Low-Resource Formality Controlled NMT Using Pre-trained LM

Priyesh Vakharia and **Shree Vignesh S** and **Pranjali Basmatkar**
Department of Computer Science
University of California, Santa Cruz
{pvakhari, ss64293, pbasmatk}@ucsc.edu

Abstract

This paper describes the UCSC's submission to the shared task on formality control for spoken language translation at IWSLT 2023. For this task, we explored the use of "additive style intervention" using a pre-trained multilingual translation model, namely mBART. Compared to prior approaches where a single style-vector was added to all tokens in the encoder output, we explored an alternative approach in which we learn a unique style-vector for each input token. We believe this approach, which we call "style embedding intervention," is better suited for formality control as it can potentially learn which specific input tokens to modify during decoding. While the proposed approach obtained similar performance to "additive style intervention" for the supervised English-to-Vietnamese task, it performed significantly better for English-to-Korean, in which it achieved an average matched accuracy of 90.6 compared to 85.2 for the baseline. When we constrained the model further to only perform style intervention on the <bos> (beginning of sentence) token, the average matched accuracy improved further to 92.0, indicating that the model could learn to control the formality of the translation output based solely on the embedding of the <bos> token.

1 Introduction

In the past decade, neural machine translation has made remarkable strides, achieving translation quality that is increasingly comparable to human-level performance across various languages. However, despite these advancements, the field of controllable machine translation remains relatively under-explored. One crucial aspect of translation variation is formality, which manifests through grammatical registers, adapting the language to suit specific target audiences. Unfortunately, current neural machine translation (NMT) systems lack the capability to comprehend and adhere to grammatical registers, specifically concerning formality. Consequently, this limitation can result in inaccuracies in selecting the appropriate level of formality, potentially leading to translations that may be deemed inappropriate in specific contexts. Recognizing the significance of formality control, we aim to build a formality-controlled machine translation system to foster smooth and reliable conversations and enhance communication across languages and cultures, facilitating more nuanced and effective linguistic exchanges.

Formality-controlled Neural Machine Translation is the IWSLT 2023 task (Nădejde et al., 2022) under the Formality track. The goal of the task is to achieve formality controlled machine translation for the English-Vietnamese (En-Vi), English-Korean (En-Ko) in a supervised setting and English-Portuguese (En-Pt) and English-Russian (En-Ru) in a zero-shot setting as detailed in (Agarwal et al., 2023). We provide an example of formal and informal translations of an English sentence into Vietnamese in Figure 1. The formal and informal tokens are in bold.

2 Related Works

Machine translation (MT) research has primarily focused on preserving the meaning between languages. However, it is widely recognized that maintaining the intended level of formality in communication is a crucial aspect of the problem (Hovy, 1987) (Hovy, 1987). This field of research was named formality-sensitive machine translation (FSMT) (Niu et al., 2017), where the target formality level is considered in addition to the source segment in determining the translated text. Further, several studies have attempted to regulate formality in MT through side constraints to control politeness, or formality (Sennrich et al., 2016); (Feely et al., 2019); (Schioppa et al., 2021a). Other studies have tried to address this with custom models trained on data with consistent formality (Viswanathan et al., 2020). Most prior research

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 321–329

English: Awesome, and now I just need your billing address, that is associated with the card.
Formal: Tuyệt vời [F]ạ[/F], giờ tôi chỉ cần địa chỉ thanh toán của [F]quý vị[/F], địa chỉ đó được liên kết với thẻ [F]ạ[/F].
Informal: Tuyệt vời, giờ tôi chỉ cần địa chỉ thanh toán của [F]bạn[/F], địa chỉ đó được liên kết với thẻ.

Figure 1: Contrastive Data Sample

has been tailored to individual languages and has labeled large amounts of data using word lists or morphological analyzers.

3 Approach

3.1 Overview

The task of formality-controlled generation can be viewed as a seq2seq machine translation task. More formally, given an input sequence x, we design a model that does the following:

$$\hat{y} = \arg\max_{y \in Y} p(y|x, l_s, l_t, f; \theta) \tag{1}$$

Where,
x is the input sequence,
l_s is the source language,
l_t is the target language,
f is the formality,
$\hat{y}$ is the formality controlled translation

We propose a single model that produces an output, given input *x*, and formality setting *f*. Despite being part of the unconstrained task, our proposed approach does not mine or develop any formality annotated data for training and just uses a pre-trained checkpoint of mBART.

3.2 Design

We looked at previous works incorporating contrasting styles Rippeth et al., 2022, and Schioppa et al., 2021b as motivation for our approach. For controlling styles, the aforementioned works use an additive intervention approach. This approach entails adding a *single style intervention vector V* to the pre-trained encoder output *Z*. The same vector *V* is added to all the tokens of the encoder outputs, thereby changing the encoder outputs uniformly.

We modify the above approach to allow for more flexibility while learning. Instead of a single intervention vector *V*, we propose a unique vector V_i for every token *i* in the input space. In short, we repurpose an Embedding layer as a style intervening layer between the encoder and the decoder. This design resulted from our original question: will allowing more flexibility in the encoder enable it to identify which tokens require stylization, thus making it more interpretable. The hypothesis that originated from this question was: by giving each token its own intervention vector V_i, the model will learn each intervention vector V_i differently based on whether the token at that time step has a contrasting translation that is dependent on the formality setting. In short, we let the model learn different V_i*'s* for each token. If true, this will provide some interpretability on which tokens the model recognizes as having a formality marker and translates them differently in formal and informal settings. This approach is visualized in Figure 2. Since our approach uses an embedding layer for style intervention, we call our approach '*style embedding intervention*.'

We learn the style embedding layer only in the formal setting and use a zero vector in the informal setting. In other words, the style embedding intervention is performed only in the formal setting, and encoder outputs are not perturbed in the informal setting. We do not have separate Embedding layers to learn each formality style, simply because, it would be difficult to switch between layers during batched training. Looking at (Schioppa et al., 2021b), the combination of a style vector and a zero vector for contrasting styles was sufficient to learn the style.

4 Experimental Apparatus

4.1 Dataset

The IWSLT formality shared task provided a formality annotated dataset (Nadejde et al., 2022). This dataset comprises source segments paired with two contrastive reference translations, one for each

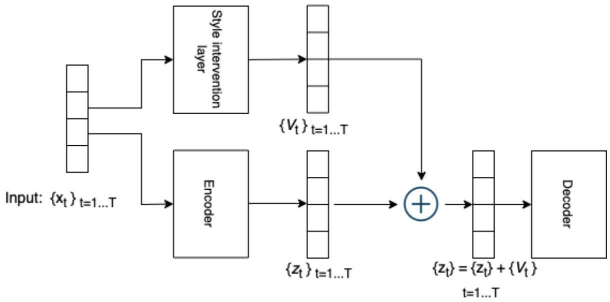

Figure 2: Approach

formality level (informal and formal) for two language pairs: EN-KO, VI in the supervised setting and two language pairs: EN-PT, RU in the zero-shot setting. The data statistics can be seen in Table 1. We use a random split of 0.2 to construct the validation dataset during model development.

4.2 Training Setup

For all our modeling experiments, we use mbart-large-50-one-to-many-mmt, a fine-tuned checkpoint of mBART-large-50 (Liu et al., 2020). This model, introduced by (Tang et al., 2020), is a fine-tuned mBART model which can translate English to 49 languages, including the languages we are interested in: KO, VI, PT, and RU.

For our baseline, we perform zero-shot inference on the mBART model for the four language pairs. The results are shown in tables 3 - 6.

Based on the findings of (Nakkiran et al., 2019) and (Galke and Scherp, 2022) we fixed our loss function to be 'cross entropy with logits' and optimizer to AdamW (Loshchilov and Hutter, 2017). We use the default learning rate of 10^{-3}, standard weight decay of 10^{-2} and set β_1, β_2 and ϵ to 0.9, 0.998 and 10^{-8} respectively.

To effectively train the transformer-based mBART model, we used a learning rate scheduler - a linear schedule with a warm-up, as introduced by (Vaswani et al., 2017). This creates a schedule with a learning rate that decreases linearly from the initial learning rate to 0 after a warm-up period. The warm-up period is set to 10% of the total training steps, during which the learning rate increases linearly from 0 to the initial learning rate set in the optimizer. All the other hyper-parameters are left at their defaults.

We trained our models using one NVIDIA A100 GPU with 80GB memory. To fit our model in this GPU we used a batch size of 16 and a max sequence length of 128. We trained for 15 epochs with an early stopping callback set at 3.

We have implemented all the models in PyTorch (Paszke et al., 2019) leveraging Huggingface (Wolf et al., 2019) transformers and evaluate libraries.

4.3 Evaluation

To assess the performance of the models, we use four metrics to evaluate the two main underlying tasks - translation quality and formality control.

For evaluating the translation quality, we use the following two metrics:

- **Bilingual Understudy Evaluation (BLEU) score**: BLEU score (Papineni et al., 2002) calculates the similarity between a machine translation output and a reference translation using n-gram precision. We use SacreBLEU 2.0 (Post, 2018) implementation for reporting our scores.

- **Cross-lingual Optimized Metric for Evaluation of Translation (COMET) score**: COMET score (Rei et al., 2020) calculates the similarity between a machine translation output and a reference translation using token or sentence embeddings. We use COMET wmt22-comet-da (Rei et al., 2022) model for reporting our scores.

For evaluating the formality control, we use the following two metrics:

- **Matched-Accuracy (M-Acc)**: A reference-based corpus-level automatic metric that leverages phrase-level formality markers from the references to classify a system-generated translation as either formal or informal. This metric was provided by the IWSLT Formality shared task organizers.

- **Reference-free Matched-Accuracy (RF-M-Acc)**: A reference-free variant of M-Acc that uses a multilingual formality classifier, based on xlm-roberta-base, fine-tuned on human-written formal and informal text, to label a system-generated hypothesis as formal or informal. This metric was provided by the IWSLT Formality shared task organizers.

In addition to this, we evaluate our generic translation quality on FLORES-200 (Goyal et al., 2022) for all language pairs under supervised and zero-shot settings. We use the devtest set of FLORES-200 and compute the BLEU and COMET scores.

Language pair	Training Data points	Testing Data points
EN-KO	400	600
EN-VI	400	600
EN-PT	0	600
EN-RU	0	600

Table 1: Data description

	Formal		Informal	
	BLEU	Matched Acc	BLEU	Matched Acc
Rippeth et al., 2022	38.3	98.4	38.3	82.7
Style embedding intervention	38	99.2	37.4	98

Table 2: Grounding our model for EN-ES data

5 Grounding results and observations

Along with the validation splits, we ground our approach by comparing our results with the 2022 formality track submission Rippeth et al., 2022. We compare our results on one language pair i.e. English-Spanish. The comparison is shown in Table 2.

As seen in Table 2, the BLEU scores between our approach - “style embedding intervention” - and the approach in Rippeth et al., 2022 - "additive style intervention" - are similar but our approach makes significant gains in Matched Accuracy, especially in the informal setting indicating improved formality control.

5.1 Style embedding layer analysis

In this section, we analyze the style embedding layer and compare the analysis with the original hypothesis - giving each token its own intervention vector V_i, the model will learn each vector differently based on whether the token at that time step has a contrasting translation that is dependent on the formality setting. Due to the unique nature of our training setup - learning zero vector in the informal setting - for our hypothesis testing, we compare the encoder vectors with and without the style embedding intervention. For this purpose, we use the dot product similarity. At each time step, we compute the dot product similarity between the encoder output before style intervention and the output after style intervention. This is equivalent to comparing the encoder outputs in the formal and the informal setting. The similarity scores are visualized in Figure 3. For a closer look, Table 8 displays the similarity scores.

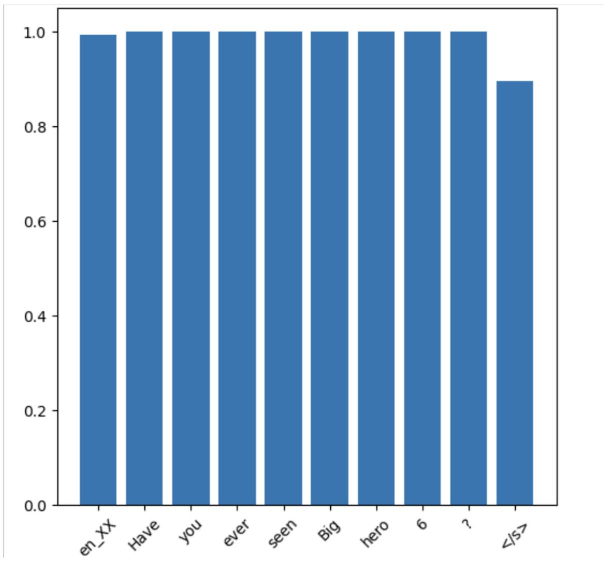

Figure 3: Similarity scores for hypothesis analysis.

As seen from the token representation similarity scores, the model does not seem to learn new information in tokens that have a contrasting setting-dependent translation - the tokens’ similarity scores are very near 1. Instead, it uses the </s>’s representation to store the style ’signal’, by creating a style vector that makes the </s>’s representation $\sim$11% different between formality settings.

Another interesting observation is the extremely slight dissimilarity produced at the beginning of the sentence or ’en_xx’ token. Did the model learn the same style information in $\sim$1% of information space in the ’en_xx’ token compared to the $\sim$11% of information space in the ’</s>’ token? To an-

Models	EN-VI				EN-KO			
	BLEU	COMET	%M-Acc	%C-F	BLEU	COMET	%M-Acc	%C-F
Baseline 1	26.7	0.3629	96	0.95	4.9	0.2110	78	0.99
Baseline 2	26.1	0.829	3	0.006	3.9	0.8445	66.7	0.979
Model 1	44.8	0.8467	99	0.989	22.2	0.8246	74.1	0.9815
Model 2	44.2	0.8702	98.6	0.9782	22.5	0.831	82.9	0.9765
Model 3	44.6	0.874	99	0.9849	23.3	0.836	85.7	0.9832
Model 4	44.3	0.8462	99.2	0.9849	23.2	0.8287	75.3	0.9815

Baseline 1: UMD-baseline
Baseline 2: Zero-Shot mBart
Model 1: single vector intervention with train-dev split of 0.1
Model 2: style embedding intervention
Model 3: bos style intervention - **Primary Submission**
Model 4: single vector intervention with train-dev split of 0.2

Table 3: Results on the official test split in the ***formal supervised*** setting for language pairs ***EN-VI*** and ***EN-KO***.

Models	EN-PT				EN-RU			
	BLEU	COMET	%M-Acc	%C-F	BLEU	COMET	%M-Acc	%C-F
Baseline 1	27.3	0.4477	96.3	0.9766	22.0	0.3492	96.20	0.92
Baseline 2	33	0.8445	54.9	0.8447	24.9	0.7604	99.4	0.9116
Model 1	27.2	0.7686	84.6	0.918	23.8	0.737	97.6	0.865
Model 2	26.6	0.7895	81.5	0.8748	18.5	0.6837	99.2	0.76
Model 3	26.6	0.7889	89.9	0.9082	18.4	0.6664	98.8	0.79
Model 4	28.2	0.7726	80.5	0.9348	24.3	0.7373	97.9	0.858

Baseline 1: UMD-baseline
Baseline 2: Zero-Shot mBart
Model 1: single vector intervention with train-dev split of 0.1
Model 2: style embedding intervention
Model 3: bos style intervention - **Primary Submission**
Model 4: single vector intervention with train-dev split of 0.2

Table 4: Results on the official test split in the ***formal unsupervised*** setting for language pairs ***EN-PT*** and ***EN-RU***.

swer this question, we added another modification to our approach - we masked out the intervention vectors for all tokens except the 'en_xx' token.

For naming purposes, we call this approach '*bos style intervention*' respectively.

6 Official Results

Along with the approach from Rippeth et al., 2022 taken as a baseline and an adapted version of it, we submit the results of our approach and of the '*bos style intervention*' approach. We analyse the performance of our models under the supervised setting and the zero-shot setting. We also generate results on the FLORES-200 test split.

6.1 Supervised Setting

We trained our models multi-lingually on EN-VI and EN-KO for the supervised setting. In the formal setting, we obtain a BLEU score of 44.6 for EN-VI and 23.3 for EN-KO on the official test split. In the informal setting, we obtain a BLEU score of 43.5 for EN-VI and 22.8 for EN-KO. Tables 3 and 5 have detailed results of all our models. Our primary model - '*bos style intervention*' - outperforms the UMD baseline significantly for both languages with around 20 BLEU increase and more than double the COMET score. This answers our hypothesis that the model can learn the formality style in the small $\sim$1% information space at the beginning of the sentence in 'en_xx' token. Moreover, we obtain higher scores on the metrics M-Acc% & C-F% that compute the degree of formality/informality induced.

Qualitative analysis of the translations, especially for KO, revealed that code-switching was a major issue. For example, some translations have

Models	EN-VI				EN-KO			
	BLEU	**COMET**	**%M-Acc**	**%C-F**	**BLEU**	**COMET**	**%M-Acc**	**%C-F**
Baseline 1	25.3	0.3452	96	0.9816	4.9	0.1697	97.6	0.995
Baseline 2	31.9	0.8352	97	0.9933	3.2	0.8311	33.3	0.020
Model 1	43.3	0.8238	98.7	0.9949	22.1	0.8115	96.3	0.889
Model 2	43.6	0.8514	98.9	0.9949	23.0	0.8256	98.3	0.9514
Model 3	43.5	0.8504	98.9	1	22.8	0.8257	98.3	0.9581
Model 4	42.5	0.8232	98.3	0.9765	22.6	0.8162	96.4	0.9028

Baseline 1: UMD-baseline
Baseline 2: Zero-Shot mBart
Model 1: single vector intervention with train-dev split of 0.1
Model 2: style embedding intervention
Model 3: bos style intervention - **Primary Submission**
Model 4: single vector intervention with train-dev split of 0.2

Table 5: Results on the official test split in the ***informal supervised*** setting for language pairs ***EN-VI*** and ***EN-KO***.

Models	EN-PT				EN-RU			
	BLEU	**COMET**	**%M-Acc**	**%C-F**	**BLEU**	**COMET**	**%M-Acc**	**%C-F**
Baseline 1	30.9	0.4161	93.2	0.9082	21.6	0.3475	84.1	0.8417
Baseline 2	33.2	0.8229	45.1	0.1552	18.8	0.7489	0.6	0.0883
Model 1	28.2	0.7606	55.6	0.378	18.8	0.7109	47.7	0.556
Model 2	28.7	0.7821	58.8	0.5092	18.6	0.6544	45.1	0.6
Model 3	28.4	0.7853	58	0.419	14.9	0.6365	51.6	0.6683
Model 4	28.8	0.7673	57	0.3305	20	0.7102	46.9	0.55

Baseline 1: UMD-baseline
Baseline 2: Zero-Shot mBart
Model 1: single vector intervention with train-dev split of 0.1
Model 2: style embedding intervention
Model 3: bos style intervention - **Primary Submission**
Model 4: single vector intervention with train-dev split of 0.2

Table 6: Results on the official test split in the ***informal unsupervised*** setting for language pairs ***EN-PT*** and ***EN-RU***.

Models	EN-VI		EN-KO		EN-PT		EN-RU	
	BLEU	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**
Model 1	29.8	0.8169	5.5	0.773	30.6	0.8082	21.4	0.794
Model 2	27.8	0.8205	4.6	0.758	30.8	0.8258	19.3	0.7686
Model 3	27.9	0.8225	4.5	0.7586	30.4	0.8264	19.1	0.7543
Model 4	30.3	0.8186	5.6	0.7752	30.9	0.814	21.5	0.7935

Model 1: single vector intervention with train-dev split of 0.1
Model 2: style embedding intervention
Model 3: bos style intervention - **Primary Submission**
Model 4: single vector intervention with train-dev split of 0.2

Table 7: Results on ***Flores-200*** test split for language pairs ***EN-VI*** & ***EN-KO*** in supervised setting and for language pairs ***EN-PT*** & ***EN-RU*** in unsupervised setting.

entire phrases or latter parts of sentences in English as shown in Figure 4.

6.2 Zero-shot Setting

We evaluate the above multi-lingually trained model on RU and PT in a zero-shot setting. In the formal setting, we obtain a BLEU score of 26.6 for

Token	Similarity Score
en_xx	0.99037
Have	0.99928
you	0.99914
ever	0.99935
seen	0.99916
Big	0.99916
hero	0.99919
6	0.99920
?	0.99910
</s>	0.89028

Table 8: Similarity scores for hypothesis analysis.

Source(EN) : Okay, I got you. Sorry about that.
Gold Translation(KO) : 네, 이해했어요. 죄송해요.
Predicted translation(KO) : 좋아요, 당신을 잡았어요. Sorry about that.

Figure 4: Similarity scores for hypothesis analysis.

EN-PT and 18.4 for EN-RU on the official test split. In the informal setting, we obtain a BLEU score of 28.4 for EN-PT and 14.9 for EN-RU. Tables 4 and 6 have detailed results of all our models. We observe that our model does not transfer the style knowledge very well. In both cases, the model is often biased toward formal translations. Moreover, our models have a slightly degraded performance in the translation quality than UMD baseline model. This cements our earlier observation that style knowledge transfer is incomplete. Qualitative analysis of the translations revealed that the zero-shot language translations also suffer from code-switching.

6.3 Testing on FLORES-200 dataset

In addition to evaluating formality, we assess the translation quality of our models by evaluating on the FLORES-200 test split. The results can be seen in Table 7.

7 Conclusion

In this paper, we presented and explored "style embedding intervention," a new approach for low-resource formality control in spoken language translation. By assigning unique style vectors to each input token, the proposed approach shows promising results in understanding and controlling the nuances of formal and informal style translation. It outperforms previous "additive style intervention" methods, specifically for the English-to-Korean translation task, resulting in an average matched accuracy improvement from 85.2 to 90.6. Further, on analysis of our "style embedding intervention" model, we find that most of the style information is learnt in the <bos> token. Constraining style addition to the <bos> token - "bos style intervention" - further improved our averaged matched accuracy from 90.6 to 92.

We also observed that in a zero-shot setting, the formality control doesn't seem to transfer well, and the model leans towards biases learnt during pre-training rather than the transferred style interventions. This is more pronounced for En-Ru translations where the model is more biased towards the formal style, with a matched accuracy of 98.8, than the informal style, with a matched accuracy of 51.6.

Future works focused on alleviating the style biases of pre-trained models might be necessary to ensure style transfer works equally well in a zero-shot setting.

We hope our work on translation models with interpretable formality control can serve as a base for other future works on interpretable models, especially in low-resource settings.

Code used for our implementation can be accessed at https://github.com/Priyesh1202/IWSTL-2023-Formality.

8 Acknowledgements

We thank Prof. Lane, Prof. Rao and Brendan King from UC Santa Cruz for their constant guidance and support.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Khalid Choukri, Alexandra Chronopoulou, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Benjamin Hsu, John Judge, Tom Ko, Rishu Kumar, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Matteo Negri, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Elijah Rippeth, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Mingxuan Wang, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

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.

Lukas Galke and Ansgar Scherp. 2022. Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4038–4051, Dublin, Ireland. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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

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.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Maria Nădejde, 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. *arXiv preprint arXiv:2205.04022*.

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.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2019. Deep double descent: Where bigger models and more data hurt. *CoRR*, abs/1912.02292.

Xing Niu, Marianna 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*, pages 2814–2819.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

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, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021a. 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.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021b. 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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Aditi Viswanathan, Varden Wang, and Antonina Kononova. 2020. Controlling formality and style of machine translation output using automl. In *Information Management and Big Data: 6th International Conference, SIMBig 2019, Lima, Peru, August 21–23, 2019, Proceedings 6*, pages 306–313. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.