Improving Neural Machine Translation Formality Control with Domain Adaptation and Reranking-based Transductive Learning

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, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,lizongyao,weidaimeng,shanghengchao,guojiaxin1,chenxiaoyu35, raozhiqiang,yuzhengzhe,yangjinlong7,lishaojun18,xieyuhao2,weibin29, zhengjiawei15,zhuming47,leilizhi,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents Huawei Translation Service Center (HW-TSC)'s submission on the IWSLT 2023 formality control task, which provides two training scenarios: supervised and zero-shot, each containing two language pairs, and sets constrained and unconstrained conditions. We train the formality control models for these four language pairs under these two conditions respectively, and submit the corresponding translation results. Our efforts are divided into two fronts: enhancing general translation quality and improving formality control capability. According to the different requirements of the formality control task, we use a multi-stage pre-training method to train a bilingual or multilingual neural machine translation (NMT) model as the basic model, which can improve the general translation quality of the base model to a relatively high level. Then, under the premise of affecting the general translation quality of the basic model as little as possible, we adopt domain adaptation and reranking-based transductive learning methods to improve the formality control capability of the model.

1 Introduction

Machine translation (MT) (Lopez, 2008; Vaswani et al., 2017) models typically return one single translation for each input sentence. This means that when the input sentence is ambiguous, the MT model must choose a translation from among various valid options, without regard to the intended use case or target audience. