Controlling Styles in Neural Machine Translation with Activation Prompt

Yifan Wang[1,2*], **Zewei Sun**[2], **Shanbo Cheng**[2], **Weiguo Zheng**[1], **Mingxuan Wang**[2]
[1] Fudan University, [2] ByteDance
isivan.wang@gmail.com, zhengweiguo@fudan.edu.cn
{sunzewei.v,chengshanbo,wangmingxuan.89}@bytedance.com

Abstract

Controlling styles in neural machine translation (NMT) has attracted wide attention, as it is crucial for enhancing user experience. Earlier studies on this topic typically concentrate on regulating the level of formality and achieve some progress in this area. However, they still encounter two major challenges. The first is the difficulty in style evaluation. The style comprises various aspects such as lexis, syntax, and others that provide abundant information. Nevertheless, only formality has been thoroughly investigated. The second challenge involves excessive dependence on incremental adjustments, particularly when new styles are necessary. To address both challenges, this paper presents a new benchmark and approach. A **m**ultiway **s**tylized **m**achine **t**ranslation (**MSMT**) benchmark is introduced, incorporating diverse categories of styles across four linguistic domains. Then, we propose a method named **style a**ctivation **p**rompt (**StyleAP**) by retrieving prompts from stylized monolingual corpus, which does not require extra fine-tuning. Experiments show that StyleAP could effectively control the style of translation and achieve remarkable performance.

1 Introduction

Natural language texts can be written in various styles while preserving the content, such as politeness, formal, classical, and many others (Hovy, 1987; Jing et al., 2020; Fu et al., 2018). Styles are crucial for communication since every sentence should fit a specific scenario and the appropriate style makes it more user-centric. A speaker needs to switch the styles of words to adapt to different conditions. Using the inappropriate style can be impolite or ridiculous in some societies and result in serious cultural conflicts (Nida and Taber, 2021).

Direction	Source Text	Target Text	Style
English-to-Chinese	On the eleventh, an egg-sized black spot appeared on the sun.	十一日，太阳上出现像鸡蛋大的黑点。	Modern
		壬辰，日有黑子如鸡卵。	Classical
Chinese-to-English	继续，继续，否则我就宣布我自己是赢家。	Keep going, keep going, or I'll declare myself the winner.	Modern
		Switch and spurs, switch and spurs, or I'll cry a match.	Early
English-to-Korean	What's up guys, I'm Cole.	안녕하세요 여러분, 전 콜입니다.	Honorific
		안녕 얘들아, 난 콜이야.	Non-hono
English-to-Portuguese	In the Andes, this glacier is the source of drinking water for this city.	Nos Andes, todo o gelo é a principal fonte de água potável para toda cidade.	European
		Nos Andes, essa geleira é a fonte de água potável para toda cidade.	Brazilian

Figure 1: Examples of stylized translation. For each language pair, two different translation styles are shown.

As a cross-lingual generation problem, machine translation performance heavily relies on the appropriate style. Therefore, many commercial translation systems provide multiple style choices, such as Portuguese (European vs. Brazilian) and English (American vs. British) in DeepL[1], Korean (Honorific vs. Non-honorific) in Papago[2], Chinese (Modern vs. Classical) in Volctrans[3].

Recently, controlling style in machine translation has also drawn much attention in the academic community (Yamagishi et al., 2016; Michel and Neubig, 2018; Feely et al., 2019). Formally, stylized machine translation refers to translating the source sentence into different styles with certain attributes while the translation quality remains satisfactory, as the cases showed in Figure 1. Many previous studies have explored the task and gained promising results (Sennrich et al., 2016; Rabinovich et al., 2017; Wang et al., 2021). However, challenge still remains in two aspects.

The first challenge is about the benchmark. The style of natural languages consists of many aspects like word preference and grammar structure. However, the well-studied benchmark datasets mainly focus on the formality and politeness of European languages. Due to this limitation, previous work re-

*Work done while Y. Wang was an intern at ByteDance.

[1] https://www.deepl.com
[2] https://papago.naver.com
[3] https://translate.volcengine.com

Findings of the Association for Computational Linguistics: ACL 2023, pages 2606–2620
July 9-14, 2023

stricts styles to a relative narrow scene. In addition, most of the test sets of the previous work have only one reference rather than multiple stylized references, which hinders the automatic evaluation for different styles. As such, a benchmark involving more diverse styles, multiple stylized references and beyond European languages is greatly needed.

The second challenge is about the iterative training framework. Most related work heavily relies on fine-tuning with new stylized data (Sennrich et al., 2016; Wang et al., 2021). Basically, they collect stylized bilingual texts and append tags before the sentence, then conduct fine-tuning to adapt the model to the given style (Sennrich et al., 2016). However, parallel data in specific styles is pretty sparse and costly to gather. Furthermore, in this way, we have to re-tune the model every time we want to add new styles, which is inconvenient.

Correspondingly, this paper contributes in terms of both benchmark and approach:

For the benchmark, we re-visit this task and push the boundary of styles to a wider range of language phenomena. We propose a dataset **MSMT**, including four directions with diverse language styles. We collect related public corpus as training sets and provide newly labeled sentences as test sets. Each source sentence has two references in different styles, which is convenient for automatic evaluation. By broadening the category and providing standard datasets, we hope to effectively push the development of this field.

For the approach, we propose **style a**ctivation **p**rompt (**StyleAP**), a method to avoid re-tuning time after time. The main idea is to extract one sentence of the target style as a prompt to guide the main sentence translation style. The intuition is straightforward. We assume that once the model has been trained on all kinds of data with various styles, it has the potential to generate any style as far as correctly activated. We can activate the ability by language model since it tends to maintain the sequence consistency (Sun et al., 2022b). And the prompt can be easily retrieved in a specific stylized monolingual corpus. In a word, we can obtain a "plug-and-play" model for any new generation style with mere stylized monolingual data instead of iterative fine-tuning. The experiments show that our approach achieves explicit style transformation while well maintaining the text semantics.

2 Related Work

2.1 Style Transfer for Machine Translation

Existing studies on style transfer mainly focus upon formality (Feely et al., 2019; Wu et al., 2021). They can be roughly divided into two groups: supervised methods and unsupervised methods. Sennrich et al. (2016) propose side constraints to control politeness and shows that substantial improvements can be made by limiting translation to the required level of politeness. Niu et al. (2017) propose a Formality-Sensitive Machine Translation (FSMT) scenario where lexical formality models are used to control the formality level of NMT product. Since the parallel sentences are of unknown formality level, some work focus on the unsupervised way. Niu and Carpuat (2020) introduce Online Style Inference (OSI) to generate labels via a pre-trained FSMT model. Feely et al. (2019) use heuristics to identify honorific verb forms to classify the unlabeled parallel sentences into three groups of different level formality. Wang et al. (2021) propose to use source token, embedding and output bias to control different styles and achieve a remarkable performance. Wu et al. (2020b) propose a machine translation formality corpus. Diverse translation is also related to this work (Sun et al., 2020; Wu et al., 2020a).

2.2 Adaptive via In-Context Learning

Recent work shows that prompting the large language models (LMs) like GPT-3 (Brown et al., 2020) with a few examples in the context can further leverage the inductive bias of LMs to solve different NLP tasks (Wang et al., 2022). This part of work shows the adaptive ability of LMs learned from analogy. Our work is inspired by it, but we work under the iterative training situation where the supervised data is pretty sparse.

As prompts play a vital role in generic in-context learning, recent work propose different prompting strategies. Ben-David et al. (2021); Sun et al. (2022a) select the representative keywords of the field for domain adaptation. Zhu et al. (2022) capture keywords of images as prompts for multimodal translation. Hambardzumyan et al. (2021) put special tokens into the input and use continuous embeddings as prompts and Li and Liang (2021) directly optimize prompts in the continuous space. Besides, there is a research direction focusing on retrieval. These methods use two main representations for generating demonstrations. As for the sparse representations, they focus on a rule-

based score such as Okapi BM25 (Robertson and Zaragoza, 2009) for retrieval. Wang et al. (2022) use this method to improve model performance on four NLP tasks. The dense representations are generated by the pre-trained autoencoder model and have higher recall performance on most NLP tasks such as machine translation (Cai et al., 2021). For the sake of accuracy and storage, we use dense representations for retrieval in this paper.

3 Task Definition & MSMT: A Multiway Stylized Translation Benchmark

Stylized machine translation refers to the translations with certain language characteristics or styles on the basis of ensuring the quality of translation. Based upon the definition, we construct a stylized machine translation benchmark including four language directions. In each language direction, we give the illustration of various styles and provide corresponding training and test sets.

Different from traditional stylized machine translation studies, each group of our test sets is consist of one single source and multiple references in parallel. For example, for English-to-Chinese direction, for each English source sentence, we have two parallel Chinese references: classical style and modern style. In this way, we can automatically evaluate the style transformation by measuring the similarity between the stylized hypothesis and stylized references.

All the data has been publicly released and the detailed number is in Table 1. In this section, we will introduce our benchmark construction.

3.1 English-to-Chinese Translation

There are two common styles for Chinese: *Classical* and *Modern*. Classical Chinese originated from thousands of years ago and was used in ancient China. Modern Chinese is the normal Chinese that is commonly used currently.

The former is adopted on especially solemn and elegant occasions while the latter is used in daily life. They vary in many aspects like lexis and syntax so can be regarded as two different styles. In this direction, we aim at translating texts from English to Chinese in both styles. Specific data usage is as follows:

- **Basic Parallel Data:** Cleaned WMT2021 corpus plus the back translation of the subset of an open source corpus containing classical Chinese and modern Chinese [4].
- **Stylized Monolingual Data:** The open source corpus containing classical Chinese and modern Chinese and the Chinese part of WMT2021.
- **Development Set:** Newstest2019.
- **Test Set:** English-Classical-Modern triplet parallel data annotated by language experts.

3.2 Chinese-to-English Translation

There are two common styles for English: *Early Modern* and *Modern*. Early Modern English in this paper refers to English used in the Renaissance such as Shakespearean plays. Modern English is the normal English used currently.

The former one is mostly seen in Shakespearean play scripts like Hamlet while the latter one is used in the daily life. They vary in many aspects like grammatical constructions such as two second person forms, thou and you. Therefore, they can be regarded as two styles. In this direction, we aim at translating texts from Chinese to English in both styles. Specific data usage is as follows:

- **Basic Parallel Data:** Cleaned WMT2021 corpus plus the back translation of a crawled corpus: *The Complete Works of William Shakespeare* [5].
- **Stylized Monolingual Data:** An open source dataset[6] containing early modern and modern English and the English part of WMT2021.
- **Development Set:** Newstest2019.
- **Test Set:** Chinese-Early-Modern triplet parallel data annotated by language experts.

3.3 English-to-Korean Translation

There are seven verb paradigms or levels of verbs in Korean, each with its own unique set of verb endings used to denote the formality of a situation. We simplify the classification and roughly divide them into two groups: *Honorific* and *Non-honorific*

The former is used to indicate the hierarchical relationship with the addressee such as from the young to the old, from the junior to the senior. The latter is used in daily conversations between friends. They vary in some lexical rules so can be regarded

[4] https://github.com/NiuTrans/Classical-Modern

[5] http://shakespeare.mit.edu/

[6] https://github.com/harsh19/Shakespearizing-Modern-English

	en→zh		zh→en		en→ko		en→pt	
Styles	Modern	Classical	Modern	Early	Honorific	Non-hono	European	Brazilian
Monolingual	22M	967K	22M	83.2K	20.5K	20.5K	168K	234K
Parallel	9.12M		9.11M		271K		412K	
Develepment	1,997		2,000		879		890	
Test	1,200		1,182		1,191		857	

Table 1: MSMT Statistical Description. The table shows the number of training, development, and test sets.

as two styles. In this direction, we aim at translating texts from English to Korean in both styles. Specific data usage is as follows:

- **Basic Parallel Data:** IWSLT2017 [7] plus the back translation of an open source dataset [8] containing honorific and non-honorific.
- **Stylized Monolingual Data:** The open source dataset and the crawled corpus from a public translation tool[9].
- **Development Set:** IWSLT17.
- **Test Set:** English-Honorific-Non-honorific triplet parallel data annotated by language experts language experts.

3.4 English-to-Portuguese Translation

There are two common styles for Portuguese: *European* and *Brazilian*. European Portuguese is mostly used in Portugal. Brazilian Portuguese is mostly used in Brazil.

They vary in some detailed aspects like pronunciation, grammar and spelling, so can be regarded as two different styles. In this direction, we aim at translating texts from English to Portuguese in both styles. Specific data usage is as follows:

- **Basic Parallel Data:** IWSLT2017.
- **Stylized Monolingual Data:** European & Brazilian part of the parallel data.
- **Development Set:** IWSLT17.
- **Test Set:** English-European-Brazilian triplet parallel data annotated by language experts.

3.5 Evaluation

Previous style evaluation relies on human resources, which is costly and slow. Since our test sets are all multiway, we can evaluate our stylized hypothesis with the corresponding reference to take both quality and style into consideration at a small cost. Moreover, human evaluation is inevitably subjective while our test sets can guarantee the comparison stability.

[7] https://wit3.fbk.eu/

[8] https://github.com/ezez-refer/Korean-Honorific-Translation

[9] https://papago.naver.com/

4 Style Activation Prompt

Prior work usually uses the fixed tag to control the generation with expected attributes (Sennrich et al., 2016). However, tag-based methods rely on a large amount of labeled parallel data, requiring relabeling and retraining of models when new styles need to be generated.

