How Transferable are Attribute Controllers on Pretrained Multilingual Translation Models?

Danni Liu and **Jan Niehues**
Karlsruhe Institute of Technology, Germany
{danni.liu, jan.niehues}@kit.edu

Abstract

Customizing machine translation models to comply with desired attributes (e.g., formality or grammatical gender) is a well-studied topic. However, most current approaches rely on (semi-)supervised data with attribute annotations. This data scarcity bottlenecks democratizing such customization possibilities to a wider range of languages, particularly lower-resource ones. This gap is out of sync with recent progress in pretrained massively multilingual translation models. In response, we transfer the attribute controlling capabilities to languages without attribute-annotated data with an NLLB-200 model as a foundation. Inspired by techniques from controllable generation, we employ a gradient-based inference-time controller to steer the pretrained model. The controller transfers well to zero-shot conditions, as it operates on pretrained multilingual representations and is attribute- rather than language-specific. With a comprehensive comparison to finetuning-based control, we demonstrate that, despite finetuning's clear dominance in supervised settings, the gap to inference-time control closes when moving to zero-shot conditions, especially with new and distant target languages. The latter also shows stronger domain robustness. We further show that our inference-time control complements finetuning. A human evaluation on a real low-resource language, Bengali, confirms our findings. Our code is here.

1 Introduction

Pretrained multilingual translation models with massive coverage (Zhang et al., 2020; Liu et al., 2020; Fan et al., 2021; Xue et al., 2021; NLLB Team et al., 2022) have become of the backbone of many translation systems. While their off-the-shelf translation quality has been constantly improving (Fan et al., 2021; Ma et al., 2021; NLLB Team et al., 2022), the flexibility of customization towards desired attributes, such as formality or grammatical gender, is another important metric.

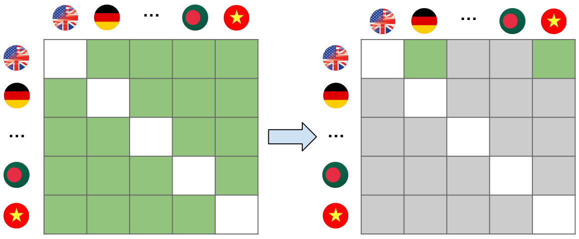

Figure 1: The number of translation directions with attribute-annotated data (**right**) is far less than that of what massively pretrained models serve (**left**).

Adapting generic systems for attribute-controlled translation relies on training data with attribute information. Creating such annotated data often requires language-specific knowledge and manual curation. This makes data acquisition challenging even for single languages. When scaling to the numerous directions served by massively multilingual models, it quickly becomes impractical, as shown in Figure 1. While prior works (Michel and Neubig, 2018; Saunders et al., 2020; Nadejde et al., 2022) showed promising results of finetuning on limited attribute-annotated data, to allow other languages without supervised data to similarly benefit from the customization possibilities, the *transferability* of the attribute controllers remains to be studied.

A straightforward way to achieve attribute control is finetuning on attribute-specific data. Recent works (Rippeth et al., 2022; Wu et al., 2023) have shown that finetuning with just hundreds of attribute-specific sentences is sufficient. However, small finetuning data also brings the risk of overfitting and catastrophic forgetting (Freitag and Al-Onaizan, 2016; Thompson et al., 2019). It is especially relevant when generalizing to new languages, where finetuning on some languages may erase the knowledge of others from pretraining (Garcia et al., 2021; Cooper Stickland et al., 2021; Liu and Niehues, 2022). While these issues may be mitigated by partial finetuning (Houlsby et al., 2019; Bapna and Firat, 2019), domain mismatch between

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 334–348
March 17-22, 2024 ©2024 Association for Computational Linguistics

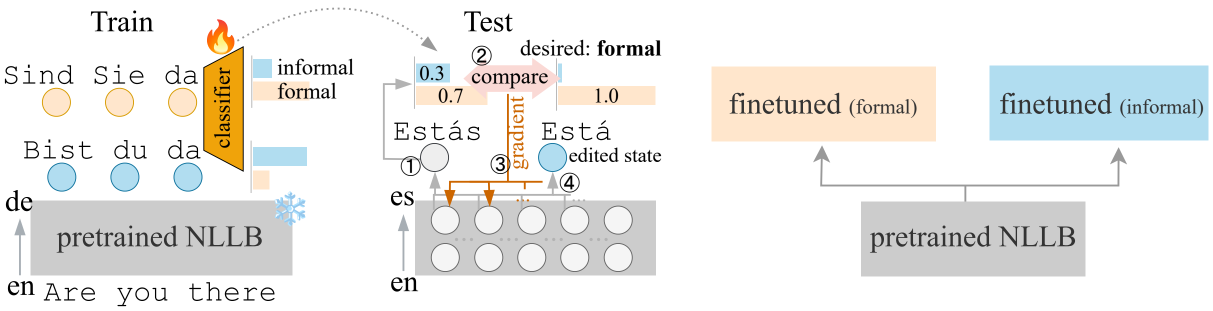

Figure 2: Left: **Inference**-time control by gradient-based classifier guidance: training classifiers for attributes on decoder activations, and using its predictions to edit inference-time model *activations* towards desired attributes. Right: Standard **training**-time control by finetuning on attribute-specific data.

the finetuning data and the test domain can still degrade translation quality. We will validate these concerns in various zero-shot conditions with different language relatedness and domains.

On the other end of the spectrum, inference-time customization is another paradigm of attribute control. In this case, the pretrained model is fully unchanged in the training stage. At inference time, the generation process is steered towards desired attributes by e.g. re-weighting entries in the output distribution (Saboo and Baumann, 2019; Yang and Klein, 2021; Landsman et al., 2022) or editing model activations (Dathathri et al., 2020). To enable cross-lingual transfer, the controller must be trained on features that are shared across languages. This precludes methods that operate on the *surface* vocabulary level. In this work, we will extend an activation-based approach (Dathathri et al., 2020) originally for decoder-only models to cross-lingual transfer on pretrained translation models.

