Towards Modeling the Style of Translators in Neural Machine Translation

Yue Wang* and **Cuong Hoang** and **Marcello Federico**
Amazon AI
hoacuong@amazon.com

Abstract

One key ingredient of neural machine translation is the use of large datasets from different domains and resources (e.g. Europarl, TED talks). These datasets contain documents translated by professional translators using different but consistent translation styles. Despite that, the model is usually trained in a way that neither explicitly captures the variety of translation styles present in the data nor translates new data in different and controllable styles. In this work, we investigate methods to augment the state-of-the-art Transformer model with translator information that is available in part of the training data. We show that our style-augmented translation models are able to capture the style variations of translators and to generate translations with different styles on new data. Indeed, the generated variations differ significantly, up to $+4.5$ BLEU score difference. Despite that, human evaluation confirms that the translations are of the same quality.

1 Introduction

Translators often translate the original content with provided guidelines for styles.[1] However, guidelines are supposed to be high level and not comprehensive. Personal stylistic choices are thus welcome as creative part of the translator's job, as long as their translation style consistency is ensured to the task. By contrast, although neural machine translation (NMT) models (Cho et al., 2014; Sutskever et al., 2014) are trained from these human translations (e.g. Europarl, TED Talks), the models do not explicitly learn to capture the rich variety of translators' styles from the data. This limits their capability to creatively translate new data with different and consistent styles as translators do. We believe that modeling the style of translators is an important yet overlooked aspect in NMT. Our contribution, to the best of our knowledge, is to fill this gap for the first time.

In particular, our work investigates ways to integrate translator information into NMT, with an emphasis on mimicking the translator's style. Our study uses the TED talk dataset, with four language pairs with translator annotations. We present and compare a set of different methods of using a discrete translator token to model and control translator-related stylistic variations in translation. Note that using a discrete token is a common approach to model and control not only specific traits in translation such as verbosity, politeness and speaker-related variances (Sennrich et al., 2016a; Michel and Neubig, 2018)) but also other aspects in NMT such as language ids (Johnson et al., 2017; Fan et al., 2020). However, our study is the first to use such a discrete token to model the style of translators. It also provides several insights regarding translation style modeling as follows.

First, we show that the state-of-the-art Transformer model implicitly learns the style of translators only to a limited extent. Moreover, methods that add translator information to the decoder surprisingly result in NMT that fully *ignores* the additional knowledge. This is regardless of whether the token is added to the bottom (i.e. the embedding layer) or to the top (i.e. the softmax layer) of the decoder. Meanwhile, methods that add the information to the encoder seem to model the translator's style effectively.

Second, we show that our best style-augmented NMT method is able to control the generation of translation in a way that mimics the translator's style, e.g. lexical and grammatical preferences, verbosity. While output produced by the style-augmented NMT can vary significantly with the translator-token values, with BLEU score variations up to $+4.5$, a human evaluation confirms that observed differences are all about style and not

*Y. Wang carried out this work during an internship with Amazon AI.

[1] See https://www.ted.com/participate/translate/guidelines as an example of translation style guidelines.

translation quality. Finally, we show that the translator information has more impact on NMT than the speaker information, which was investigated by Michel and Neubig (2018).

2 Related Work

Style itself is a broad concept (Kang and Hovy, 2019). It includes both simple high-level stylistic aspects of language such as verbosity (Marchisio et al., 2019; Agrawal and Carpuat, 2019; Lakew et al., 2019), formality (Niu et al., 2017; Xu et al., 2019), politeness (Mirkin et al., 2015) and complex aspects such as demography (Vanmassenhove et al., 2018; Moryossef et al., 2019; Hovy et al., 2020) and personal traits (Mirkin and Meunier, 2015; Rabinovich et al., 2017; Michel and Neubig, 2018).

Our study focuses on capturing the personal style of translators. The closest work to our study is thus the work of Michel and Neubig (2018), where they study instead the effects of using the speaker information in NMT. In our results, we show that the translator information has indeed more impact to NMT than the speaker information.

Finally, another distantly related research line tries to improve the diversity in the top rank translations of an input (Li et al., 2016; Shen et al., 2019; Agrawal and Carpuat, 2020). In fact, adding the translator information to NMT also provides means to generate translations with significantly different stylistic variations.

