Mitigating Stylistic Biases of Machine Translation Systems via Monolingual Corpora Only

Anonymous ACL submission

Abstract

The advent of neural machine translation (NMT) has revolutionized cross-lingual communication, yet preserving stylistic nuances remains a significant challenge. While existing approaches often require parallel corpora for style preservation, we introduce Babel, a novel framework that enhances stylistic fidelity in NMT using only monolingual corpora. Babel employs two key components: (1) a style detector based on contextual embeddings that identifies stylistic disparities between source and target texts, and (2) a diffusion-based style applicator that rectifies stylistic inconsistencies while maintaining semantic integrity. Our framework integrates with existing NMT systems as a post-processing module, enabling style-aware translation without requiring architectural modifications or parallel stylistic data. Extensive experiments on five diverse domains (law, literature, scientific writing, medicine, and educational content) demonstrate Babel's effectiveness: it identifies stylistic inconsistencies with 88.21% precision and improves stylistic preservation by 150% while maintaining a high semantic similarity score of 0.92. Human evaluation confirms that translations refined by Babel better preserve source text style while maintaining fluency and adequacy. Our implementation and datasets are available at https://anonymous.4open.science/r/Babel-3EB2/.

1 Introduction

Machine translation technology has revolutionized cross-language communication, yet the preservation of stylistic nuances remains a significant challenge. Style, encompassing elements from formality and tone to domain-specific conventions, is crucial for maintaining the intended impact and appropriateness of translated text. Consider these examples of stylistic deviations in translation: when translating formal legal documents from Chinese to English, commercial translation systems often fail to maintain the authoritative tone and standardized legal terminology - translating “甲方应当” (formal legal term for “Party A shall”) as the casual “Party A needs to” rather than the proper legal phrasing ”Party A shall”. Similarly, in literary translation, the poetic style of classical Chinese literature is frequently lost - a line like “春花秋月何时了” (literally “when will spring flowers and autumn moon end”) might be translated prosaically as ”when will the seasons end” rather than preserving its lyrical quality with something like “when shall cease the dance of spring blooms and autumn moons”. When translating Yoda’s dialogues from Star Wars into Chinese, the iconic OSV syntax (“Much to learn, you still have”) is frequently normalized to SVO structures “你还有很多要学习” (“You still have much to learn”), diluting the character’s idiosyncratic speech patterns that are deeply tied to his wisdom and alien identity. Such stylistic flattening not only reduces translation fidelity but also diminishes narrative cohesion and audience immersion.

Several studies have addressed this problem (Hovy et al.), and a few methods have been proposed for style preservation in translation (Hu et al., 2017; Zhang et al., 2018). However, these methods exhibit significant limitations. First, existing translation systems often have a limited scope when it comes to the types of styles they can support, typically offering only a binary distinction between formal and informal styles. This oversimplification fails to account for the rich tapestry of stylistic diversity found in human language. Second, most methods require parallel text data specific to certain languages or domains, which is impractical for many applications because obtaining sufficient parallel corpora is challenging in many real-world scenarios.

We propose Babel, a novel framework that addresses these limitations by enabling style-aware translation without relying on parallel corpora. Babel introduces two key innovations: 1) A style detector utilizing contextual embeddings to identify and characterize stylistic attributes in both source and target languages, trained on monolingual corpora; and 2) A diffusion-based style applicator that can modify translated text to match source text style while preserving semantic content, guided by user-provided style examples.

To evaluate our approach, we construct Babel-Corpus, a comprehensive evaluation dataset spanning five diverse domains: law, literature, scientific writing, medicine, and educational content. The corpus focuses on Chinese-English translation, motivated by the significant need for accurate style preservation between these widely-used languages - while over one billion people speak each language, less than 1% of Chinese speakers are proficient in English (Fishman, 2020; chi), making machine translation both essential and challenging. Extensive experiments demonstrate that Babel effectively identifies stylistic inconsistencies in commercial translation systems with 88.21% precision, as verified through human evaluation. The framework improves stylistic consistency by 150% while maintaining semantic fidelity, achieving an average similarity score of 0.92. Results indicating that Babel enhances translation quality across different domains and styles, offering style-aware machine translations.

Our main contributions include:

- The first framework for style-aware translation that operates without parallel corpora, significantly expanding the practical applicability of stylistic translation.
- A novel approach combining style detection and diffusion-based style application for translation refinement.
- The Babel-Corpus dataset, facilitating research in style-aware translation.
- Comprehensive evaluation demonstrating significant improvements in stylistic consistency across domains and translation systems.

2 Background

2.1 Text Style

The concept of style in text refers to the distinct manner in which semantics are expressed, shaped by individual characteristics and pragmatic protocols (Jin et al.). Style is inherent to personal utterances and can manifest through various stylistic devices, such as metaphors, word choices, and syntactic structures. According to Kang and Hovy, style encompasses both personal attributes (e.g., personality, gender) and interpersonal dynamics (e.g., humor, romance). Linguistic or rule-based definition of style theoretically constrains what constitutes a style and what not, such as a style guide (e.g., American Psychological Association 2020 (Association)) that requires that formal text not include any contraction, e.g., “isn’t”.

With the rise of deep learning methods, the data-driven definition of style leverages the variability of attributes across datasets to define style, a necessity driven by the requirements of deep learning models. For instance, the Yelp review dataset (Mukherjee et al., 2013) categorizes reviews based on ratings into positive or negative corpora, an attribute more content-related than stylistic in the traditional sense. Given the complexities and ambiguities inherent in defining and distinguishing styles, particularly through data-driven methods, the employment of neural network classifiers becomes crucial (Hovy et al.). These classifiers can effectively learn to identify and discriminate between different styles by processing diverse datasets, thereby accommodating the broader and more flexible data-driven definitions of style.

2.2 Diffusion Model

Diffusion models, a class of generative models, have gained significant attention in the field of machine learning due to their ability to generate high-quality samples from complex data distributions. The fundamental principle behind diffusion models is to learn a reversible process that gradually adds noise to the data, transforming it into a simple distribution, and then learns to reverse this process to generate new samples. The diffusion process can be described as a Markov chain of latent variables $\mathbf{x}_t$, where $t \in 0, 1, \ldots, T$ denotes the time step. The forward process begins with the original data $\mathbf{x}_0$ and progressively adds Gaussian noise to obtain the latent variables:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (1)$$

where $\beta_t \in (0, 1)$ is a variance schedule that determines the amount of noise added at each step. The reverse process used to generate new samples is defined as a Gaussian transition probability:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (2)$$

where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are learned functions parameterized by a neural network θ. In practice, the variance $\boldsymbol{\Sigma}_\theta$ is usually taken directly as β_t as an approximation for ease of calculation.

The training objective of a diffusion model is to maximize the log-likelihood of the model's predicted distribution by finding the probability distribution of the Markov chain transitions in the reverse process through maximum likelihood estimation, which the loss function can be expressed as:

$$\mathcal{L} = \mathbb{E}_q \left[\log p_\theta(\mathbf{x}_0)\right] \quad (3)$$

During inference, new samples are generated by starting from a sample of the prior Gaussian distribution $N(0, \mathbf{I})$ and iteratively applying the reverse process to obtain $\mathbf{x}_0$. Diffusion models have been successfully applied to various domains, including image generation (Ho et al., 2020), audio synthesis (Kong et al., 2020), and text generation (Li et al., 2022b), demonstrating their versatility for generating realistic and diverse samples.

3 Babel System

3.1 Problem Statement

In this paper, we aim to develop a framework that detects and corrects stylistic inconsistencies in machine translation outputs. The preservation of style in translation is crucial for maintaining the intended impact and appropriateness of translated text. As illustrated by the examples in §1, stylistic deviations can significantly impact translation quality across different domains. For instance, when legal documents lose their formal register or literary texts their poetic qualities, the translations fail to serve their intended purpose despite being semantically accurate. To address these stylistic inconsistencies, we need to tackle two fundamental challenges: (1) How to detect and characterize the stylistic attributes of text in different languages? and (2) How to ensure stylistic consistency between source and translated text while preserving semantic meaning? The first challenge requires developing a robust definition of style that captures domain-specific conventions and linguistic patterns. We adopt a data-driven approach to style modeling. Our task can be formalized as follows: Given a set of style-labeled monolingual texts as training data, how can we develop a model that accurately identifies stylistic attributes in both source and target languages? The second challenge involves developing methods to modify translation outputs to match source text style. Since we treat commercial translation systems as black boxes, we approach this as a post-processing task. Specifically: How can we efficiently generate style-refined translations that maintain both stylistic fidelity to the source text and semantic accuracy?

3.2 Overview

Workflow. Our proposed framework, Babel, introduces novel *style detector* and *style applicator* modules to determine text style, detect stylistic inconsistency bugs, and obtain accurate and stylistically consistent translation text. Figure 1 presents the workflow of Babel. The workflow consists of two steps: style testing and style repairing. Firstly, Babel takes the original text and its translated counterpart as input and uses two style detectors, trained on the respective language corpora, to determine their style attributes. If Babel detects a stylistic inconsistency between the original and translated texts, it identifies a style issue in the output and initiates the repair process. Using a style applicator pre-trained on target language corpora, Babel then adjusts the style of the translated output to match the style attributes of the original text. Lastly, we utilize the existing semantic similarity model to verify if the optimized text aligns semantically with the original translated text. If a significant discrepancy is detected, the repair process must be reiterated.

Intuition. The style detector and style applicator modules are trained on language-specific corpora, enabling the system to learn and adapt to the unique stylistic characteristics of each language. This data-driven approach ensures that the system can effectively identify and correct stylistic inconsistencies in various languages and domains. The workflow of Babel is divided into two steps: style testing and style repairing. This separation allows for a more efficient and accurate process, as the system first identifies potential style issues before initiating the repair process. This approach minimizes unnecessary computations and ensures that the system only modifies the translated text when a stylistic inconsistency is detected. Besides, this design choice allows for better maintainability, reusability, and flexibility, as each module can be independently improved, replaced, or adapted

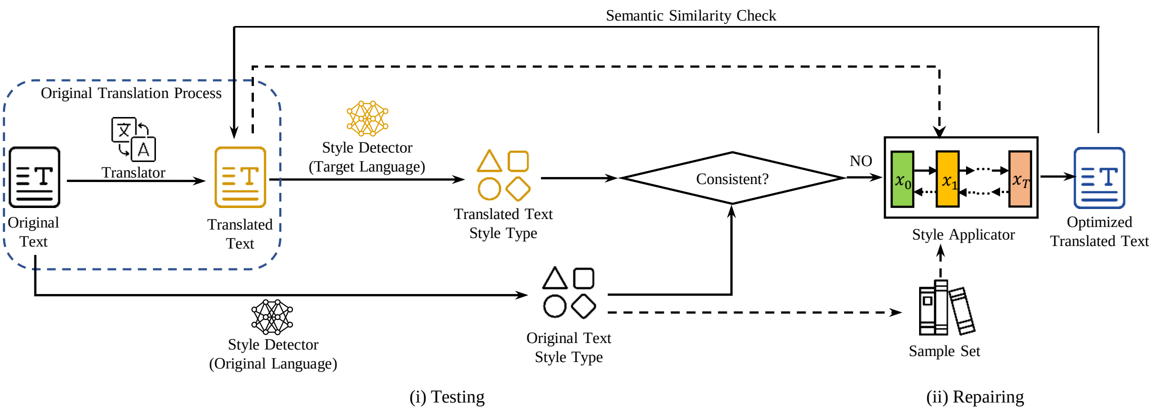


Figure 1: Overview of Babel.

to different languages or style attributes.

3.3 Style Detector

The goal of the style detector is to determine the style attributes of the original text. The primary challenge is accurately identifying and matching stylistic features across different languages, as each language has distinct stylistic norms and expressions. To address this challenge, we train a model to recognize and classify various stylistic features in texts, facilitating the alignment of stylistic attributes between source and target languages.

We start by collecting a diverse corpus of texts in both the source and target languages. These texts are annotated with various style attributes predetermined by the user or by their source, such as *law style, wikipedia style, early childhood education style*, etc. Unlike parallel texts, which are translations of the same content, our corpus only requires stylistic annotations, simplifying the data collection process. The collected texts are then tokenized and cleaned to normalize them into a common format.

Using BERT (Devlin et al.), we extract features that capture the stylistic essence of the texts. As a mainstream pre-trained transformer model, BERT can effectively capture contextual information and nuances in text, which are essential for style recognition. We train two style classification models separately for the source and target languages. For the source language, we fine-tune a BERT model on the source language corpus annotated with style attributes. The model learns to classify the text based on these annotations, identifying patterns and stylistic markers specific to the source language. Similarly, we fine-tune another BERT model for the target language, ensuring that the model can recognize and classify stylistic features in the target language independently of the source language. Once trained, the BERT models are used to classify the style of new texts. For a given original text, the source language BERT model identifies its style attributes. We then search the target language and match it with similar stylistic features. The approach of using BERT for style classification leverages its deep contextual understanding, as demonstrated in various NLP tasks. BERT's ability to capture fine-grained textual nuances makes it an excellent choice for distinguishing stylistic attributes.

Note that this similarity relationship is defined by the user. As we discussed in §2.1, style is highly subjective, making an objective definition challenging and impractical. Instead of attempting to universally define styles, we allow users to provide samples of their desired styles, lowering the barrier for customization. For example, if translating a Chinese medical text, users can provide samples of formal medical writing in both languages to maintain appropriate clinical terminology and professional register. Similarly, when translating academic content, users can provide samples of scientific writing to preserve the formal academic style and technical precision. Users can thus customize the style correspondence and the corpus according to their specific needs. The independent training on source and target languages addresses the challenge of differing stylistic norms, ensuring that each model is well-tuned to its specific language. By systematically addressing the challenges of style identification and classification in both source and target languages using BERT, our style detector effectively aligns stylistic attributes across languages, forming a robust foundation for subsequent style application.

3.4 Style Applicator

After detecting a stylistic inconsistent issue in the output of the translation system, it is essential to generate a revised output that maintains the original semantic content while ensuring stylistic consistency with the input, which is the goal of the style applicator. The style applicator consists of two key processes: training (style extraction) and inference (style application). The training process focuses on extracting style in the embedding space, while the inference process emphasizes the application of identified style attributes to the target language text. The main challenge is to effectively separate the style from the content while preserving the semantic integrity of the original text, and then accurately apply the desired style to the translated text.

3.4.1 Training process

The objective of the training process is to simulate style loss during translation within the same language, and prepare the model to extract and capture the stylistic essence of sentences while preserving their semantic content. To imitate the style loss observed in translation, we use a paraphrase model to generate paraphrases of the input text. These paraphrases retain the original meaning but have reduced stylistic elements, simulating the effect of translation where the core content remains intact, but the style may be neutralized. This step is crucial for preparing the model to neutralize and extract the stylistic essence of sentences while preserving their semantic content. The diffusion is performed in the embedding space, where the text is represented in a numerical format that captures its meaning. Operating in the embedding space helps maintain the semantic integrity of the sentences. We adopt a noise schedule that decreases to zero at a significantly slower rate compared to the cosine and square root schedules, thus preserving information more effectively. For NLP tasks, this feature is crucial as it helps to maintain the semantic information of the original text. By making small adjustments at each step, the model turns data from the noisy state to the desired state. During this process, the model learns to preserve semantic content and reconstruct the original embeddings as closely as possible. The detailed mathematical formulations are provided in §C.1.

3.4.2 Inference process

After completing the training, the diffusion model can then be used to attach attributes to the text, a process we refer to as inference process. The inference process starts with sampling initial noisy data and iteratively removes the noise to construct the improved sentences. The advantage of our style applicator is that the generated text can be gradient-guided based on user-supplied style samples, directing the output to a specific target style. Given a set of user-supplied style samples and a style embedding model, we can guide the generation process through style similarity computation and gradient updates. After estimating optimized text at each step, we proceed backward in time to iteratively acquire states with proceeding time steps. The style applicator embeds these tokens using the word embedding model and subsequently adds noise to generate the latent representation for the preceding diffusion time step. After iterating this process until the final step, we eventually get the desired output. The mathematical details of this process can be found in §C.2.

3.4.3 Summary

The style applicator is trained using neutralized paraphrases and reconstructs styled text during inference, guided by user input to achieve specific style targets. During the training process, the goal is to simulate the loss of style during translation within the same language, preparing the model to capture the essence of sentences' style while maintaining their semantic content. The text is represented numerically in the embedding space, where a diffusion process gradually adds noise. The model is trained to reconstruct the original text from noise, preserving its meaning. In inference, the trained model applies styles to new text by sampling noisy data and iteratively refining it. User-supplied style samples guide this process to achieve the desired stylistic attributes. In summary, the style applicator can effectively separate style from input texts and apply new styles to translated texts while maintaining semantic integrity.

4 Experiment

Our evaluation experiments examine both the effectiveness and efficiency of Babel in detecting and repairing stylistic inconsistencies in machine translation outputs. We evaluate Babel in two scenarios: (1) finding and fixing stylistic inconsistency issues, assessing its precision and repair success rate through both automatic metrics and human evaluation, and (2) measuring its computational efficiency and analyzing the impact of key parameters.

4.1 Setup

Datasets Due to the lack of comprehensive public datasets with parallel text in multiple languages and styles, we extracted 1000 data points from commonly used Chinese and English datasets in five domains, creating a dataset that lacks parallel texts but contains domain (style) information, as shown in Appendix A.

Translation Systems We consider four mainstream state-of-the-art machine translation systems: Google Translate (Goo), Baidu Translate (Bai), Youdao Translate (You), and Transformers (Opus-MT (Tiedemann and Thottingal, 2020)).

Evaluation metrics We evaluate Babel from three perspectives: the number of repaired issues (bias ratio), the overall state of repair (style score), and the ability to maintain semantics (semantic textual similarity).

Human Evaluation While automatic evaluation offers a preliminary assessment of the quality of repaired translations, it is insufficient for accurately gauging the quality of revised texts. To further validate the effectiveness of our approach, we conduct a human evaluation on the test set. We engage three annotators who are native Chinese speakers with proficiency in English, as well as two annotators who are native English speakers with proficiency in Chinese (see §B.3). All annotators possess advanced degrees, with a minimum of an undergraduate qualification, and include professionals in the fields of linguistics, translation studies, and literature.

4.2 Effectiveness in Finding Stylistically Inconsistent Issues

Experiment Design: To evaluate whether the translated texts generated by translation systems maintain the original style, we conducted the following steps. First, for each test sentence, we generated the corresponding translated text using translation systems. Then we assessed the style of these translated texts using a style detector trained to identify specific stylistic attributes. Additionally, as mentioned in §4.1, we randomly sample 250 of input sentences to manually evaluate whether our style detector works well. That is, we sample 50 samples from each dataset, and their corresponding 200 translated texts each after being translated by the four translation systems. Then samples are distributed to annotators. The annotators are asked to rate each output for stylistic consistency on a likert scale from 1 to 5. The generated sentence is marked as "stylistically consistent" when it is scored 4 or 5, otherwise it is marked as "stylistically inconsistent".

Results: The Babel's evaluation results are presented in Table 1. The first and second column list the four translation systems and corresponding five domains. The third and fourth column show the ratio of stylistic bias of these translation systems and the average style scores. The remaining columns list the revised text (see §4.3). The experiment results demonstrate that the stylistic bias issue is widespread in translation systems and Babel can effectively find these biased sentences.

Table 1: Effectiveness in finding stylistically inconsistent issues and repairing them. *Score* is short for *Style Score*, and *STS* is short for *Semantic Textual Similarity*.

Translation System	Domain	Bias ratio	Score	Revised Bias ratio	Revised Score	STS
Google	Law	17.54%	0.72	7.87%(-55.13%)	0.77(+6.94%)	0.91
	Literature	12.34%	0.75	7.67%(-37.84%)	0.78(+4.00%)	0.88
	Wikipedia	5.98%	0.73	2.34%(-60.87%)	0.79(+8.22%)	0.93
	Medicine	15.67%	0.74	5.98%(-61.84%)	0.80(+8.11%)	0.91
	Education	14.21%	0.76	11.56%(-18.65%)	0.81(+6.58%)	0.95
Baidu	Law	18.34%	0.71	6.78%(-63.03%)	0.76(+7.04%)	0.90
	Literature	8.54%	0.77	4.89%(-42.74%)	0.82(+6.49%)	0.87
	Wikipedia	7.33%	0.72	5.67%(-22.65%)	0.79(+9.72%)	0.92
	Medicine	7.54%	0.70	3.45%(-54.24%)	0.75(+7.14%)	0.93
	Education	13.89%	0.73	11.33%(-18.43%)	0.78(+6.85%)	0.94
Youdao	Law	16.47%	0.72	10.21%(-38.01%)	0.77(+6.94%)	0.91
	Literature	8.90%	0.78	6.54%(-26.52%)	0.83(+6.41%)	0.90
	Wikipedia	10.67%	0.74	5.22%(-51.08%)	0.79(+6.76%)	0.93
	Medicine	9.87%	0.75	7.65%(-22.49%)	0.81(+8.00%)	0.94
	Education	11.45%	0.76	9.10%(-20.52%)	0.82(+7.89%)	0.95
Opus-MT	Law	15.78%	0.72	10.56%(-33.08%)	0.77(+6.94%)	0.92
	Literature	7.45%	0.76	5.23%(-29.80%)	0.83(+9.21%)	0.89
	Wikipedia	6.34%	0.74	4.21%(-33.60%)	0.79(+6.76%)	0.92
	Medicine	19.95%	0.71	17.82%(-10.68%)	0.76(+7.04%)	0.93
	Education	13.66%	0.73	11.47%(-16.03%)	0.78(+6.85%)	0.94

Table 2: Correlation between Babel evaluation and manual inspection on stylistically inconsistent issues finding.

Translation System	Domain	TP	TN	FP	FN	Precision	FPR
Google	Law	22	23	5	0	81.48%	17.86%
	Literature	21	24	4	1	84.00%	14.29%
	Wikipedia	20	25	3	2	86.96%	10.71%
	Medicine	19	26	2	3	90.48%	7.14%
	Education	20	24	2	4	90.91%	7.69%
Baidu	Law	23	22	4	1	85.19%	15.38%
	Literature	22	23	3	2	88.00%	11.54%
	Wikipedia	23	25	0	2	100.00%	0.00%
	Medicine	20	23	1	6	95.24%	4.17%
	Education	19	23	3	5	86.36%	11.54%
Youdao	Law	21	22	6	1	77.78%	21.43%
	Literature	19	24	6	1	76.00%	20.00%
	Wikipedia	20	24	3	3	86.96%	11.11%
	Medicine	20	24	1	5	95.24%	4.00%
	Education	22	25	0	3	100.00%	0.00%
Opus-MT	Law	24	21	3	2	88.89%	12.50%
	Literature	22	20	3	5	88.00%	13.04%
	Wikipedia	21	22	2	5	91.30%	8.33%
	Medicine	15	28	6	1	71.43%	17.65%
	Education	22	25	0	3	100.00%	0.00%

On average, Google Translate had the most stylistic bias issues, accounting for 13.15% of the total output, followed by the open-source Opus-MT model (12.64%), Youdao Translate (11.47%), and Baidu Translate (11.13%).

Overall, Google Translate has the most stylistic bias issues, due to its poor performance on the literature and medicine datasets, which we speculate is due to the lack of Chinese Internet information in its training corpus. In contrast, Baidu and Youdao (NetEase), as Chinese ISPs with a large amount of Chinese online communities corpus to train their translation models, have fewer stylistic bias problems on the literature and medicine data.

Manual Inspection: The examination results based on manual inspection are presented in Table 2. The first and second column list the four translation systems and their corresponding evaluation datasets. The third to sixth columns represent true positive (TP), true negative (TN), false positive (FP) and false negative (FN) of the Babel evaluation results, respectively. The remaining three columns list the precision, recall, and false positive rate (FPR), which are important indicators for assessing the quality of the test. A false positive means the Babel judges a translation as stylistically inconsistent but manual inspection is consistent. A false negative means the Babel judges a translation as stylistically consistent but manual inspection is inconsistent. Overall, Babel exhibits a false positive rate of 10.41%, with a precision of 88.21%. These metrics indicate that Babel is effective in identifying stylistically inconsistent issues.

4.3 Effectiveness in Repairing Stylistically Inconsistent Issues

Experiment Design: After Babel identifies translated sentences with stylistic inconsistencies, we apply our method to fix them and evaluate how many translations can be successfully repaired. For each sentence, we generate four candidate translations, based on previous work in translation systems (Horvitz et al.; Han et al.). We use a style detector to assess these candidates, considering them repaired if their style matches the original sentence and the semantic loss is within an acceptable threshold.

As in §4.2, we conduct a manual evaluation. We randomly select 50 samples per dataset, each containing a source sentence, translations from different systems, and their revised versions by Babel. These samples are double-blind evaluated by five annotators who rate each on a Likert scale from 1 to 5 for style accuracy (Acc), semantic preservation (Sem), and fluency (Flu). We calculate inter-annotator agreement using Fleiss' s kappa. A sentence is marked as "successful" if it scores 4 or 5 on all three criteria. This evaluation is stricter than in §4.2, as it also considers semantic preservation and fluency in addition to stylistic consistency.

Table 3: Manual inspection results. We show average human ratings for style accuracy (Acc), semantic preservation (Sem) and fluency of sentences (Flu) on a 1 to 5 Likert scale. "Suc" denotes the overall success rate. We consider a generated output "successful" if it is rated 4 or 5 on all three criteria (Acc, Sem, Flu).

Translation System	Domain	Original Texts				Revised Texts			
		Acc	Sem	Flu	Suc	Acc	Sem	Flu	Suc
Google	Law	3.8	4.0	4.6	35%	4.2	3.8	4.4	50%(+43%)
	Literature	3.2	3.8	4.4	27%	4.0	3.8	4.4	38%(+41%)
	Wikipedia	3.4	4.4	4.4	18%	4.0	4.4	4.2	71%(+294%)
	Medicine	2.4	3.8	4.2	7%	3.2	4.0	4.6	22%(+214%)
	Education	3.0	3.6	4.2	21%	3.6	4.0	4.4	43%(+105%)
Baidu	Law	3.8	4.0	4.4	38%	4.4	3.8	4.4	49%(+29%)
	Literature	3.2	3.4	4.0	15%	3.8	3.8	4.0	31%(+107%)
	Wikipedia	2.8	4.2	3.8	8%	3.6	4.4	4.2	40%(+400%)
	Medicine	2.8	3.8	4.0	7%	3.2	4.2	3.8	16%(+129%)
	Education	3.2	3.8	4.0	14%	3.6	4.0	3.8	29%(+107%)
Youdao	Law	3.6	4.0	3.8	22%	4.2	3.8	4.0	42%(+91%)
	Literature	3.2	3.6	3.6	10%	3.8	3.8	4.2	45%(+350%)
	Wikipedia	3.2	4.0	3.8	8%	3.8	4.4	4.0	52%(+550%)
	Medicine	3.0	3.6	4.0	18%	3.2	4.0	3.8	23%(+28%)
	Education	2.8	4.0	4.4	16%	3.6	3.6	4.0	30%(+88%)
Opus-MT	Law	3.4	3.6	3.8	20%	4.2	3.8	3.6	33%(+65%)
	Literature	3.0	3.4	3.6	14%	4.0	3.0	3.4	21%(+50%)
	Wikipedia	3.4	3.8	3.2	9%	4.0	3.6	3.6	23%(+156%)
	Medicine	2.2	3.4	3.6	5%	2.8	3.2	3.8	8%(+60%)
	Education	1.8	3.2	3.8	6%	3.4	3.0	3.8	12%(+100%)

Results: The comparison results are presented in Table 1. The first and second column list the four translation systems and corresponding five domains. The third and fourth column show the ratio of stylistic bias of the original translated texts and the average style scores. The remaining columns list the the ratio of stylistic bias of the translated texts revised by Babel and the average style scores. It can be observed that across all translation systems, stylistic bias issues significantly decrease after improvements via Babel, with a corresponding increase in Style Scores. Specifically, on average, Google Translate shows the greatest reduction in issues by 46.87%, followed by Baidu Translate (40.22%), Youdao Translate (31.72%), and Opus-MT (24.64%), with an average decrease of 35.86%. For the Style Score, all four systems show improvements. On average, Baidu Translate has the highest increase of 7.45%, followed by Opus-MT (7.36%), Youdao Translate (7.20%), and Google Translate (6.77%), averaging a 7.20% increase. Moreover, Babel maintains high semantic consistency between the modified texts and the original translations. Across the four systems, the lowest Semantic Textual Similarity (STS) score reaches 0.87, with an average of 0.92.

Manual Inspection: The manual inspection results are shown in Table 3. The first two columns list the translation systems and their datasets, while the next four columns show the mean scores for the original translations, and the last four show scores for texts revised by Babel. Significant improvements were observed across all systems. Specifically, Youdao Translate saw the largest increase in style translation success, rising from 15% to 38% (221% improvement); Baidu Translate increased from 16% to 33% (154% improvement); Google Translate rose from 22% to 45% (139% increase); and Opus-MT improved from 11% to 19% (86% increase). On average, Babel repair increased the style translation success rate from 16% to 34%, a 150% improvement.

We show the efficiency of Babel in Appendix D and the impact of configurable parameters in Appendix E. Besides, examples of the repairing by Babel are shown in Appendix F.

5 Related work and discussion

Text Style Transfer Text style transfer has seen significant development, beginning with Hu et al. (2017)'s VAE framework using attribute classifiers for sentiment and tense transformation. A major advancement came from Shen et al. (2017), who introduced non-parallel text corpora with cross-aligned autoencoders, though their back-translation approach risked content distortion. Zhang et al. (2018) addressed data scarcity through pseudo-parallel data generation using SMT, while Fu et al. (2018) explored adversarial learning with both multiple and single decoder approaches for style disentanglement. To improve generation quality, Dai et al. (2019) proposed a Transformer-based architecture that eliminates explicit style disentanglement steps. In contrast to these methods that require explicit style labels and operate within fixed style categories, Babel enables text stylization using only user-supplied samples and can be adaptively trained on user-provided datasets.

Recent advances in text style transfer have leveraged Large Language models through various approaches, including model fine-tuning (Mukherjee et al.; Dementieva et al.), in-context learning (Chen; Zhang et al.; Pan et al.; Mai et al.), and prompt engineering (Luo et al.; Liu et al.). While these methods demonstrate impressive performance, they face practical limitations: fine-tuning demands substantial computational resources, while prompt-based methods often rely on carefully crafted, sensitive prompts that can lead to inconsistent results. Unlike LLMs-based methods, Babel maintains stable performance without requiring extensive computational resources, and avoids the brittleness often associated with complex prompting strategies.

Machine Translation Testing The research community has proposed various automated testing techniques to evaluate machine translation systems, primarily focusing on translation robustness. Early work by Pesu et al. (2018) introduced metamorphic testing using multiple intermediate languages, while Heigold et al. (2017) evaluated robustness against character-level perturbations. Several approaches leverage word replacement strategies: He et al. (2020) proposed structure-invariant testing (SIT) using BERT-based word substitutions, Sun et al. (2020, 2022) developed TransRepair and CAT for context-aware word replacements, and Gupta et al. (2020) introduced PatInv to verify translation consistency under semantic perturbations. Other methods explore structural aspects of translations: He et al. (2021) presented referential transparency testing (RTI) using noun phrase extraction, Ji et al. (2021) employed constituency invariance relations, and Zhang et al. (2024) introduced syntactic tree pruning. Beyond robustness testing, Chen et al. (2022) developed NMTSloth to detect efficiency bugs, and Sun et al. (2024) proposed FairMT to evaluate demographic fairness in translations. Unlike these approaches, Babel is the first work to specifically address stylistic biases in machine translation systems.

6 Conclusion

In this paper, we presented Babel, the first framework that automatically tests and repairs stylistic inconsistent issues in translation. As a black-box post-processing method, Babel takes the input text and the corresponding translated text to identify any stylistic discrepancies between the two. If inconsistencies are found, Babel performs stylistic repairs using a diffusion model, enhanced by user-supplied customized samples. Our evaluation results demonstrate that Babel effectively and efficiently mitigate stylistic bias of mainstream commercial translation systems, while maintaining semantic integrity.

Limitations

External Validity The threats to external validity lie in the implementation of the dataset we used and the selected machine translation systems. The limited dataset may not adequately characterize the diversity and linguistic stylistic features of texts to be translated in real-world scenarios. To address this concern, we sampled from corpora that are popular in both English and Chinese communities, and the dataset size is five times larger than existing translation testing work (He et al., 2020, 2021). For the selected machine translation systems, we chose state-of-the-art systems from both industry (Google Translate, Baidu Translate, and Youdao Translate) and academia (Transformer-based Opus-MT). Additionally, we have released our implementation (Ano), which can be easily extended to incorporate more datasets and machine translation systems.

Internal Validity The threats to internal validity primarily stem from the evaluation metrics used in the experiments. We measured both the style scores and semantic similarity of texts to assess improvements in retaining linguistic style and semantics. Specifically, we utilized language models to calculate semantic textual similarity. Furthermore, to verify the accuracy of these assessments, we employed manual evaluation to explore the correlation between automated assessment results and human understanding.

Ethical Considerations

While Babel aims to improve translation quality through style preservation, we acknowledge several important ethical considerations. Machine translation systems, including our framework, can perpetuate and potentially amplify societal biases present in training data (Sheng et al., 2021; Weidinger et al., 2022). The preservation of style, while beneficial for maintaining appropriate register and domain conventions, could also maintain problematic stylistic elements such as gender bias in formal writing or cultural stereotypes in literary translations.

The usage of domain-specific corpora raises additional ethical concerns. Legal and medical texts often contain sensitive information, requiring careful consideration of privacy and data protection (Carlini et al., 2021). While we have carefully selected public domain texts for our experiments, deployments of similar systems must ensure appropriate data handling protocols. Furthermore, the ability to modify translation style could be misused to generate misleading content - for instance, making informal or unreliable sources appear more authoritative by adopting formal academic or legal style (Bagdasaryan and Shmatikov, 2022).

The modular nature of our framework, which allows integration with various style classifiers, presents both opportunities and risks. While this flexibility enables adaptation to different domains and use cases, it could potentially be exploited to generate harmful content if inappropriate style models are used. We recommend implementing safeguards such as:

- Careful curation of training corpora to minimize harmful biases.
- Implementation of detection mechanisms for potential misuse.
- Clear documentation of intended use cases and limitations.

In the process of refining and improving this paper, we utilized ChatGPT and Claude for suggesting improvements in language clarity. These tools aided in enhancing the writing process but were used under human oversight to ensure that the content adheres to the ethical and scholarly standards expected in academic research.

References

ACOSharma/literature · Datasets at Hugging Face. https://huggingface.co/datasets/ACOSharma/literature.

AlekseyKorshuk/fairy-tale-books · Datasets at Hugging Face. https://huggingface.co/datasets/AlekseyKorshuk/fairy-tale-books.

Anonymized Repository - Anonymous GitHub. https://anonymous.4open.science/r/Babel-3EB2/README.md.

Baidu translator.

English levels in china: Quality of spoken english, signage, etc.

Google translator.

Twang2218/chinese-law-and-regulations · Datasets at Hugging Face. https://huggingface.co/datasets/twang2218/chinese-law-and-regulations.

Youdao translator.

2024. Sentence-transformers/all-MiniLM-L6-v2 · Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

American Psychological Association. Publication manual of the american psychological association,(2020). 428.

Eugene Bagdasaryan and Vitaly Shmatikov. 2022. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1532–1532.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulnar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, pages 2633–2650.

Dhivya Chandrasekaran and Vijay Mago. 2022. Evolution of Semantic Similarity – A Survey. *ACM Computing Surveys*, 54(2):1–37.

Jianlin Chen. LMStyle Benchmark: Evaluating Text Style Transfer for Chatbots. *Preprint*, arXiv:2403.08943.

Simin Chen, Cong Liu, Mirazul Haque, Zihe Song, and Wei Yang. 2022. Nmtsloth: understanding and testing efficiency degradation of neural machine translation systems. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1148–1160.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. Exploring Methods for Cross-lingual Text Style Transfer: The Case of Text Detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

Joshua A Fishman. 2020. Who speaks what language to whom and when? In *The bilingualism reader*, pages 55–70. Routledge.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Shashij Gupta, Pinjia He, Clara Meister, and Zhendong Su. 2020. Machine translation testing via pathological invariance. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 863–875.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive Simplex-based Diffusion Language Model for Text Generation and Modular Control. *Preprint*, arxiv:2210.17432.

Pinjia He, Clara Meister, and Zhendong Su. 2020. Structure-invariant testing for machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 961–973.

Pinjia He, Clara Meister, and Zhendong Su. 2021. Testing machine translation via referential transparency. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 410–422. IEEE.

Georg Heigold, Günter Neumann, and Josef van Genabith. 2017. How robust are character-based word embeddings in tagging and mt against wrod scramlbing or randdm nouse? *arXiv preprint arXiv:1704.04441*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. ParaGuide: Guided Diffusion Paraphrasers for Plug-and-Play Textual Style Transfer. *Preprint*, arxiv:2308.15459.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. “You Sound Just Like Your Father” Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Pin Ji, Yang Feng, Jia Liu, Zhihong Zhao, and Baowen Xu. 2021. Automated testing for machine translation via constituency invariance. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 468–479. IEEE.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep Learning for Text Style Transfer: A Survey. 48(1):155–205.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Dongyeop Kang and Eduard Hovy. Style is NOT a single variable: Case Studies for Cross-Stylistic Language Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387. Association for Computational Linguistics.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Jonathan Li, Rohan Bhambhoria, and Xiaodan Zhu. 2022a. Parameter-efficient legal domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 119–129, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022b. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, volume 38 of *AAAI'24/IAAI'24/EAAI'24*, pages 18689–18697. AAAI Press.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. Prompt-Based Editing for Text Style Transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750. Association for Computational Linguistics.

Huiyu Mai, Wenhao Jiang, and Zhi-Hong Deng. Prefix-Tuning Based Unsupervised Text Style Transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14847–14856. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013. What yelp fake review filter might be doing? In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 409–418.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. Text Detoxification as Style Transfer in English and Hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 133–144. NLP Association of India (NLPAI).

Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. Unsupervised Text Style Transfer via LLMs and Attention Masking with Multi-way Interactions. *Preprint*, arXiv:2402.13647.

Daniel Pesu, Zhi Quan Zhou, Jingfeng Zhen, and Dave Towey. 2018. A monte carlo method for metamorphic testing of machine translation services. In *Proceedings of the 3rd International Workshop on Metamorphic Testing*, pages 38–45.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4275–4293.

Zeyu Sun, Zhenpeng Chen, Jie Zhang, and Dan Hao. 2024. Fairness testing of machine translation systems. *ACM Transactions on Software Engineering and Methodology*.

Zeyu Sun, Jie M Zhang, Mark Harman, Mike Papadakis, and Lu Zhang. 2020. Automatic testing and improvement of machine translation. In *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, pages 974–985.

Zeyu Sun, Jie M Zhang, Yingfei Xiong, Mark Harman, Mike Papadakis, and Lu Zhang. 2022. Improving machine translation systems via isotopic replacement. In *Proceedings of the 44th international conference on software engineering*, pages 1181–1192.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT —Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. volume abs/1711.07010.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. Distilling Text Style Transfer With Self-Explanation From LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Quanjun Zhang, Juan Zhai, Chunrong Fang, Jiawei Liu, Weisong Sun, Haichuan Hu, and Qingyu Wang. 2024. Machine translation testing via syntactic tree pruning. *ACM Transactions on Software Engineering and Methodology*.

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

A Babel-Corpus

The selected data has distinct stylistic features and maintains the stylistic correspondence between Chinese and English, making it suitable for evaluating our method. Our preprocessing steps include metadata removal, formatting standardization, tokenization, sentence segmentation, and ensuring that each data point contained at least one complete sentence.

Ethical and privacy considerations In the compilation of our dataset, we have been vigilant in addressing ethical and privacy concerns to ensure that the data utilized does not infringe upon individual rights or breach any legal regulations. Our dataset is derived from publicly available resources, and we have taken the following measures to uphold ethical standards and protect privacy:

- All personal identifiers have been removed from the data to prevent the identification of individuals. This process includes the removal of names, addresses, and any other unique identifiers that could be linked to specific individuals.
- Where applicable, we have obtained necessary permissions and consents from the original data providers or authors of the texts to use the data for research purposes.
- The dataset has been reviewed by an independent ethics committee to ensure that it meets the ethical standards required for academic research.

B Detailed setup of Babel

B.1 Software and Hardware

We conduct our experiments on a server with 64 cores Intel Xeon 2.90GHz CPU, 256 GB RAM, and 4 NVIDIA 3090 GPUs running the Ubuntu 20.04 operating system.

B.2 Translation Systems

Google Translate. Google Translate is a multilingual neural machine translation system developed by Google. It supports over 100 languages, has a vast user base with over 500 million users and translates more than 100 billion words daily.

Baidu Translate. Baidu Translate is optimized for translations between Chinese and other languages, leveraging Baidu's AI and big data technologies. It serves millions of users in China, providing text, voice, and image translation services.

Youdao Translate. Youdao Translate, developed by NetEase, integrates rich dictionary resources and NMT technology for accurate translations, particularly beneficial for educational purposes. It is widely used by students and educators in China, with millions of active users.

Opus-MT. Opus-MT is an open-source neural machine translation model based on the Transformers architecture, supported by the Marian NMT toolkit. It is popular among researchers and developers for its flexibility and customization options, and there were 1.55M downloads on huggingface last month.

Table 4: Datasets being extracted to Babel-Corpus.

Domain	English	Chinese
Law	Law Stack Exchange (Li et al., 2022a)	ChineseLaw and Regulations (Twa)
Literature	Classic Literature in ASCII (ACO)	Chinese Literature (Xu et al., 2017)
Wikipedia	wikitext (Merity et al., 2016)	wiki2019zh (Xu, 2019)
Medicine	PubmedQA (Jin et al., 2019)	CBLUE (Zhang et al., 2022)
Early Childhood Education	Fairy Tale Books (Ale)	CFT (Cui et al., 2016)

B.3 Human Evaluation

We engaged ten annotators for the evaluation process: three native Chinese speakers (live in China) proficient in English, and two native English speakers proficient in Chinese (live in Singapore). All annotators hold advanced degrees, with at least an undergraduate qualification, and have professional expertise in linguistics, translation studies, or literature. These qualifications ensured that the evaluation was conducted by individuals with the necessary expertise to assess translation quality accurately. Annotators were recruited based on their professional backgrounds and were not compensated for their participation, as the study was conducted in-house with experts who volunteered due to their academic and professional interests. Consent was obtained from all annotators before their involvement, with instructions provided that explained how their evaluation data (ratings and feedback) would be used for research purposes. The instructions also outlined the evaluation criteria and expectations for the task, ensuring full transparency. No risks to participants were identified, and participation was voluntary. The data collection protocol was exempt from formal ethics review, as it involved professional annotators in a controlled research setting, and all procedures adhered to ethical standards for transparency and informed consent.

C Technical Details

C.1 Training Process Formulation

To imitate the style loss observed in translation, we use a paraphrase model $P(\cdot)$, such as PEGASUS (Zhang et al., 2020), to generate paraphrases $\mathbf{p}$ of the input text $\mathbf{r}$. Formally, we have:

$$\mathbf{p} = P(\mathbf{r}) \tag{4}$$

These paraphrases retain the original meaning but have reduced stylistic elements, simulating the effect of translation where the core content remains intact, but the style may be neutralized. This step is crucial for preparing the model to neutralize and extract the stylistic essence of sentences while preserving their semantic content.

The diffusion is performed in the embedding space, where the text is represented in a numerical format that captures its meaning. Operating in the embedding space helps maintain the semantic integrity of the sentences. Formally, we have:

$$\mathbf{x}_t = \sqrt{\beta_t} E(\mathbf{r}) + \sqrt{(1-\beta_t)}\boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I}) \tag{5}$$

where $E(\cdot)$ is an embedding model. In terms of the schedule of noise, we follow the paradigm of Horvitz et al., that is:

$$\beta_t = \sqrt{\frac{T-t}{T}} \tag{6}$$

This schedule decreases to zero at a significantly slower rate compared to the cosine and square root schedules, thus preserving information more effectively. For NLP tasks, this feature is crucial as it helps to maintain the semantic information of the original text.

Algorithm 1 Style Applicator of Babel

Input: $\mathbf{r}$: initial translated texts
Input: $\mathbf{y}$: user-supplied sample texts
Input: $D_{\theta^*}(\cdot)$: trained diffusion model
Input: T: number of total diffusion step
Input: λ: number of total diffusion step
Output: $\mathbf{r}^*$: optimized translated texts

```
procedure ApplicateStyle
    x_T ← SampleFrom(N(0, I))    ▷ Sample a random gaussian noise
    for t ← T to 1 do
        l̂_t ← D_θ*(x_t, t, r)
        r̂_t ← SampleFrom(Top-p(Softmax(l̂_t)))
        J ← GetSimilarity(r̂_t, y)
        l̂*_t ← l̂_t − λGetGradient(J)
        r̂*_t ← SampleFrom(Top-p(Softmax(l̂*_t)))
        x_{t−1} ← ForwardDiffusion(r̂*_0)
    l̂_0 ← D_θ*(x_0, 0, r)
    r̂_0 ← SampleFrom(Top-p(Softmax(l̂_0)))
    J ← GetSimilarity(r̂_0, y)
    l̂*_0 ← l̂_t − λGetGradient(J)
    r̂*_0 ← SampleFrom(Top-p(Softmax(l̂*_0)))
    return r̂*_0
```

After completing these preparations above, we perform the training process on the paraphrased text, as mentioned in §2.2. Formally, we train the model $D_\theta(\cdot)$ by minimizing the cross entropy between the posterior distribution of the model at each diffusion time step and the actual embeddings:

$$\mathcal{L}(\theta) = \mathcal{E}\left[\log p_\theta(\mathbf{r}|D_\theta(\mathbf{x}_t, t, \mathbf{p}))\right] \tag{7}$$

where $\mathcal{L}(\cdot)$ is the loss function, $\mathcal{E}(\cdot)$ represents the cross entropy function, $\mathbf{r}$ is the original text, t represents the time step, and $\mathbf{p}$ represents the paraphrase. By making small adjustments at each step, the model turns data from the noisy state to the desired state. During this process, the model learns to preserve semantic content and reconstruct the original embeddings as closely as possible.

C.2 Inference Process Formulation

After completing the training, the diffusion model $D_{\theta^*}(\cdot)$ can then be used to attach attributes to the text, a process we refer to as *inference process*.

The inference process starts with sampling initial noisy data $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively removes the noise to construct the improved sentences. For each time step t ($t \in [T, 1]$), the style applicator estimates an optimized text:

$$\hat{\mathbf{r}}_t \sim \text{top-p}(\text{softmax}((D_{\theta^*}(\mathbf{x}_t, t, \mathbf{r}))) \tag{8}$$

where $\mathbf{r}$ represents initial translated texts output by translation system.

The advantage of our style applicator is that the generated text can be gradient-guided based on user-supplied style samples, directing the output to a specific target style. Given a set of user-supplied style samples $[\mathbf{y_1}, \cdots, \mathbf{y_n}]$ and a style embedding model $E_s(\cdot)$, we can obtain the style guidance function

$$J = \frac{\sum_{i=1}^{n} d(E_s(\hat{\mathbf{r}}_t), E_s(\mathbf{y_i}))}{n} \tag{9}$$

where $d(\cdot, \cdot)$ represents cosine similarity. So, we get the final style-guided textual inference equation:

$$\hat{\mathbf{r}}_t^* \sim \text{top-p}(\text{softmax}((D_{\theta^*}(\mathbf{x}_t, t, \mathbf{r})) - \lambda \nabla J)) \tag{10}$$

After estimating $\hat{\mathbf{r}}_t^*$, we proceed backward in time to iteratively acquire states with proceeding time steps. Similarly to the training process, the style applicator embeds these tokens using the word embedding model $E(\cdot)$ and subsequently adds noise to generate the latent representation for the preceding diffusion time step:

$$\mathbf{x}_{t-1} = \sqrt{\beta_{t-1}} E(\hat{\mathbf{r}}_t^*) + \sqrt{(1 - \beta_{t-1})}\boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{11}$$

After iterating this process until $t = 0$, we eventually get the desired output $\hat{\mathbf{r}}_0^*$.

C.3 Model Training

The two core models of Babel are configured as follows:

Style Detector. We start with the publicly available BERT-base-cased checkpoint[1] and BERT-base-chinese checkpoint[2], both equipped with a classification head. Our model is trained for 200 steps on the Babel-Corpus (train-test ratio is set to 8:2), with a batch size of 16 and a learning rate of 2e-5. The style classification threshold h is set to 0.5, and we explore the impact of this parameter on style detection in Appendix E.

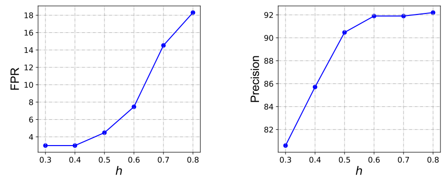
Style Applicator. We employ the publicly available SSDLM RoBERTa-large checkpoint (Horvitz et al.) and train our model for 200K steps on train set of Babel-Corpus, with a batch size of 128, total time steps of 800, and learning rate of 1e-5. During inference, we use temperature $\tau = 0.3$ and guidance strength $\lambda = 1000$. We investigate effects of these parameters in Appendix E.

User Customization. Users can customize the style they wish to address by providing samples of bilingual texts that exhibit the desired style. This process involves collecting a sufficient number of bilingual texts within the same style domain and fine-tuning them using a script we provide (Ano). For detailed information on the training cost, please refer to Appendix D. It is important to note that, although this paper focuses on Chinese-English bilingual style repair due to resource constraints, Babel is theoretically applicable to any bilingual style text repair. To adapt Babel for other language pairs, such as English-German, users need to provide English and German text samples of the target style and replace the base models in the style detector and style applicator. Specifically, bert-base-chinese should be replaced with a German BERT model, such as BERT-base-german-cased[3], and the SSDLM RoBERTa

[1] https://huggingface.co/google-bert/bert-base-cased
[2] https://huggingface.co/google-bert/bert-base-chinese
[3] https://huggingface.co/google-bert/bert-base-german-cased

Table 5: The time overhead of Babel. The values in the table are averaged over the entire dataset, in units of seconds.

Translation System	Testing Cost	Repairing Cost
Google	1.81	4.14
Baidu	1.60	3.97
Youdao	1.74	3.69
Opus-MT	1.67	3.73



(a) h-FPR **(b)** h-precision

Figure 2: Effect of h on the average performance of Babel's testing process.

model should be substituted with a German large-language model, like xlm-roberta-german[4].

C.4 Evaluation metrics

Bias Ratio. We utilized the style detector to quantify the number of stylistic bias in the outputs of each translation system and to determine the proportion of bias relative to the total sample (refer to §4.2 for the detailed methodology). To ensure the validity of the style detector, we conducted a manual evaluation for confirmation (see Appendix E).

Style Score. To evaluate the overall stylistic bias of the translation system, we calculate the average style scores of all its outputs. These style scores are derived from the confidence provided by the style detector. Due to varying sentence lengths and stylistic distinctiveness across datasets, this score lacks absolute significance and is meaningful only when comparing different translation systems on the same dataset.

Semantic Textual Similarity. STS (Semantic Textual Similarity) (Chandrasekaran and Mago, 2022) is a criterion that assesses how similar two texts are in terms of meaning. Since our focus lies in assessing the ability to repair translations without parallel texts, we calculate the STS score between the revised text and the initial translated text to gauge Babel's proficiency in preserving semantic integrity. We use one of the most commonly used models for this task, *all-MiniLM-L6-v2* (Sen, 2024), for this assessment.

D Efficiency in Testing and Repairing stylistically inconsistent Bias

Experiment Design: To assess efficiency, we meticulously measure the time Babel expends during both the testing and repair phases for stylistic inconsistencies. For each translation system involved, we calculate the average duration required by Babel to complete a single cycle of stylistic bias detection and subsequent rectification. This comprehensive timing analysis enables us to determine the operational speed of Babel, ensuring it efficiently addresses style bias without significantly detracting from user experience, thereby maintaining seamless workflow integration.

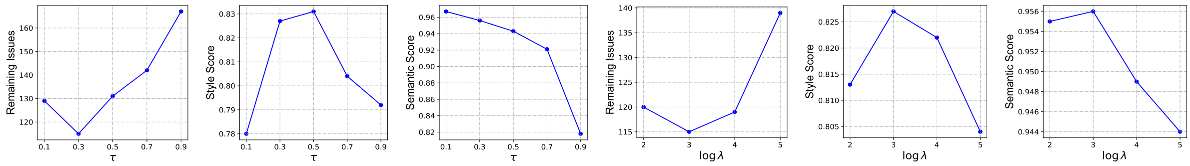
Results: The results are presented in Table 5. On average, testing a single original translation text with Babel requires only 1.7 seconds, while repairing a problematic text takes just 3.9 seconds. This demonstrates that Babel is efficient in both testing and repairing, improving translation style without significantly impacting user experience.

Training cost: To estimate the computational cost for users adding a new style, we measured the time required to fine-tune a fifth domain style fix on a Babel that already supports four domains. To mitigate the impact of individual datasets, we performed fine-tuning on each of the five domains separately and calculated the average time as the result. Our experiments, conducted using the computational resources described in Appendix B, indicate that fine-tuning a pair of datasets, each containing 1,000 samples, takes an average of 37,261 seconds (approximately 10.5 hours). It is important to note that the computational cost may vary substantially depending on the size of the datasets and the specific languages involved.

E Impact of Configurable Parameters

Experiment Design: Babel leverages three hyperparameters: detection threshold h, inference temperature τ, and guidance strength λ, to find and repair stylistic consistent issues. The detection threshold h determines the point at which a sentence's style score is categorized as a stylistic inconsistent issue, with a lower h indicating a greater tolerance for style inconsistencies. The inference temperature τ represents the maximum lexical deviation allowed from the initial translation when generating the revised sentence, with a higher τ granting greater freedom to modify the initial translation. The parameter λ denotes the strength of

[4]https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-german



(a) τ-issues **(b)** τ-style score **(c)** τ-semantic score **(d)** λ-issues **(e)** λ-style score **(f)** λ-semantic score

Figure 3: Effect of τ and λ on the average performance of Babel's repairing process.

user-supplied style guidance samples for generating revised sentences.

We conduct experiments to investigate and understand how different values of these configurable hyperparameters affect the performance of Babel in finding and repairing stylistic consistent issues. Specifically, we evaluate Babel's performance using output from Google Translate, testing a range of h values from 0.3 to 0.8, τ values from 0.1 to 0.9, and λ values from 1e2 to 1e5.

To assess the impact of h on the detection of style problems, we analyze changes in precision and false positive rate of the detector as h varies, using manual labeled samples as detailed in §4.2. For the style applicator, we evaluate the effects of varying τ and λ on repair effectiveness, measuring changes in the number of repaired issues, overall style scores, and semantic textual similarity (STS) values post-repair.

Results: Figure 2 illustrates the impact of parameter related to the style detector on its performance, whereas Figure 3 demonstrates the influence of parameters related to the style applicator.

Impact of h: The parameter h influences the sensitivity of style detector in identifying stylistic inconsistencies. As shown in Figure 2(a), FPR increases from 3% to 18.3% as h increases from 0.3 to 0.8, indicating that higher h value leads to a more radical detection of style issues. Concurrently, precision increases from 80.6% to a peak of 90.5% at $h = 0.5$, then slightly increases to 92.2% at $h = 0.8$. The results indicate that both metrics increase as h rises, initially grows more slowly and then accelerates, while the precision increases rapidly at first and then plateaus. This pattern suggests an optimal balance point at $h = 0.5$, where precision is nearly maximized while the false positive rate is reasonably low.

Impact of τ: The parameter τ plays a crucial role in the repair of stylistic inconsistent issues. Figure 3(a), (b) and (c) shows the effect of τ. As τ increases, the number of remaining issues after repair initially decreases, reaching an optimal value at $\tau = 0.3$, and then increases. Concurrently, the overall style score of the revised output follows a similar trend, achieving optimal performance at $\tau = 0.5$. The semantic score, however, consistently decreases with increasing τ, with a more rapid decline observed at higher τ values. Considering the trade-off between style and semantic score, Babel selects 0.3 as the default value of τ.

Impact of λ: The parameter λ affects the weight given to style preservation during the repair process. Figure 3(d), (e) and (f) shows the effect of λ. The figure demonstrates that as λ increases, the number of residual issues after repair initially decreases and then increases. Similarly, both the overall style score and semantic score of the revised output follow an increasing trend initially, reaching an optimal point at $\lambda = 1000$, before declining. Consequently, $\lambda = 1000$ is selected as the default value for optimal performance.

Analysis: From Figure 2 and Figure 3, we can observe that the configurable parameters h, τ, and λ have a significant impact on the performance of Babel in detecting and repairing stylistic inconsistencies. In terms of detection accuracy, increasing h initially improves precision while maintaining a reasonable false positive rate, suggesting an optimal balance at $h = 0.5$. For repair performance, the parameter τ shows that allowing moderate lexical deviations ($\tau = 0.3$) optimizes the number of corrected stylistic issues, while a higher τ value can detrimentally affect semantic integrity. The guidance strength parameter λ demonstrates that moderate user-supplied guidance ($\lambda = 1000$) enhances both stylistic and semantic scores, with performance declining at higher values. Consequently, to achieve optimal detection and repair, we set $h = 0.5$, $\tau = 0.3$ and $\lambda = 1000$ as default values in Babel.

Summarization: We have proved the advancement of Babel through the above experimental evaluations. Overall, Babel is capable of detecting over 80% of stylistic inconsistencies in translations and successfully enhances approximately

83% of these inconsistent outputs. The additional computational expense of Babel remains relatively modest, averaging no more than 6 seconds, which makes it feasible for integration into a wide range of commercial translation systems.

F Qualitative examples

To illustrate the effectiveness of Babel in preserving domain-specific styles, we present a collection of example translations in Figure 4. These examples span our five domains (legal, literary, scientific writing, medical, and educational content) and demonstrate both Chinese-to-English and English-to-Chinese translations. Each row shows an original text, its direct translation from a commercial system (Google Translate, DeepL, Bing, or Opus-MT), and Babel's style-refined version. For instance, in legal texts, Babel transforms casual expressions like "根据这里的规定" into proper legal language "受限于本协议之规定", maintaining formal register. In literary translation, it preserves poetic elements, transforming literal translations like "The mountains end at the plains" into more literary renderings like "Where mountain meets the boundless plain". The examples highlight how Babel preserves domain-appropriate terminology and conventions while maintaining semantic accuracy. Bold text indicates specific stylistic elements that were improved in the repair process.

Categories	Original Text	Translation	Repaired Translation
Law	Subject to the provisions herein	根据这里的规定	受限于本协议之规定
	若合同任何条款与法律相抵触，以法律规定为准。	If any terms of the contract conflict with the law, the law shall prevail. (By Bing)	In the event that any provision hereof conflicts with applicable laws, such laws shall prevail and govern.
Literature	Her heart was a secret garden and the walls were very high.	她的心是个秘密花园，围墙很高。(By DeepL)	她心似秘园，重重围墙高耸。
	山随平野尽，江入大荒流。	The mountains end at the plains, and the river flows into the wilderness. (By DeepL)	Where mountain meets the boundless plain, the mighty river seeks the wild domain.
Textbook	The experiment demonstrated a significant correlation.	实验显示出明显的相关性。(By Bing)	实验结果表明存在显著相关关系。
	恐龙是因为**小行星**撞击地球而灭绝的。	Dinosaurs died out because a **small star** hit the Earth. (By Opus-MT)	Dinosaurs became extinct due to the impact of an **asteroid** on Earth.
	化学方程式必须**遵守**质量守恒定律。	Chemical equations must **obey** the law of conservation of mass. (By DeepL)	Chemical equations must **adhere to** the law of conservation of mass.
Medicine	The patient **presents** with symptoms of a common cold.	病人**展现(show)**了普通感冒的症状。(By Bing)	患者**出现(appear)**了普通感冒的症状。
	建议进行进一步检查。	Suggest further examination. (By Bing)	Further diagnostic evaluation is recommended.
Early Childhood Education	The little rabbit **hopped** merrily through the forest.	小兔子在森林里快乐地**单脚跳(jump on one leg)**。(By Google)	小兔子在森林里欢快地**蹦蹦跳跳(bouncing around)**。
	小猫钓鱼，总是**心不在焉**。	The kitten is fishing but always **absent-minded**. (By Google)	The kitten goes fishing, but **can't focus**.

Figure 4: Example of stylistic inconsistent issues and repaired translation generated by Babel.