Task Formalization We focus on the following task: Given a pretrained many-to-many multilingual translation model covering N languages and $N(N-1)$ translation directions, along with parallel data on k ($k \ll N(N-1)$) translation directions where the target translation corresponds to specific attributes (e.g., formality level), we aim to customize the pretrained model to translate with desired attributes for as many directions as possible. We refer to the subsequent model as an *attribute controller*. Specifically, after learning on the k sets of parallel data with attribute annotation, to what extent can we transfer the attribute controller to the remaining $N(N-1)-k$ translation directions?

2 Background and Related Work

Attribute-Controlled Translation Previous works investigated controlling various attributes of machine translation outputs, for instance politeness (Sennrich et al., 2016; Niu et al., 2018; Feely et al., 2019), gender (Vanmassenhove et al., 2018; Saunders et al., 2020), length (Takase and Okazaki, 2019; Lakew et al., 2019; Marchisio et al., 2019; Niehues, 2020), or style in general (Michel and Neubig, 2018; Schioppa et al., 2021; Vincent et al., 2023; Wang et al., 2023). As existing works mainly focus on supervised conditions with at least some supervised data, how these approaches generalize to new languages remains unclear. In face of data scarcity, one approach is to use synthetic data by pseudo-labeling (Rippeth et al., 2022; Lee et al., 2023). In our work, by building upon massively multilingual translation models, we *do not assume* the scalability of creating synthetic data for all languages served by the backend model, *nor do we assume* a classifier that can a priori distinguish attribute classes for zero-shot languages.

Multilinguality for Controllable Generation Our work is also related to controllable text generation in general. Despite steady progress in this field (Keskar et al., 2019; Krause et al., 2021; Yang and Klein, 2021; Liu et al., 2021), how the controller generalizes across languages is likewise less explored. With the recent surge of large language models (LLMs), attribute-controlled translation has also been addressed by prompting multilingual language models in a few-shot manner (Sarti et al., 2023; Garcia et al., 2023). Notably, Sarti et al. (2023) reported promising few- and zero-shot attribute control results using multilingual LLMs. In this work, we take a different perspective by using a pretrained dedicated encoder-decoder translation model as backend, and transferring the attribute control capabilities with lightweight add-ons. As currently open LLMs still lag behind dedicated translation models (Zhu et al., 2023; Sarti et al., 2023) especially on low-resource languages (Robinson et al., 2023), we believe improving the attribute control capabilities of massively multilingual conventional models is still highly relevant.

Multilingual Domain Adaptation Attribute control can be viewed as a light domain adaptation task. Prior works (Cooper Stickland et al., 2021; Vu et al., 2022) adapting pretrained multilingual models have reported catastrophic forgetting of languages absent from the finetuning stage. Our results on finetuning for zero-shot attribute control (§6.1) shows a different picture. One potential reason is that, compared to adapting to fully new domains such as medical or law texts, the attribute control task can be learned with less data. This in turn requires less intense finetuning and is therefore less vulnerable to forgetting.

3 Transferring Attribute Controllers for Multilingual Translation

To generalize to new translation directions, an ideal controller should be *attribute*- rather than *language*-specific. That is, its representation for different attribute labels varies little with specific languages.

Inference-Time Control by Classifier Guidance: Our first approach builds upon the observation that the activations of pretrained multilingual models capture commonalities of different languages (Pires et al., 2019; Liu et al., 2020). An attribute classifier trained on these activations can then potentially transfer across languages, which we use at inference time to steer the generation for languages without attribute-annotated data. The control takes effect on inference-time model *activations* instead of *parameters*, as shown in Figure 2. Specifically, we first train an attribute classifier while *freezing* the pretrained model, and then edit the model activations towards the wanted attribute based on the predicted label at inference time. This idea has shown success in controllable image synthesis (Dhariwal and Nichol, 2021) and text generation (Li et al., 2022). To the best of our knowledge, no prior work has explored it for cross-lingual transfer.

Specifically, we extend the approach by Dathathri et al. (2020) to encoder-decoder models. For machine translation, Given a frozen pretrained model, we run forward passes with attribute-annotated[1] parallel data $(\mathbf{X}, \mathbf{Y})^c$ for $c \in [C]$, where $\mathbf{X}$ and $\mathbf{Y}$ are the source and target sentences with individual sentence pairs $(\mathbf{x}, \mathbf{y})^i \in (\mathbf{X}, \mathbf{Y})$, and C is the number of attribute labels.

While freezing the translation model's parameters, we train a classifier that maximizes $P(c \mid \mathbf{h})$, where c is the ground-truth attribute label and $\mathbf{h}$ is the last decoder layer's hidden states after forced-decoding parallel data $(\mathbf{x}, \mathbf{y})$:

$$\mathbf{h} = \text{decoder}(\mathbf{y}, \text{encoder}(\mathbf{x})). \tag{1}$$

Like with a standard model, the output distribution is then $\text{softmax}(\mathbf{W}\mathbf{h})$, where $\mathbf{W}$ maps the hidden states $\mathbf{h}$ to the vocabulary distribution.

At inference time step t, the hidden state is:

$$\mathbf{h}_t = \text{decoder}(y_{t-1}, \mathbf{A}_{t-1}), \tag{2}$$

where y_{t-1} is the token from the previous step, and $\mathbf{A}_{t-1}$ is the model activations. $\mathbf{A}_{t-1}$ contains activation key-value pairs[2] from the decoder self-attention and cross-attention for steps 1 to $t-1$, and is cached in most Transformer decoding implementations (Ott et al., 2019; Wolf et al., 2020).

Based on all available decoder states till $t-1$, we predict an attribute label: $\text{argmax}_c P(c \mid \mathbf{h}_{1,\ldots,t-1})$. Following Dathathri et al. (2020), we meanpool the states from timestep 1 to $t-1$ for the prediction. It also empirically showed better performance than 1) using a token-level classifier without pooling and 2) operating on the cumulative sum of hidden states from all time steps so far.[3]

As $\mathbf{h}_{1,\ldots,t-1}$ is only determined by $\mathbf{A}_{t-1}$, we can rewrite $P(c \mid \mathbf{h}_{1,\ldots,t-1})$ as $P(c \mid \mathbf{A}_{t-1})$. Comparing the prediction to the desired attribute c^*, we can derive gradients measuring how much the current activations satisfy the desired c^*. The gradients, $\nabla_{\mathbf{A}_{t-1}} P(c^* \mid \mathbf{A}_{t-1})$, are then back-propagated for several iterations with given step sizes, resulting in updated activations $\tilde{\mathbf{A}}_{t-1}$, which further leads to modified decoder hidden state:

$$\tilde{\mathbf{h}}_t = \text{decoder}(y_{t-1}, \tilde{\mathbf{A}}_{t-1}). \tag{3}$$

A new output token y_t (that more likely satisfies the control) is generated from $\tilde{\mathbf{h}}_t$ by $\text{softmax}(\mathbf{W}\tilde{\mathbf{h}}_t)$.

Finetuning-Based Control: A more common way to realize control is finetuning the pretrained model on attribute-specific parallel data, as done in domain adaptation (Freitag and Al-Onaizan, 2016). To transfer to directions without annotated data,

[1] Only the target side needs attribute labels.

[2] Note these are not the key/value projection weights of the Transformer, but the activations after applying the projections.

[3] In initial experiments training an English-German formality classifier, the accuracy on the dev set was 86.7% (meanpool), 66.1% (token-level) and 73.0% (cumulative sum).

Task	Directions	# Sent. per lang. per att.
Formality control (formal/informal)		
train	en→{de, es, fr, hi, it}	400
test (supervised)	en→{de, es, fr, hi, it}	600
test (new tgt)	en→{pt, ru, ko}	600
test (new src)	{de, fr, hi, it}→es	366-572
Grammatical gender control (feminine/masculine)		
train	en→es	194
test (supervised)	en→es	552-556
test (new tgt)	en→{it, fr}	515-546
test (new src+tgt)	{es, fr}→it, {es, it}→fr	271-365

Table 1: Data overview. Codes: German (de), Spanish (es), French (fr), Hindi (hi), Italian (it), Korean (ko), Portuguese (pt), Russian (ru), source (src), target (tgt).

the adaptation step must mostly learn the desired *attributes* rather than the specific *languages* in finetuning, so as not to forget the languages without annotated data. On our tasks, naive finetuning already works effectively: We finetune the full model on each attribute, resulting in one specialized model per attribute as shown in Figure 2.[4] Partial finetuning e.g. with adapters (Bapna and Firat, 2019; Philip et al., 2020) is a more parameter-efficient approach. We do not explore partial finetuning in this work, as it does not fully align with our focus on the transferability of attribute controllers.

4 Experimental Setup

We experiment on two attribute control tasks: formality and grammatical gender control. As outlined in Table 1, the training data has English on the source side. For the target languages, there is one set of translations per attribute. The low data volume not only reflects the practical challenge of data acquisition, but is also an established condition in existing benchmarks (Nadejde et al., 2022).

4.1 Formality Control (In-Domain)

The training data come from CoCoA-MT (Nadejde et al., 2022)[5], where the test domain overlaps with training. For zero-shot conditions, we transfer controllers trained on different language pairs to new translation directions. Specifically, we investigate the following two cases:

Transfer to New Target Languages We train the attribute controllers on one or multiple target languages to assess the impact of multilinguality on transfer. We compare the following settings:

- **Single-direction**: We use en→es and de as representative Romance and Germanic languages;
- **Multilingual**: We train on all languages in the training data: en→{de, es, fr, hi, it}.

For the new target languages, we choose three directions from the IWSLT 2023 formality control shared task[6] (Agarwal et al., 2023): en→pt (**close**), en→ru (**related**), and en→ko (**distant**) for their different degrees of relatedness to the languages in training. Among them, en→ko has 400 sentences of supervised data. We use it to establish the oracle performance in the presence of supervised data.

Transfer to New Source Languages We re-align the CoCoA-MT test set using English as pivot, creating a new test set with non-English source and target sentences.[7] Unlike translating from English, here the source sentences also contain formality information. This allows testing if the model can: 1) *preserve* the source formality level; 2) *change* the source formality level when steered so.

4.2 Gender Control (Out-of-Domain)

For the formality control setup above, the data for training the attribute controller come from the same domain as the test set. To evaluate domain generalization, for grammatical gender control, we train the controller on texts with very different styles from the test data. For *training* the attribute controller, we use the en-es set from Saunders et al. (2020)[8] with artificial sentences of very simple grammatical structure up to 7 words. In contrast, for the *test* set we use MuST-SHE (Bentivogli et al., 2020), which consists of TED talks with much longer sentences and more versatile styles. More dataset details are Appendix A.2. Besides transfer to new target languages like previously (§4.1), we also explore the following setting:

Transfer to New Source & Target Languages The MuST-SHE test set comes in en-{es, fr, it}. Like previously, we re-align them using English as pivot, creating non-English source and target

[4] We tried prepending attribute tags to the source sentences (Chu et al., 2017; Kobus et al., 2017), but this was not enough to make the pretrained model to be attribute-aware. A potential reason is that the pretrained model tends ignore the source tags as noise, and that the low amount of finetuning data cannot re-establish the importance of the tags.

[5] We excluded Japanese, where our pretrained model has very low translation accuracy on formality-annotated words (<40%, whereas all 5 other languages score >60%).

[6] https://github.com/amazon-science/contrastive-controlled-mt/tree/main/IWSLT2023

[7] The original test sets only have English input. As the English sentences mostly overlap, we create new pairs of two non-English languages by matching their English translations.

[8] https://github.com/DCSaunders/tagged-gender-coref#adaptation-sets

sentences. In this case, both the source and target sentences have the same gender. As the attribute training data is in en→es, we evaluate {es, fr}→it and {es, it}→fr for the transfer to new translation directions where both the source and target languages differ from training.

4.3 Models and Evaluation

Models We use two types of backend models. For the main experiments, we use the pretrained NLLB-200 distilled 600M model (NLLB Team et al., 2022), which covers 200 languages for many-to-many translation. We also train a Transformer-base (Vaswani et al., 2017) from scratch to verify if observed phenomena are specific to models with massive multilingual pretraining. The Transformer-base model covers all languages in our experiments and is trained on OPUS-100 (Zhang et al., 2020). Details of these data are in Appendix A.1. Training and inference details are in Appendix B.