3 NMT with Translator Information

NMT reads an input sequence $\mathbf{x} = x_1, ..., x_n$ in the source language with an encoder and then produces an output sequence $\mathbf{y} = y_1, ..., y_m$ in the target language. The generation process is performed in a token-by-token manner and its probability can be factored as $\prod_{j=1}^{m} P(y_j \mid \mathbf{y}_{<j}, \mathbf{x})$, where $\mathbf{y}_{<j}$ denotes the previous sub-sequence before j-th token. The prediction for each token over the vocabulary $\mathcal{V}$ is based on a softmax function as follows:

$$P(y_j|\mathbf{y}_{<j}, \mathbf{x}) = \text{softmax}(\mathbf{W}_V \mathbf{o}_j + \mathbf{b}_V). \quad (1)$$

Here, $\mathbf{o}_j \in R^d$ is an output vector with size d (e.g. 512 or 1024), encoding both the context from the encoder and the state of the decoder at time j. Meanwhile, $\mathbf{W}_V \in R^{|\mathcal{V}| \times d}$ and $\mathbf{b}_V \in R^{|\mathcal{V}|}$ are a trainable projection matrix and bias vector.

We adjust NMT in different ways as below to let it mimic and control the translator's style.

Source Token. In our first approach, we insert the translator token T as the beginning of each input sentence. The translator token is thus assigned with an embedding vector like any other source token. Hence, the embedding sequence E_{enc} for the MT encoder becomes:

$$E_{enc} = [e(T), e(x_1), ..., e(x_n)], \quad (2)$$

where $e(\cdot)$ is an embedding lookup function.

Token Embedding. We also consider adding the embedded translator token $e(T)$ to every token embedding in the encoder and/or decoder as follows:

$$E_{enc} = [e(T) + e(x_1), ..., e(T) + e(x_n)], \quad (3)$$
$$E_{dec} = [e(T) + e(y_1), ..., e(T) + e(y_m)]. \quad (4)$$

Our motivation is to reinforce the influence of the translator token in MT.

Output Bias. Following Michel and Neubig (2018), we add the translator token information to the output bias at the final layer of the decoder (FULL-BIAS variant). Specifically, the method directly modulates the word probability over vocabulary $\mathcal{V}$ as follows:

$$P(y_j|\mathbf{y}_{<j}, \mathbf{x}, T) = \text{softmax}(\mathbf{W}_V \mathbf{o}_j + \mathbf{b}_V + \mathbf{b}_T). \quad (5)$$

Here, $\mathbf{b}_T \in R^{|\mathcal{V}|}$ is the translator-specific bias vector, which can be thought of as a translator-token embedding with dimension $|\mathcal{V}|$ rather than d. We also explore another variant, named FACT-BIAS, as in Michel and Neubig (2018). This variant instead learns the translator bias through the factorization:

$$\mathbf{b}_T = \mathbf{W}\mathbf{s}_T, \quad (6)$$

with parameters $\mathbf{W} \in R^{|\mathcal{V}| \times k}$ and $s_T \in R^{k \times 1}$ where $k << |\mathcal{V}|$.

Note that while the above methods digest the translator token at an earlier stage, this one consumes translator signals in a late fusion manner.

4 Experiments

4.1 Dataset and Models

We run experiments with the WIT[3] public dataset of TED talks (Cettolo et al., 2012), with four language pairs: English-German (en-de), English-French (en-fr), English-Italian (en-it) and English-Spanish (en-es). The dataset contains both speaker and translator information for each talk and translation, thus allowing to measure the effects of translators and speakers.

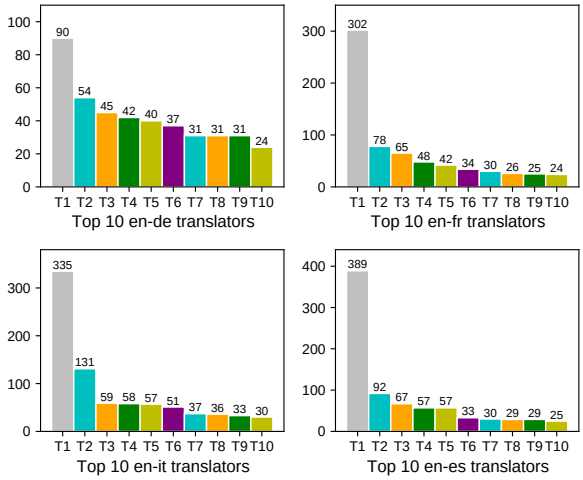

Figure 1: TED talks translated by top 10 translators.

	en-de	en-fr	en-it	en-es
#talks	425	674	827	808
avg sent/talk	107.44	118.78	118.86	115.10
std dev	64.75	60.06	59.95	56.23
#train	36,594	67,554	83,968	79,200
#val	4,066	7,506	9,329	8,800
#test	5,000	5,000	5,000	5,000

Table 1: Data statistics for four language pairs.