We go back to a standard NMT model. During the generation process, the NMT model tries to maximize the probability of the generation sentence. When predicting the i-th token, the model searches from the vocabulary to maximize:

$$P(y_i|x, y_{j<i})$$

where $y_{j<i}$ means the past words, indicating that the previous inference results can affect the subsequent generation. Therefore, our intuition is to control the generation style by taking advantage of stylized language model. We suggest that once the basic model has been trained on the data in various kinds of styles, we can activate the ability with the contextual influence.

Specifically, we retrieve an instance as a prompt from the stylized corpus and use it to instruct the NMT model to generate the sentence with the same attributes. To adapt to the prompt training, we extract every sentence in the basic parallel data and retrieve one similar sentence as the prompt. The whole framework is in Figure 2. We introduce the details of our proposed method as follows.

4.1 Prompt Retrieval

The prompt retrieval procedure aims at finding the proper sentence prompt. First, we construct a candidate datastore D that contains many (r, y) key-value pairs, where r is the representation of y. In this paper, we use a multilingual pre-trained language model XLM-R (Conneau et al., 2019) to obtain the sentence representation. By calculating

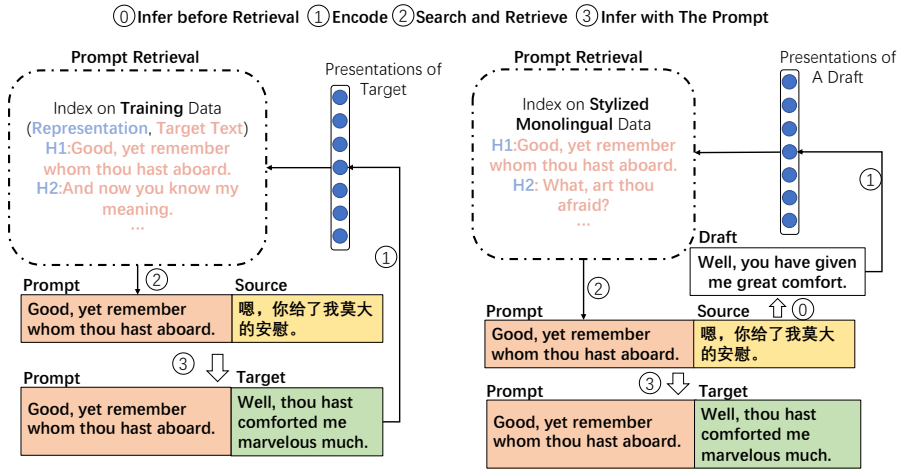

Figure 2: The proposed model training and inference process. We provide an example and mark the entire retrieval process in order. During training process, it includes encoding, retrieving and predicting, while it has an extra predicting operation during inference.

the similarity between the query representation h and keys, we can extract the needed sentence y:

$$y = \arg\min_{r \in \mathcal{D}} Distance(h, r)$$

where the search tool is Faiss(Johnson et al., 2021), a library for efficient similarity search and clustering of dense vectors.

4.2 Training

In the training stage, the goal is to retrieve a similar prompt with the current sentence to make the model adapt to the inference pattern. Specifically, we iterate each target-side sentence in the basic parallel data as a query and retrieve the most similar sentence as its prompt.

After obtaining the prompt, we concatenate the prompt and the query sentence by a special token as:

$$prompt, [s], src \rightarrow prompt, [s], trg$$

We train the model with this kind of data and normal data together to learn the prompt-based generation as well as basic translation.

4.3 Inference

In the inference stage, we first translate the source sentence roughly. Then the draft hypothesis is used as the query. The candidate datastore is constructed with the monolingual data in the given style. After retrieving the prompt, we append it to the beginning of the source sentence with the special token. After the second inference, the hypothesis can be obtained by splitting the token.

4.4 Advantages

We conclude the advantages of StyleAP as follows:

- StyleAP does not need any architecture modification and is easy to deploy.
- StyleAP does not need to assign various tags to all kinds of styles.
- Extra tuning is no longer needed when it comes to a new style. We only need to retrieve the prompt from the new monolingual stylized corpus and then generate the given style.

5 Experiments

In this section, we will introduce the details of our experiments.

5.1 Setup

We first compare our method with other baseline models on four tasks. Then, we design a manual evaluation to assess whether our method maintains translation quality and achieves diversity. All experiments are implemented in the following settings.

5.1.1 Data & Preprocessing

In the previous section, we introduce our provided stylized NMT benchmark MSMT. Our designed experiments and analysis are based upon this benchmark. The statistics information of this benchmark is shown in the Table 1.

We use SentencePiece(Kudo and Richardson, 2018) to jointly learn an unsupervised tokenizer. We preprocess the training data and filter the parallel sentences with length greater than 256. We set

	en→ zh		zh→en		en→ko		en→pt		
Styles	Modern	Classical	Modern	Early	Honorific	Non-hono	European	Brazilian	Average
Baseline	25.00	13.86	26.73	14.28	20.65	17.48	31.30	32.86	22.77
Transfer	24.87	20.88	11.05	7.46	<5	<5	32.84	32.59	<20
Tag-tuning	28.43	21.21	**27.16**	14.48	21.05	**21.11**	33.67	33.84	25.11
StyleAP	**29.73**	**24.98**	26.76	**17.72**	**21.65**	20.67	**33.82**	**34.27**	**26.20**

Table 2: BLEU results on the multiple stylized references. The experiment of en→ko Translation Transfer fails and yields non-sense results due to the data scarcity. Overall, StyleAP achieves the best results.

hyper-parameters min frequency 5 and max vocabulary 32k.

5.1.2 Implementation Details

Here, we introduce more details of our experiment settings. Our experiment is implemented on the open source Seq2Seq tool Neurst[10] (Zhao et al., 2021). Our seq2seq model uses a transformer-base structure with 6 encoder layers and 6 decoder layers, attention with a layer size of 512, and word representations of size 512. We apply post-layer normalization (Ba et al., 2016), adding dropout to embeddings and attention layers with a dropout rate of 0.1. We tie the source and target embeddings. The main training parameters are as follows. We use Adam Optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We use label smoothed cross entropy as a criterion with a label smoothing rate of 0.1. We set batch size per GPU 4096 and batch by tokens. And, we use four A100 GPUs to train our model from scratch. We save checkpoints every 1000 steps and stop training when there is no improvement in the continuous 50 checkpoints.

5.1.3 Comparing Systems

We use sacreBLEU (Post, 2018) as our metrics and compare StyleAP with three common systems:

- **Baseline:** Transformer that is trained on the raw parallel data.
- **Transfer:** A two-phases processing: Translate first and conduct style transfer (Syed et al., 2020). We train the translation model with normal parallel data and train the transfer model with stylized data.
- **Tag-tuning:** A tag-based model which is generally used in other work (Sennrich et al., 2016). They add a special token as the tag at the start of the source text with the known style. In this way, the model can generate different styles with different tokens. This method needs explicit extra fine-tuning.

[10] https://github.com/bytedance/neurst

	en→zh		zh→en		en→ko		en→pt	
Styles	M	C	M	E	H	N	E	B
Base	3.5	**3.6**	**3.6**	**3.7**	1.9	1.9	2.7	2.8
StyleAP	**3.6**	**3.6**	3.5	3.6	**2.0**	**2.0**	**2.9**	**2.9**

Table 3: The quality of all systems, ranging from 0 to 4. StyleAP maintains a comparable quality even the style is transformed.

5.2 Results

As is shown in Table 2, we calculate the BLEU score on the test set to compare StyleAP with the mentioned baselines. The Transfer methods have many drawbacks. Not only does it need two-phases training, but also yields poor results. The attempt in English-to-Korean direction even fails. Tag-tuning gains some improvements and even achieves the best performance in some directions. But overall, StyleAP obtains the best results and outperform the other methods. At the same time, StyleAP needs no extra tag or extra tuning when it comes to new styles. After acquiring the ability to translate with style prompt, StyleAP can handle various styles.

5.3 Human Evaluation

We also design a human evaluation to manually check the style transfer ratio as well as the translation quality preservation during the transfer. The quality score ranges from 0 to 4. The style transfer ratio means the percentage of the hypothesis that meets the required style, ranging from 0 to 100. Refer to the appendix for the specific scoring criteria and rules.

5.3.1 Quality Preservation

The quality results are in Table 3. StyleAP achieves a comparable performance with the baseline model, which means little semantic loss within the stylized translation.

5.3.2 Style Transfer Ratio

As is shown in Figure 3, StyleAP significantly enhances the transfer ratio. The only unsatisfactory is the Chinese-to-English Translation in the Early

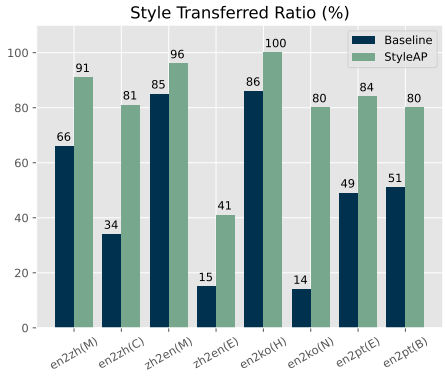

Figure 3: The percentage of successfully transferred sentences, ranging from 0 to 100. StyleAP significantly enhances the generation ratio of certain styles.