Control Evaluation For *formality* control, we report matched accuracy (M-Acc; %) following Nadejde et al. (2022). For *gender* control, we use the official evaluation script (Bentivogli et al., 2020) for accuracy (%). For formality, as the test set is the same for both formalities, the baseline M-Acc for the two formality labels add up to 1.0. This is not the case for gender control.

Quality Evaluation We use COMET↑ (Rei et al., 2020)[9] as the main translation quality metric, and additionally report BLEU↑[10] to compare to prior works. Note that BLEU is impacted by n-gram matches on the correct formality or gendered words, while COMET is less susceptible to the artifact. For COMET score comparisons, we run paired T-tests and bootstrap resampling using `comet-compare`. We use "*" or "†" to mark systems better or worse than the base pretrained model at $p = 0.05$.

Human Evaluation To test the transfer to real low-resource languages, we conduct a human evaluation on Bengali, which was marked as low-resource in the NLLB-200 training data (NLLB Team et al., 2022). Details on the evaluation are in Appendix C.

Baselines Few existing works experimented on the same data conditions as ours. An exception is the "`mBART-large Gold Finetuned`" model by Rippeth et al. (2022), who finetuned mBART (Liu et al., 2020) on parts of CoCoA-MT (Nadejde et al., 2022) for formality control. Their results overlap with our supervised results on en→{de, es, hi} and zero-shot results on en→ru. Other than this, the majority of prior works used *more relaxed* data conditions than ours, e.g., using an existing attribute classifier that covers *zero-shot* languages for pseudo-labeling (Lee et al., 2023) or hypothesis reranking (Wu et al., 2023). We report these results in Appendix D. Overall, our model's performance is comparable to the leading systems.

	Model	Formal	Informal	Avg.	BLEU	COMET$_{2022}$
	base	45.6	54.4	–	35.7±1.0	82.1
en→de	+CG	95.0	89.6	92.3	38.4±1.1	81.6†
	+FT	100.0	100.0	100.0	43.6±1.2	83.8*
	Rippeth et al.	93.6	77.4	85.5	37.4	–
	base	29.7	70.3	–	40.0±1.1	83.9
en→es	+CG	72.9	92.4	82.7	41.2±1.2	84.4*
	+FT	100.0	95.9	98.0	46.0±1.2	85.5*
	Rippeth et al.	96.7	82.7	89.7	38.3	–
	base	76.8	23.2	–	36.0±1.1	80.8
en→fr	+CG	99.8	77.2	88.5	38.8±1.2	80.9
	+FT	100.0	99.3	99.7	43.0 ±1.1	83.0*
	base	96.7	3.3	–	24.0±0.9	75.5
en→hi	+CG	99.3	30.7	65.0	24.3±0.9	75.0†
	+FT	99.6	99.2	99.4	36.4±1.0	81.7*
	Rippeth et al.	98.5	64.7	81.6	28.7	–
	base	3.2	96.8	–	41.3±1.1	84.9
en→it	+CG	18.7	99.5	59.1	40.6±1.1	84.1†
	+FT	98.6	99.3	99.0	49.6±1.1	86.0*

Table 2: Formality control results in *supervised* condition (controllers trained on formality-annotated data).

	Model	Feminine	Masculine	Global	BLEU	COMET$_{2022}$
	base	58.8	86.7	73.6	45.0±1.2	84.9
en→es	+CG	75.0	89.7	82.8	44.7±1.2	84.7
	+FT	90.2	89.7	86.9	43.7±1.2	84.0†

Table 3: Grammatical gender control results in *supervised* condition (cross-domain: controller trained on gender-annotated data from a different domain).

5 Supervised Conditions

Table 2 and Table 3 show formality and gender control results respectively with supervised controllers on NLLB-200. Overall, both finetuning and CG are able to steer the output towards given attributes, while maintaining the original translation quality or at the cost of a slight degradation.

Finetuning more effective than classifier guidance in supervised conditions: A comparison of scores in Table 2 and Table 3 clearly shows FT is more effective than CG. For formality control, FT consistently scores nearly 100% M-Acc.

[9] with `Unbabel/wmt22-comet-da` (×100 for readability)

[10] using sacreBLEU (Post, 2018) with confidence intervals: bs:1000|rs:12345|c:mixed|e:no|tok:13a|s:exp|v:2.3.1

		Pretrained Massively Multilingual					Transformer-base				
	Model	**Formal**	**Informal**	**Avg.**	**BLEU**	**COMET$_{2022}$**	**Formal**	**Informal**	**Avg.**	**BLEU**	**COMET$_{2022}$**
en→pt	base	47.7	52.3	–	41.7±1.1	85.1	35.8	64.2	–	38.7±1.1	82.2
	+CG (de)	75.6	74.2	74.9	43.0±1.1	84.9	50.0	72.8	61.4	38.7±1.1	81.8†
	+FT (de)	99.0	45.5	72.3	40.4±1.0	85.3	79.2	71.2	75.2	39.6±1.1	82.7*
	+CG (es)	85.4	83.6	84.5	43.8±1.0	85.0	53.3	79.8	66.6	38.8±1.1	81.7†
	+FT (es)	99.8	28.7	64.3	40.3±1.0	85.2	93.9	80.1	87.0	40.5±1.0	82.5*
	+CG (multi)	84.8	80.0	82.4	43.7±1.1	84.9	55.9	80.8	68.4	39.0±1.1	81.8†
	+FT (multi)	99.5	51.0	75.3	42.3±1.0	85.9*	95.8	81.9	88.9	41.4±1.0	83.1*
	+CG +FT (multi)	100.0	83.2	**91.6**	42.1±1.0	85.7*	97.8	93.7	**95.8**	41.0±1.0	82.4
en→ru	base	55.0	45.0	–	30.3±1.0	83.7	43.9	56.1	–	24.2±1.0	75.9
	+CG (de)	87.3	77.7	82.5	32.2±1.0	83.1	67.2	71.8	69.5	24.6±0.9	75.0†
	+FT (de)	99.5	84.7	92.1	33.0±1.1	84.2*	84.0	69.3	76.7	25.0±1.0	75.8
	+CG (es)	86.8	73.9	80.4	32.4±1.0	83.2	61.7	76.8	69.5	24.8±1.0	75.0†
	+FT (es)	98.3	60.6	79.5	32.8±1.1	84.1*	83.5	68.6	76.1	26.1±1.0	76.6*
	+CG (multi)	87.3	78.2	82.8	32.2±1.0	83.2	72.2	80.9	76.6	25.0±1.0	75.0†
	+FT (multi)	99.8	79.6	89.7	33.0±1.1	84.2*	87.5	69.8	78.7	25.9±1.0	77.0*
	+CG +FT (multi)	100.0	93.0	**96.5**	33.1±1.0	84.4*	96.2	91.3	**93.8**	26.2±1.0	76.2
	Rippeth et al. (2022)	100.0	13.8	56.9	23.5	–	–	–	–	–	–
en→ko	base	50.9	49.1	–	15.7±0.7	82.6	32.0	68.0	–	10.6±0.6	74.0
	+CG (de)	67.0	64.6	65.8	15.7±0.7	82.1†	45.2	78.2	61.7	10.4±0.6	73.4†
	+FT (de)	67.8	54.2	61.0	12.8±0.6	84.1*	42.7	66.4	54.6	10.7±0.6	74.0
	+CG (es)	68.9	61.6	65.3	15.1±0.8	82.1†	46.3	77.6	62.0	10.7±0.6	74.1
	+FT (es)	64.4	47.3	55.9	14.0±0.7	84.4*	47.4	62.7	55.1	11.7±0.6	75.2*
	+CG (multi)	67.0	61.7	64.4	15.5±0.8	82.2	46.0	78.1	62.1	10.6±0.6	74.1
	+FT (multi)	68.5	46.2	57.4	13.4±0.7	84.7*	48.3	68.4	58.4	11.0±0.6	74.4
	+CG +FT (multi)	70.0	63.5	**66.8**	13.2±0.7	84.2*	58.9	81.8	**70.4**	10.8±0.6	73.4†
	+oracle CG (ko)	70.3	62.6	66.5	15.2±0.7	81.7†	58.9	82.3	70.6	11.2±0.6	74.5*
	+oracle FT (ko)	79.4	93.5	86.5	22.2±0.9	86.2*	86.7	97.9	92.3	19.1±0.9	74.0*

Table 4: Zero-shot formality control results. **Best** and second best results under the same data condition are marked.

It also substantially improves the quality scores due to adapting towards the specific domain of the attribute-annotated data, which is the same as the test domain in this case. On the other hand for CG, while it also improves the formality accuracy, the scores lag behind finetuning in both accuracy and quality. The gap is especially prominent on hi and it, where the underlying NLLB model has a strong bias towards a single formality: the accuracy for the rare formality is nearly zero (3.3% and 3.2% respectively). This is likely to do with NLLB's training data, which might be skewed towards one single formality for some languages. In this case, CG can only partly recover the ability to generate translation in the formality NLLB is unfamiliar with. These results indicate that CG is only effective when the underlying model does not suffer from an absolute bias towards one attribute.

Classifier guidance more robust to domain mismatch: As motivated in §4.2, the gender control results in Table 3 allow us to assess the impact of domain mismatch between the controller training data and the test data, a very realistic scenario in practice. Here, while finetuning achieves higher accuracy for gendered words, it also degrades translation quality by 0.9 COMET. This provides further evidence that the previously improved COMET scores (Table 2) are results of finetuning on in-domain data. In contrast, the translation quality with CG does not significantly differ from NLLB by the T-tests, suggesting its stronger domain robustness. We hypothesize it is because CG operates on the last decoder layer's hidden states, which are just one projection away from the output vocabulary. These representations likely contain more word-level than domain information, which is precisely needed in the task of attribute control.

6 Zero-Shot Conditions

6.1 New Target Languages

Now we transfer the trained controllers to target languages unseen when training the controllers, i.e., those without attribute annotation. In Table 4 and Table 5, we report the results on formality and gender control respectively. In Table 4, we also compare the single-direction and multilingual controllers as motivated in §4.1.

	Model	Feminine	Masculine	Global	BLEU	COMET
en→it	base	53.8	88.9	73.1	35.1±1.0	84.1
	+CG	72.3	92.8	83.6	35.4±1.1	83.7
	+FT	83.6	91.2	87.8	34.4±1.0	83.5†
	+CG +FT	88.6	94.5	**91.8**	33.4±1.0	82.6†
en→fr	base	55.3	88.4	72.4	38.3±1.3	82.6
	+CG	67.8	90.3	79.4	38.7±1.2	82.5
	+FT	78.9	90.8	85.0	38.2±1.2	82.0†
	+CG +FT	87.0	91.9	**89.5**	37.4±1.2	81.9†

Table 5: Zero-shot grammatical gender control results on *new target* languages with *domain mismatch*.

Gap between finetuning and classifier guidance shrinks in zero-shot conditions: While finetuning was consistently leading in supervised conditions (§5), now under zero-shot conditions with unseen target languages, the gap shrinks. For formality control, on Korean, the most distant language, CG consistently achieves stronger control results than finetuning, indicating more robustness when transferring to unfamiliar settings. Overall in Table 4, for the main experiments on NLLB-200, CG outperforms FT in 7 of the 9 pairwise comparisons ({de, es, multi} × 3 target languages). With gender control results in Table 5, finetuning achieves stronger control accuracy (avg. +4.9% abs.) but degrades translation quality (−0.6 COMET) due to domain mismatch. On the other hand, CG retains the translation quality. This confirms the previous finding (§5) on its stronger domain robustness.

Multilingual controllers help when the base model is not massively multilingual: In Table 4, controllers trained on multiple translation directions (multi) are compared to those trained on single directions (en→es or de). On Transformer-base, multi consistently outperforms its single-direction counterparts, regardless whether the controller is finetuning- or CG-based. In contrast, for the pretrained NLLB, there is no clear distinction between the multilingual systems and rest. This indicates that NLLB does not further benefit from multilinguality in the controller training stage, likely because it already underwent a massively multilingual pretraining stage. This shows that massively multilingual models are a useful basis for attribute control especially when annotated resources are limited to single languages.