We construct training, validation and test sets for each translation direction as follows. We first extract all talks that are translated by the 10 most popular translators (see Figure 1) and split them into parallel sentences. From the data of each translator, we then sample 500 sentences for testing, and, from the remaining data, 90% for training and 10% for validation. All training, testing, and validation sentence pairs are put together and annotated with training and speaker labels. Table 1 shows the data statistics for four language pairs.

For preprocessing, we employ Moses (Koehn et al., 2007) tool[2] for tokenization and apply subword-nmt[3] (Sennrich et al., 2016b) to learn subword representations.

We choose Transformer (Vaswani et al., 2017) as the baseline and employ Fairseq (Ott et al., 2019) for our implementations. Our Transformer model is comprised of 6 layers of encoder-decoder network, where each layer contains 16 heads with a self-attention hidden state of size 1024 and a feed-forward hidden state of size 4096. We employ Adam optimizer (Kingma and Ba, 2015) to update model parameters. We warm up the model by linearly increasing the learning rate from 1×10^{-7} to 5×10^{-4} for 4000 updates and then decay it with an inverse square root of the rest training steps by a rate of 1×10^{-4}. We apply a Dropout of 0.3 for en-de and 0.1 for both en-fr and en-it.

For all MT systems, we load weights from pretrained models to set up a better model initialization. Specifically, we employ models pretrained on WMT data for en-de and en-fr (Ott et al., 2018), and pretrain models for en-it and en-es using our large in-house out-of-domain data, as there are no previous pretrained models for these pairs. We finetune models on TED talk data for 10 epochs[4] and select the best model based on the validation loss.

During inference, we employ beam search with a beam size of 4 and add a length penalty of 0.4.

We use the BLEU score (Papineni et al., 2002) to evaluate translation accuracy.

4.2 Results

4.2.1 Adding Translator Token

We first compare methods to integrate the translator token into the Transformer. Notice that we report performance of the model in two settings: (i) when fed with the oracle translator label (as at training time) and (ii): when fed with randomly assigned labels. Intuitively, if a model really leverages the translator information, we expect to see a performance drop in the randomized setting. Results are shown in Table 2.

Our findings are as follows. First, it is surprisingly ineffective to add the translator token into the decoder, whether to the input (DEC-EMB) or to the softmax (FULL-BIAS, FACT-BIAS). In most cases, our randomization experiment shows that the model simply ignores the information.

Second, methods adding the token to the encoder (SRC-TOK, ENC-EMB) are significantly more effective. Translation accuracy is also consistently better (at most by 0.4 BLEU) than with the Transformer baseline, indicating the translator token is useful. For those models, randomizing translator labels results in visible drops in BLEU score (up to 1.0 BLEU), indicating that the translator information has an important effect to the model.

[2] https://github.com/moses-smt/mosesdecoder

[3] https://github.com/rsennrich/subword-nmt

[4] We try finetuning with more epochs and observe no further improvements.

Model	en-de	en-fr	en-it	en-es
BASE	32.70_{11}	48.20_{14}	42.59_{08}	50.02_{03}
SRC-TOK	32.73_{09}	$\mathbf{48.59}_{06}$	$\mathbf{42.86}_{11}$	50.20_{18}
Rand (Δ)	-0.12	-1.01	-0.32	-0.21
ENC-EMB	$\mathbf{32.86}_{09}$	48.41_{21}	42.79_{04}	$\mathbf{50.25}_{13}$
Rand (Δ)	-0.33	-0.96	-0.43	-0.64
DEC-EMB	32.71_{13}	48.16_{12}	42.53_{07}	49.92_{14}
Rand (Δ)	-0.02	+0.01	0	+0.10
FULL-BIAS	32.65_{08}	48.18_{12}	42.61_{09}	49.97_{07}
Rand (Δ)	-0.02	0	+0.03	-0.01
FACT-BIAS	32.63_{03}	48.23_{10}	42.64_{02}	50.02_{07}
Rand (Δ)	+0.07	-0.02	-0.02	+0.01

Table 2: Average BLEU scores from 3 random seeds. Subscripts denote the standard deviation (e.g., 32.70_{11} $\Rightarrow$ 32.70 ± 0.11). Best results for each column are in bold. "Rand (Δ)" denotes the absolute performance change after randomizing translator tokens.

