, S. Mirkin, L. Specia, and S. Wintner, "Personalized machine translation: Preserving original author traits," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguist.*, Valencia, Spain, Association for Computational Linguistics, 2017, pp. 1074–1084.
[11] P. Michel and G. Neubig, "Extreme adaptation for personalized neural machine translation," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist.*, Melbourne, Australia, Association for Computational Linguistics, 2018, pp. 312–318.
[12] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence to sequence models," in *Proc. Workshop Stylistic Var.*, Copenhagen, Denmark, Association for Computational Linguistics, 2017, pp. 10–19.
[13] Y. Zhang, T. Ge, and X. Sun, "Parallel data augmentation for formality style transfer," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, Association for Computational Linguistics, 2020, pp. 3221–3228.
[14] Y. Wang, Y. Wu, L. Mou, Z. Li, and W. Chao, "Formality style transfer with shared latent space," in *Proc. 28th Int. Conf. Comput. Linguist.*, Barcelona, Spain, International Committee on Computational Linguistics, 2020, pp. 2236–2249.
[15] H. Lai, A. Toral, and M. Nissim, "Thank you BART! rewarding pre-trained models improves formality style transfer," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Joint Conf. Nat. Lang. Process.*, Association for Computational Linguistics, 2021, pp. 484–494.
[16] B. Li *et al.*, "Deliberate then generate: Enhanced prompting framework for text generation," 2023, *arXiv:2305.19835*.
[17] H. Lai, A. Toral, and M. Nissim, "Multidimensional evaluation for text style transfer using ChatGPT," 2023, *arXiv:2304.13462*.
[18] K. Papineni, S. Roukos, T. Ward, and W. -J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, Philadelphia, PA, USA, Association for Computational Linguistics, 2002, pp. 311–318.
[19] S. Shen *et al.*, "Minimum risk training for neural machine translation," in *Proc. 54th Annu. Meet. Assoc. Comput. Linguist.*, Berlin, Germany, Association for Computational Linguistics, 2016, pp. 1683–1692.
[20] S. Kiegeland and J. Kreutzer, "Revisiting the weaknesses of reinforcement learning for neural machine translation," in *Proc. 2021 Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, Association for Computational Linguistics, 2021, pp. 1673–1681.
[21] D. Donato, L. Yu, W. Ling, and C. Dyer, "Mad for robust reinforcement learning in machine translation," *ArXiv preprint*, 2022.
[22] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, Eds., Curran Associates, Inc., Dec. 6–12, 2020, vol. 33, pp. 1877–1901.
[23] C. Wang *et al.*, "ESRL: Efficient sampling-based reinforcement learning for sequence generation," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 17, pp. 19107–19115, 2024. doi: 10.1609/aaai.v38i17.29878.
[24] N. Ho, L. Schmid, and S. -Y. Yun, "Large language models are reasoning teachers," 2022, *arXiv:2212.10071*.
[25] C. Wang, Y. Lu, Y. Mu, Y. Hu, T. Xiao and J. Zhu, "Improved knowledge distillation for pre-trained language models via knowledge selection," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 6232–6244.
[26] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.
[27] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[28] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., Curran Associates, Inc., Long Beach, CA, USA, Dec. 4–9, 2017, pp. 5998–6008.
[29] C. Hu *et al.*, "RankNAS: Efficient neural architecture search by pairwise ranking," 2021, *arXiv:2109.07383*.

[30] S. Rao and J. Tetreault, "Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer," in *Proc. 2018 Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, New Orleans, Louisiana, Association for Computational Linguistics, 2018, pp. 129–140.
[31] T. Xiao and J. Zhu, "Introduction to transformers: An NLP perspective," 2023, *arXiv:2311.17633*.
[32] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., Long Beach, CA, USA, Dec. 4–9, 2017, pp. 4299–4307.
[33] J. Wang *et al.*, "Is ChatGPT a good NLG evaluator? A preliminary study," 2023, *arXiv:2303.04048*.
[34] R. Rei *et al.*, "COMET-22: Unbabel-IST, 2022 submission for the metrics shared task," in *Proc. Seventh Conf. Mach. Transl. (WMT)*, Abu Dhabi, United Arab Emirates (Hybrid), Association for Computational Linguistics, 2022, pp. 578–585.
[35] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Doha, Qatar, Association for Computational Linguistics, 2014, pp. 1746–1751.
[36] Z. Du *et al.*, "GLM: General language model pretraining with autoregressive blank infilling," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist.*, 2022, pp. 320–335.
[37] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[38] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *8th Int. Conf. Learn. Rep., ICLR 2020*, Addis Ababa, Ethiopia, Apr. 26–30, 2020.