Classifier guidance is complementary with finetuning: When applying CG on top of the finetuned models, we see the *strongest* control accuracy for both formality and gender control. This observation is consistent whether the base model is the pretrained NLLB or the normal Transformer-base. Compared to finetuning alone, the addition of CG also does not degrade translation quality on NLLB. On the more challenging case of gender control which involves domain mismatch, adding CG to finetuning does not impact translation quality on fr and causes a slight degradation on it. This is likely linked to poor hyperparameter choices in CG: due to time constraints we directly used the hyperparameters when applying CG alone, which are too strong for models already finetuned for attribute control. We are optimistic for improved scores under more fitting hyperparameters.

Model	Quality	Formality	Win	Win & Tie
	(1-5)	(1-3)	(%)	(%)
(1) NLLB-200	4.25±0.75	2.69±0.46	–	–
(2) CG (multi) formal	4.00±0.79	2.63±0.48	56.3	81.3
(3) CG (multi) inf.	4.44±0.70	2.38±0.69	62.5	93.8
(4) FT (multi) formal	4.31±0.85	2.63±0.48	43.8	68.8
(5) FT (multi) inf.	4.13±1.05	2.44±0.49	62.5	93.8

Table 6: Human evaluation on Bengali, with quality on a 5-point scale↑ and formality on a 3-point scale (↑: formal) with standard deviations. Last two columns show pairwise comparison of formality scores to baseline NLLB-200 given the same source sentences (winning: scoring more in the direction of the desired formality).

Finetuning did not erase knowledge on other languages: To our surprise and different from results in domain adaptation (Cooper Stickland et al., 2021; Vu et al., 2022), finetuning did not erase the pretrained model's knowledge on the target languages absent in supervised finetuning, as reflected by the translation quality scores (Table 4, 5). This is not specific to NLLB, but also observed on the Transformer-base trained with random initialization on a few translation directions. Therefore, this phenomenon is not a result of massively multilingual pretraining, but more likely linked to the light finetuning strength with limited number of updates and small learning rates.

Comparison to oracle data condition: In the bottom rows of Table 4, we report the oracle performance of using 400 sentences as supervised data for training the controllers. Our strongest zero-shot results match the performance of oracle CG, but still lag far behind the upper-bound of finetuning on in-domain data with attribute annotation (oracle FT). We believe this gap is magnified as Korean is not only linguistically distant from the languages

		Source Formal					Source Informal				
	Model	Formal	Informal	Avg.	BLEU	COMET$_{2022}$	Formal	Informal	Avg.	BLEU	COMET$_{2022}$
X→de	base	77.8	22.2	–	23.9±0.5	79.0	48.5	51.5	–	24.7±0.5	79.3
	+CG	98.6	71.5	85.1	25.9±0.5	78.7	94.0	87.7	90.9	27.0±0.5	79.0
	+FT	100.0	100.0	100.0	30.1±0.7	80.7*	100.0	99.7	99.9	30.0±0.6	80.7*
X→es	base	57.8	42.2	–	29.8±0.5	82.7	20.3	79.7	–	29.9±0.5	82.7
	+CG	86.7	73.3	80.0	30.5±0.8	82.3	67.5	93.7	80.6	31.1±0.6	82.3
	+FT	99.6	77.4	88.5	32.8±0.7	83.9*	99.8	97.8	98.8	33.2±0.7	83.9*
X→fr	base	97.0	3.0	–	29.5±0.6	79.1	87.7	12.3	–	30.3±0.6	79.6
	+CG	99.8	40.4	70.1	30.6±0.6	78.9	99.9	59.5	79.7	32.5±0.6	79.6
	+FT	99.9	99.4	99.7	34.2±0.7	81.0*	100.0	100.0	100.0	35.6±0.6	81.5*
X→hi	base	98.2	1.8	–	20.2±0.4	73.2	98.4	1.6	–	20.8±0.4	73.6
	+CG	99.2	9.8	54.5	20.3±0.4	73.0	99.2	12.3	55.7	20.8±0.4	73.4
	+FT	99.4	99.3	99.4	26.5±0.6	75.3*	99.7	99.5	99.6	27.7±0.6	75.8*
X→it	base	23.0	77.0	–	27.6±0.6	83.5	1.5	98.5	–	28.0±0.6	83.6
	+CG	45.8	88.1	67.0	28.1±0.6	82.9	17.1	99.4	58.3	28.0±0.6	82.9
	+FT	99.2	88.1	93.7	32.4±0.7	84.4*	98.2	99.2	98.7	32.8±0.7	84.5*

Table 7: Zero-shot formality control results on *new source* languages, using controllers trained on English as source. Sources are {de, es, fr, hi, it}. Colored columns indicate source formality agreeing with desired target formality.

	Model	Feminine	Masculine	Global	BLEU	COMET$_{2022}$
	base	79.4	89.3	85.2	30.0±1.5	83.3
es→it	+CG	87.6	92.3	90.4	29.5±1.4	82.9†
	+FT	90.9	90.5	90.7	30.0±1.3	82.9†
	base	75.4	90.4	84.2	28.1±1.4	82.6
fr→it	+CG	85.1	94.1	90.4	27.7±1.4	82.3
	+FT	90.4	93.6	92.3	28.6±1.4	82.5
	base	83.2	87.0	85.3	31.2±1.4	79.9
es→fr	+CG	86.8	88.8	87.9	31.3±1.4	79.8
	+FT	89.2	88.5	88.8	31.4±1.5	79.7
	base	76.1	87.2	84.3	31.5±1.3	80.4
it→fr	+CG	86.4	89.1	87.9	31.5±1.3	80.5
	+FT	90.6	88.9	89.6	31.8±1.4	80.4

Table 8: Zero-shot grammatical gender control results on *new source and target* languages.

used in training, it also differs in the notion of formality: Korean involves multiple levels of formality instead of a binary informal-formal distinction. For the zero-shot transfer, this means transferring a controller trained for binary control to a multi-class problem with an unknown class mapping, which is naturally more challenging.

Human Evaluation on Bengali: The results are in Table 6. First, adding attribute control does not appear to impact translation quality. Second, pairwise comparisons with the baseline show both CG and finetuning are effective in formality control, where CG has slightly higher win ratio than FT against the baseline. Third, the impact on formality scores is more prominent when steering towards informal translation. This likely because the baseline translations already have a high level of formality. Moreover, the rare usage of the lowest formality level in Bengali (Appendix C) could explain the relatively high formality scores for the systems steered towards "informal" (rows (3) and (5)).

6.2 New Source and Target Languages

New source languages easier than new target languages: In Table 7, we report the results of transferring controllers trained with English source to new source languages. Contrasting these scores with the target-side zero-shot results in Table 4, it is clear that transferring to new source languages is a much easier task. This is expected, as attribute-controlled translation primarily places lexical constraints on the target side. Once the controller can generate translations with the correct attribute, swapping the source language does not pose a large challenge. Even when the source formality disagrees with the desired output formality (uncolored columns in Table 7), the controllers are able to steer the translations toward the required attributes.

NLLB struggles to preserve source attributes: Contrasting the colored "base" cell in Table 7 with its uncolored counterpart, we see that NLLB does have some notion of formality in the source sentences, as source sentences with the correct formality improves accuracy on the desired formality (57.8 vs. 42.2% and 79.7 vs. 20.3%). However, the signals in the input alone are insufficient for generating the correct formality. This is confirmed by another zero-shot experiment when both the source and target languages are new (Table 8). Here the sources already contain the correct grammatical genders. Despite this, NLLB cannot fully utilize the signals in the source, especially on the feminine gender. Its accuracy (76.1-83.2%) still lags behind the masculine class (87.0-90.4%). Both CG and finetuning substantially improve the accuracy and mostly close the gap between the two grammatical classes. This shows both approaches strengthen the source signals that are otherwise neglected.

7 Conclusion

To generalize attribute-controlled translation to data-scarce conditions, we asked the question "how *transferable* are attribute controllers on pretrained multilingual translation model?". We use a novel classifier guidance method to extend a pretrained NLLB-200 model for attribute control and contrast its performance to finetuning-based control.

Our results led to the following recommendations for upgrading existing multilingual translation systems with attribute control capabilities: **1)** Given in-domain target sentences annotated with attributes, even as few as the lower hundreds, finetuning is the primary choice. **2)** In case of *distant* new target languages or strong *domain* mismatches between the attribute-annotated data and test data, decoding with classifier guidance is more promising. Otherwise finetuning is recommended. **3)** In case specific resource constraints preclude finetuning or hosting multiple specialized variants of the underlying model, we then recommend inference-time control by classifier guidance. **4)** In case the underlying translation model is not massively multilingual, finetuning the model or training the controller on multiple target languages is beneficial.

Limitations

More Fine-Grained Attributes Our classifier guidance approach works with discrete labels, making it not directly applicable to use-cases with more fine-grained or continuous attributes. In particular, although the gender classifier training incdlues a gender-neutral class, in evaluation we were only able to test two genders, limited by the availability of test data. As more test datasets with fine-grained attributes become available, our approach can be further improved and validated for these use-cases.

Inference Speed Decoding speed is a main downside of our classifier guidance approach. This is a result of multiple gradient-based updates of model activations at each decoding time step. Despite the promising zero-shot results, further speed-up is necessary is make it realistic for deployed systems.

Acknowledgement

We thank Supriti Sinhamahapatra for help with the human evaluation, and the anonymous reviewers for helpful feedback. This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. Part of this work was supported by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF).

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Parnia Bahar, Patrick Wilken, Javier Iranzo-Sánchez, Mattia Di Gangi, Evgeny Matusov, and Zoltán Tüske. 2023. Speech translation with style: AppTek's submissions to the IWSLT subtitling and formality tracks in 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 251–260, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

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*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Xavier Garcia, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 372–378. INCOMA Ltd.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

David Landsman, Jerry Zikun Chen, and Hussain Zaidi. 2022. BeamR: Beam reweighing with attribute discriminators for controllable text generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 422–437, Online only. Association for Computational Linguistics.

Seungjun Lee, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2023. Improving formality-sensitive machine translation using data-centric approaches and prompt engineering. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 420–432, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *NeurIPS*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2022. Learning an artificial language for knowledge-sharing in multilingual translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 188–202, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.

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 (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Jan Niehues. 2020. Machine translation with unsupervised length-constraints. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 21–35, Virtual. Association for Machine Translation in the Americas.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. 2022. Controlling translation formality using pre-trained multilingual language models. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 327–340, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Ashutosh Saboo and Timo Baumann. 2019. Integration of dubbing constraints into machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 94–101, Florence, Italy. Association for Computational Linguistics.

Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Priyesh Vakharia, Shree Vignesh S, and Pranjali Basmatkar. 2023. Low-resource formality controlled NMT using pre-trained LM. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 321–329, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Sebastian Vincent, Robert Flynn, and Carolina Scarton. 2023. MTCue: Learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8210–8226, Toronto, Canada. Association for Computational Linguistics.

Thuy-trang Vu, Shahram Khadivi, Xuanli He, Dinh Phung, and Gholamreza Haffari. 2022. Can domains be transferred across languages in multi-domain multilingual neural machine translation? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 381–396, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023. Controlling styles in neural machine translation with activation prompt.

In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhanglin Wu, Zongyao Li, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Xiaoyu Chen, Zhiqiang Rao, Zhengzhe Yu, Jinlong Yang, Shaojun Li, Yuhao Xie, Bin Wei, Jiawei Zheng, Ming Zhu, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Improving neural machine translation formality control with domain adaptation and reranking-based transductive learning. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 180–186, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

A Dataset Statistics

A.1 OPUS-100 Data for Transformer-Base

The data overview is in Table 9. For tokenization, we use the SentencePiece (Kudo and Richardson, 2018) model from NLLB-200[11] (NLLB Team et al., 2022). The model is trained to translate from and into English.

Direction	# Sentences	# Tokens (en)	# Tokens (X)
en-es	1,000,000	15,482,094	16,422,413
en-de	1,000,000	17,952,717	20,142,507
en-fr	1,000,000	21,495,343	26,634,530
en-hi	534,319	8,723,899	10,913,496
en-it	1,000,000	14,435,382	15,524,589
en-ko	1,000,000	11,290,102	9,552,148
en-pt	1,000,000	13,879,742	14,410,909
en-ru	1,000,000	16,638,782	19,630,699

Table 9: Overview of OPUS-100 data we used to train the Transformer-base.

A.2 Details on Domain Mismatch Data

For the grammatical gender control experiments with domain mismatch (§4.2), the training domain differs from the test sets in both style and length. An overview is shown in Table 10.

During training, an example tuple of (input, output, attribute label) is: ("the actor finished her work.", "La actriz terminó su trabajo.", 0: feminine) ("the actor finished his work.", "El actor terminó su trabajo.", 1: masculine). The training sentences are all artificial sentences following this simple subject-verb-objective structure. This differs significantly from the test sets with public speaking texts.

Split	Style	Avg. # output words per sent.
Train	artificial sentences	5.5
Test (supervised)	TED talks	25.4
Test (new tgt lang.)	TED talks	25.2
Test (new src & tgt lang.)	TED talks	26.2

Table 10: Details on domain mismatch training setup.

B Training and Inference Details

We implemented our approaches in FAIRSEQ (Ott et al., 2019) at `https://github.com/dannigt/attribute-controller-transfer`.

B.1 Inference

Preprocessing For CoCoA-MT (Nadejde et al., 2022), many test inputs contain multiple sentences. When directly decoding, NLLB-200 (NLLB Team et al., 2022) suffered from severe under-translation,

[11] `https://github.com/facebookresearch/fairseq/tree/nllb/#preparing-datasets-for-training`

where the output translation only contains one sentence. We therefore split the input by sentence boundaries and decode sentence by sentence.

Hyperparameters When decoding, we use a beam size of 4 and length penalty of 1.0.

Evaluation To evaluate BLEU and COMET scores, we concatenate the hypotheses and references from different attributes. It is also the case when reporting the multi-source results in Table 7.

B.2 Details on Finetuning

When finetuning NLLB-200, we use a batch size of $16k$ target tokens. For bilingual systems, we train for 30 updates. When training multilingually, we train for 60 updates. We use a learning rate of 0.0001 with an inverse squared root schedule and 20 warmup steps. Dropout is set to 0.1.

B.3 Details on Classifier Guidance

Attribute Classifier Training The classifier operates on meanpooled decoder hidden states and consists of two feedforward layers with ReLU activation in between. The first layer projects from the 1024 Transformer hidden dimension to 256, the second layer from 256 to C, the number of attribute classes. In our experiments, C is 2 for formality control (formal, informal) and 3 for gender control (feminine, masculine, neutral)[12].

We train the classifier on a frozen NLLB-200 600M model with an effective batch size of $32k$ target tokens. The learning rate is 0.002 with an inverse square root schedule and 20 warm-up steps. We use the Adam (Kingma and Ba, 2015) optimizer with betas of $(0.9, 0.98)$. Dropout and label smoothing are set at 0.1. For formality control, we train the monolingual classifiers for 100 updates and multilingual for 250 updates. For the gender control, we train for 25 updates due to the small dataset and simplicity of the training data.

Hyperparameters For the classifier guidance hyperparameters, on the en→de training data of CoCoA-MT, we searched among step size $[0.05, 0.1, 0.5]$, and number of iterations $[3, 5]$. We used 5 iterations and 0.1 step size for formality control, and 5 iterations and 0.05 step size for grammatical gender control. We do not use KL regularization and postnorm fusion as in Dathathri et al. (2020), since they degraded performance in initial experiments.

Decoding Speed Decoding with our approach is slow due to the repeated gradient updates. For instance on formality control, decoding on the test sets of 600 sentences takes around 30 minutes.

C Details on Human Evaluation

We randomly sampled 16 source English sentences containing second person pronouns from the CoCoA-MT test set, and collected 5 translations for each: from baseline NLLB-200, as well as from CG (`multi`) and FT (`multi`) for both formalities[13]. A native speaker rated the 80 hypotheses.

During the evaluation, we learned that there are three levels of formality in Bengali, where: 1) the lowest formality level is only used between very close relations; 2) the next higher level is used between families or acquaintances; 3) the highest level is used between unfamiliar persons or those between higher social distances. We therefore asked the annotator to match each formality category to one integer point. That is, 1, 2, and 3 correspond to very informal, informal, and formal respectively. We also learned that the lowest formality level is only used between very close relations and therefore rare.

While scoring, the annotator was presented with the English source sentences and their Bengali translations together in random order, and asked to score translation quality on a 5-point scale (1 being the worst) and formality scores on a 3-point scale (1 being the least formal).

D Comparison to Prior Works Trained on Different Data Conditions

Here we compare our results to prior works that used *more relaxed* data conditions than ours for the zero-shot tasks. In Table 11, first four systems are submissions to the unconstrained zero-shot track of the IWSLT 2023 formality control shared task (Agarwal et al., 2023). We compare to submissions in the unconstrained track, as our models would fall under this track due to the use of pretrained models. The scores of other systems are from Table 48 of Agarwal et al. (2023). We grayed out our COMET scores, as we are unsure whether our evaluation used the same underlying model as the organizers

[12] As our test set only covers two genders, we only report scores on two genders.

[13] Due to time constraints, we could not include the combination of finetuning and classifier guidance in the evaluation.

(we used `wmt22-comet-da`). Overall, our model's performance is comparable to the leading systems.

	Formality	BLEU	COMET	M-Acc
en→pt				
Ours	formal	40.3	85.3	100
	informal	43.9	86.0	83
Wu et al. (2023)	formal	45.4	77.4	100
	informal	49.1	78.5	100
Bahar et al. (2023)	formal	34.6	60.9	99
	informal	42.4	67.9	64
Lee et al. (2023)	formal	31.0	52.5	100
	informal	19.9	24.9	68
Vakharia et al. (2023)	formal	26.6	40.5	90
	informal	28.4	42.5	58
en→ru				
Ours	formal	33.2	84.4	100
	informal	33.0	84.4	93
Bahar et al. (2023)	formal	35.4	61.7	99
	informal	33.0	60.3	98
Wu et al. (2023)	formal	33.7	58.0	100
	informal	32.4	55.6	100
Lee et al. (2023)	formal	25.8	44.5	100
	informal	26.3	41.8	100
Vakharia et al. (2023)	formal	18.4	-17.1	99
	informal	14.9	-27.7	52
Vincent et al. (2023)	formal	unknown	unknown	100
	informal	unknown	unknown	99

Table 11: Comparison to prior works with different data conditions.