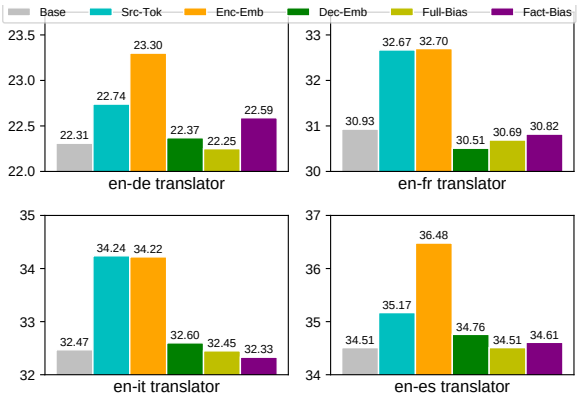

Figure 2: Translator classification accuracy. ENC-EMB yields the best result in most cases.

4.2.2 Style Imitation

Following the common practice in evaluating the style imitation (e.g. see (Michel and Neubig, 2018; Hovy et al., 2020)), we train a classifier to predict the translator style of the output of various models. We employ a Logistic Regression classifier based on both uni-gram and bi-gram word features. The classifier, trained on NMT training data, is applied on the outputs of NMT models. Figure 2 shows the results of this experiment.

As can be seen, the standard Transformer learns the style of translators only to a limited extent. The style of translation outputs are less consistent with the original translator's style, i.e. accuracy is between 20% and 35%). Meanwhile, the classification accuracy is significantly higher (up to +12% relative) under SRC-TOK and ENC-EMB. This confirms that explicitly incorporating translator information at the sentence level allows for transferring some of her/his personal traits into the translations.

Meanwhile, we notice higher accuracy achieved with the reference translations (e.g. 42% in EN-ES), suggesting there is room for improvement.

4.2.3 Stylistic Variations

We analyzed stylistic variations using different translator token labels. In particular, we evaluate model outputs on *en-fr* after translating the entire test set with the same translator token labels. As in Table 3, translator-informed NMT can produce quite different outputs, resulting in BLEU score variations up to +4.5, (i.e. between $T7$ and $T3$, $T8$, $T10$). We also observe differences in BLEU (albeit smaller) when testing with the WMT 2014 test set. In particular, BLEU score variations are up to $+0.84$ between $T7$ and $T5$. We also compute the symmetric-BLEU distances between any two of the translators using their predictions for both TED and WMT test set and visualize their heatmaps in Figure 3. We observe that a similar BLEU distance between various translators in both test sets. Besides, T7 has a farther distance with others but its gap is closer on WMT than TED. These findings verify the consistency of translator styles in data from different domains.

Then, we asked 3 professional translators to grade the quality of translation produced with the labels $T7$ and $T3$ on the TED talks. The evaluation is on a 1-6 scale (higher is better) on a random sample of 100 sentences. This resulted in average scores of 4.867 and 4.860 for T3 and T7, respectively. A similar human evaluation with $T7$ and $T5$ labels was also run on a random sample of 100 sentences of the WMT 2014 test set. It provided the same conclusion: average scores are very similar: 4.99 and 5.0 for T5 and T7 respectively. Both evaluations confirm that there is no difference in translation quality when using different token labels, i.e. the low BLEU score of T7 is only an effect due to stylistic differences.

Table 4 shows examples of translations generated with labels T3 and T7. As we can observe, the translations show different use of grammars, words and verbosity.[5]

[5]Note that one could argue that it is not just about style here but also translation fidelity. We thank a reviewer for pointing it out.

Dataset	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
TED	48.02	46.85	48.07	47.82	47.49	47.50	43.50	48.01	48.07	48.12
WMT	42.19	42.34	42.08	42.32	42.46	42.34	41.62	42.27	41.77	42.35

Table 3: BLEU scores when translating the test set with a specific translator id.