Modern style. The reason is that many sentences in the test set are very short like "What about that?" vs "What of that?". The styles for this kind of extremely short sentences are not meaningful.

5.3.3 Conclusion

The quality scores are comparable with the baseline model, while the transferred ratios are much higher than the baseline model in the four tasks. That indicates that our method can effectively translate the source text into a sentence with specific attributes without quality loss.

6 Analysis

6.1 Retrieval Strategy Matters

There are many retrieval methods to select a similar sentence from stylized monolingual sentences. We conduct a detailed comparison in English-to-Chinese inference with the following strategies:

- **Source:** Directly use the source text representation generated from the pre-trained multilingual language model.
- **Random:** Randomly choose a prompt from candidates.
- **Fixed:** Use the same prompt for all samples.

The results are shown in Table 4. We can see that the our strategy performs the best and the other retrieval methods face different levels of the BLEU loss. The retrieval strategy plays an important role in the translation.

	Modern	Classical
StyleAP	29.73	24.98
-Source	25.51	18.07
-Random	24.72	15.65
-Fixed	24.52	13.75

Table 4: Our retrieval strategy achieves the best results of BLEU.

	Modern	Classical
StyleAP	29.73	24.98
StyleAP (U)	28.72	23.76
Tag-tuning	28.43	21.21

Table 5: Unsupervised StyleAP still gains the better performance than Tag-tuning.

6.2 Even Unsupervised Prompts Works

In the training phase, we assume that the retrieval range lies within the specific styles. However, one condition that is more close to the real world is that we need to retrieve prompts from more general data, which may cause the style mismatch between the sentence and the prompt. Therefore, we also conduct a unsupervised prompt retrieval in training for English-to-Chinese direction.

As is shown in Table 5, the unsupervised version of StyleAP slightly drops in terms of BLEU but still outperforms Tag-tuning. It is worth mentioning that we have none of the style label of parallel data in this setting. General parallel data and monolingual stylized data is all you need. This again shows the universality and robustness of StyleAP.

6.3 Consistent Performance across Sizes

We are also interested in the situation of unbalanced data and even few stylized data. We implement a comparative experiment of stylized monolingual data on the en→zh task. We control the amount of labeled data at four levels: 1 million, 100 thousand, 10 thousand, and 1 thousand. For the tag-based method, we train the tag-based model from scratch at each level. For a fair comparison, we use the same stylized labeled data as the tag-based method but only as the target monolingual part for retrieval.

The results are shown in Figure 4 , where the horizontal axis represents the sample size, and the vertical axis represents the BLEU score. For the classical direction, our method performs better in all situations. Even when we only use 1,000 labeled stylized monolingual sentences, there is still an

Source	现在， 亲爱的奶妈， 哦上帝， 你 为什么 看起来 这么伤心? (Now) (good sweet Nurse) (Oh Lord) (you) (why) (look) (so sad)
Ref (E)	Now, good sweet Nurse, **O** Lord, why **look'st thou** sad?
Baseline	Now, good sweet Nurse, Oh Lord, why do you look so sad?
Prompt	What say'st **thou**, my dear nurse?
Tagged	Now, dear nurse, **O** God, why look you so sad?
StyleAP	Now, sweet nurse, **O** God, why **dost thou** look so sad?

Table 6: An example of using StyleAP to translate Chinese sentences to the Early Modern English style. The first two rows are the Chinese sentence and corresponding English translation of each Chinese word, respectively.

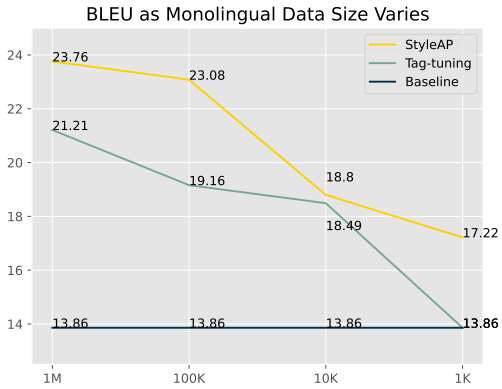

Figure 4: The impact of the size of stylized corpus on the en→zh task. StyleAP shows a consistent performance across all sizes, even the extremely few one.

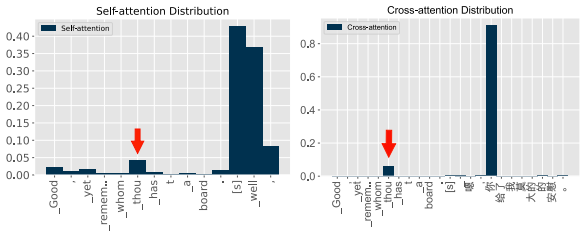

Figure 5: Attention score histograms, left for self-attention, right for cross-attention. The sample comes from the hypothesis "Well, thou hast given me great comfort." when predicting the token "_thou" by our model. The figure shows the style effect with the attention mechanism.

improvement compared with the baseline model. On the contrary, the tag-based method performs poorly in few data and even has a lower BLEU score than the baseline model.

In conclusion, our method has an overall better performance than the tag-based method at different levels. Especially for extremely few samples, our method still gains significant improvements.

6.4 Attention Score Interprets the Effect

We are also interested in how the retrieved prompt affects the translation style. Here is an example of Chinese-to-English task in Figure 5. The model is translating a Chinese sentence meaning "*Yeah, you have given me great comfort.*" into English and the next token is "thou", which means "you" in early modern English.

We show the average attention scores in the Transformer decoder, left for self-attention and right for cross-attention. For self-attention, except for some adjacent tokens, the model mainly pays attention to the token "thou" in the prompt which corresponds to the generating token. For the cross-attention, the model concentrates on the corresponding Chinese token "Ni" (means "you" in Chinese) and the token "thou" in the prompt again.

This result suggests that our retrieval prompt could affect the generation process through the attention mechanism.

6.5 Case Study

Finally, one stylized Chinese-to-English translation example is listed in Table 6 to show the effectiveness of our method more intuitively. As in these examples, the Chinese word "Ni" ("you" in English) is translated into "you" in Baseline and the Tag-tuning method. However, under the guidance of the prompt which uses the early modern English word "thou", StyleAP translates the word into "thou" correspondingly. Obviously, StyleAP can explicitly affect the translation style with prompts.

7 Conclusions

In cross-lingual generation fields, most studies focus on the translation quality but ignore the style issue, which happens to be important in the real-world communication. However, the previous studies face two major challenges including the bench-

mark as well as the approach. For those purposes, we re-visit this task and propose a standard stylized NMT benchmark MSMT with four well-defined tasks to push the boundary of this field. We also propose a new translation style controlling method with activation prompt. With stylized prompts that are retrieved from the stylized monolingual corpus, we successfully guide the translation generation style without iterative fine-tuning. Through automatic evaluation and human evaluation, our method achieves a remarkable improvement over baselines and other methods. A series of analysis also show the advantages of our method.

Limitation

One limitation of StyleAP is that one extra inference is needed for retrieval. It is mainly due to the monolingual retrieval accuracy is higher than that of crosslingual retrieval (refer to Section 6.1). In the future, we will try stronger multilingual model to mitigate this effect.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. Weiguo Zheng is the corresponding author.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2021. PADA: A prompt-based autoregressive approach for adaptation to unseen domains. *CoRR*, abs/2102.12206.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7307–7318. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. Controlling japanese honorifics in english-to-japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation, WAT@EMNLP-IJCNLP 2019, Hong Kong, China, November 4, 2019*, pages 45–53. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: word-level adversarial reprogramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4921–4933. Association for Computational Linguistics.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2020. Neural style transfer: A review. *IEEE Trans. Vis. Comput. Graph.*, 26(11):3365–3385.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 312–318. Association for Computational Linguistics.

Eugene Nida and Charles Taber. 2021. The theory and practice of translation:(fourth impression). In *The Theory and Practice of Translation*. Brill.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8568–8575. AAAI Press.

Xing Niu, Marianna J. Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2814–2819. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 35–40. The Association for Computational Linguistics.

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xinyu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8976–8983. AAAI Press.

Zewei Sun, Qingnan Jiang, Shujian Huang, Jun Cao, Shanbo Cheng, and Mingxuan Wang. 2022a. Zero-shot domain adaptation for neural machine translation with retrieved phrase-level prompts. *CoRR*, abs/2209.11409.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022b. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. Association for Computational Linguistics.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9008–9015. AAAI Press.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3170–3179. Association for Computational Linguistics.

Yue Wang, Cuong Hoang, and Marcello Federico. 2021. Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1193–1199. Association for Computational Linguistics.

Xuanfu Wu, Yang Feng, and Chenze Shao. 2020a. Generating diverse translation from model distribution with dropout. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1088–1097. Association for Computational Linguistics.

Xuanxuan Wu, Jian Liu, Xinjie Li, Jinan Xu, Yufeng Chen, Yujie Zhang, and Hui Huang. 2021. Improving stylized neural machine translation with iterative dual knowledge transfer. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3971–3977. ijcai.org.

Yu Wu, Yunli Wang, and Shujie Liu. 2020b. A dataset for low-resource stylized sequence-to-sequence generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9290–9297. AAAI Press.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in japanese-to-english neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation, WAT@COLING 2016, Osaka, Japan, December 2016*, pages 203–210. The COLING 2016 Organizing Committee.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2021. Neurst: Neural speech translation toolkit. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021*, pages 55–62. Association for Computational Linguistics.

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Yuyang Huang, Liwei Wu, and Mingxuan Wang. 2022. Beyond triplet: Leveraging the most data for multimodal machine translation. *CoRR*, abs/2212.10313.

A Appendix

In this section, we supplement human evaluation criterion in this paper and more stylized translation cases. Table 7 fully illustrates the score standard of our language experts. Table 8 shows more English examples of stylized translation.

	Accuracy	Criterion	Description
Quality	4	The translation faithfully reflects the semantics and the translation is fluent.	There is no errors and no modification required.
	3	The translated text basically reflects the semantics of the original text and is basically fluent(the subject, predicate, object and other grammatical components are in correct order), but there are a few non-keywords that are improperly used or inappropriately matched, etc. There are slight mistakes, which will not affect the understanding of the original text such as improper use of words, punctuation, capitalization, irregular date format, etc.	The meaning is basically correct, but there are partial errors, which will cause certain difficulties in understanding.
	2	The translation can reflect the semantics of the original text, the translation has one or more general errors, and the translation is basically fluent (the order of grammatical components such as subject, predicate and object is correct), but there are key words that express the semantics of improper translation, omission or mistranslation of non-keywords, etc.	The meaning is basically correct, but there are partial errors, which will cause certain difficulties in understanding.
	1	The translated text cannot reflect the semantics of the original text, and there are multiple serious translation errors in the translation text. One of the following situations exists: a) The translation text contains the main components of the original text, but fails to form a fluent composition due to sequence problems, logical errors, serious grammatical errors (including tenses), etc. The translation; b) The translation is basically fluent, but there are translation errors such as negation and double negation, serious omission of translation, mistranslation of keywords, and more translation of content that is not in the original text.	There are serious errors that have a greater impact on understanding.
	0	The translated text cannot express the meaning of the original text at all: a) The translation text is obscure and difficult to understand, and the content expressed in the original text cannot be judged by the translation text; b) A string of repeated words and garbled characters appear; totally different/completely unrelated; d) the entire sentence is not translated.	The translated text is almost completely wrong or completely incomprehensible.
Style	1	The translation has the corresponding style.	
	0	The translation has not the corresponding style.	

Table 7: Human evaluation criterion.

Modern English	Baseline	Retrieved Prompt	StyleAP
I swear to you, You have a good heart, and believe me, I'll tell her that.	I swear to you, you are kind, trust me, and I'll tell her.	I tell **thee**, I, that **thou** hast marred her gown.	I swear to **thee**, **thou** art kind, and believe me, I'll tell her.
Now I'll tell you so you don't have to ask.	Now I tell you, so you need not ask.	To tell **thee thou** shalt see me at Philippi.	Now I tell **thee**, so **thou** shalt not ask.
You're not paying attention to me.	You did not notice me at all.	God **mark thee** to his grace!	**Thou dost** not mark me.
If you were ever yourself, and this sadness was yours, you and your sadness were all for Rosaline.	If you were once yourself, This sorrow is yours, and both you and your sorrow are for Rosalin.	If e'er **thou** wast **thyself** and these woes thine, **thou** and these woes were all for Rosaline.	If **thou** werest thyself, This sorrow was **thy**, **Thou** and **thy** sorrow were all for Rosaline.
Therefore, the fact that you're awake this early tells me you've been upset with some anxiety.	Therefore you wake so early, which tells me you are uneasy about some anxiety.	Unless **thou tell'st** me where **thou** hadst this ring, **Thou** diest within this hour.	Therefore, **thou awaken'st** so early, That tells me **thou** art uneasy with some anxiety.

Table 8: Examples from test sets and the results of the baseline and StyleAP.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4,5

☑ B1. Did you cite the creators of artifacts you used?
3,4,5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3,4,5

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3,4,5

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3,4,5

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3,4,5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3,4,5

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5,6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
5

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
A

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.