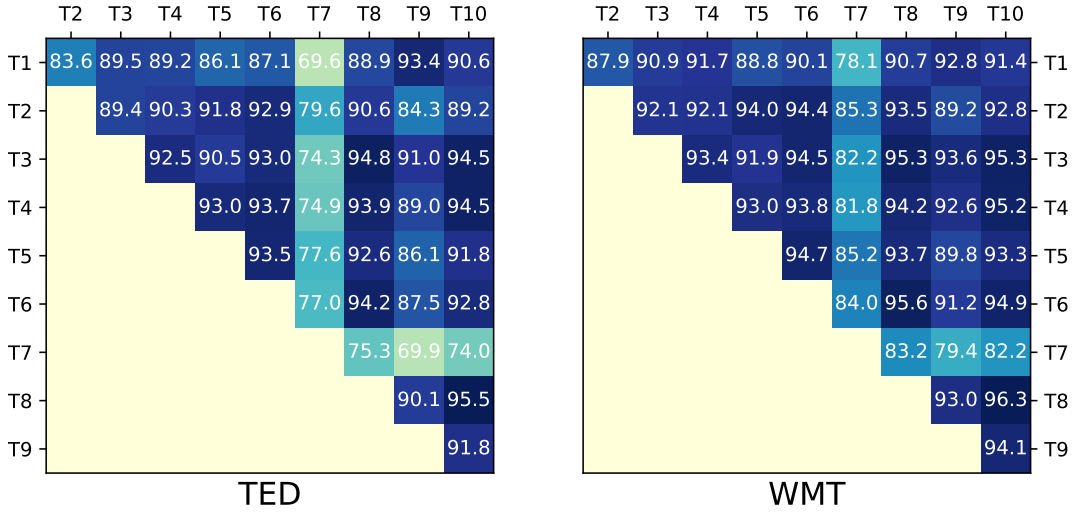

Figure 3: Heatmap visualization for symmetric-BLEU distances between translators.

Verbosity	**Src:** And I'm not the first person to ask this question. **T3:** Je ne suis pas la première personne à poser cette question. **T7:** Je ne suis pas la première à poser cette question.
	Src: And then everybody kind of runs out and goes out. **T3:** Et puis tout le monde s'enfuit.. **T7:** Tout le monde s'enfuit.
Word	**Src:** Same story for fairness. **T3:** Même histoire pour l'équité. **T7:** Même histoire d'équité.
Grammar	**Src:** I had just tweeted, "Pray for Egypt". **T3:** J'avais tweeté : "Priez pour l'Egypte". **T7:** Je venais de tweeter, "Priez pour l'Egypte."

Table 4: Examples of stylistic differences: T3 and T7 have different preferences of grammars and words in translation. Their translations are also different in the verbosity (Using T7 results in consistently less verbose output than as of using T3), which is indeed also what translations by T3 and T7 differ in the training data.

4.2.4 Translator vs. Speaker Effects

Finally, we compared the effect of the translator token with that of the speaker token, which was proposed in Michel and Neubig (2018) to perform extreme personalization. Results on all four directions (see Table 5) show that the translator token has more impact. [6] Given that speaker and author style has received much more attention in the literature, we hope that this final result will spark more interests on the style of translators.

[6]One probable reason is that the speaker signal is more sparse than the translator signal, i.e. each speaker is represented by one TED talk, while translators by multiple talks.

Model	en-de	en-fr	en-it	en-es
BASE	32.70_{11}	48.20_{14}	42.59_{08}	50.02_{03}
ENC-EMB Speaker	32.80_{13}	48.18_{10}	42.23_{09}	49.25_{08}
ENC-EMB Translator	$\mathbf{32.86}_{09}$	$\mathbf{48.41}_{21}$	$\mathbf{42.79}_{04}$	$\mathbf{50.25}_{13}$

Table 5: Comparison between ENC-EMB on Translator and Speaker sides. Results are similar for SRC-TOK.

5 Conclusion

We designed various ways of incorporating translator information into NMT, in order to model and control the generation of translation with different translator styles. We show that resulting style-augmented NMT produces significantly different stylistic variations, mimicking professional translators. Human evaluation confirms that the generated variations are all of same translation quality.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in 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 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Sweta Agrawal and Marine Carpuat. 2020. Generating diverse translations via weighted fine-tuning and hypotheses filtering for the Duolingo STAPLE task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 178–187, Online. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

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, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Dongyeop Kang and Eduard Hovy. 2019. xSLUE: A Benchmark and Analysis Platform for Cross-Style Language Understanding and Evaluation. *arXiv:1911.03663 [cs]*. ArXiv: 1911.03663.

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*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Antonino Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. *CoRR*, abs/1910.10408.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

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 Volume 1: 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.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1102–1108. The Association for Computational Linguistics.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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 NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. 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 5719–5728. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations.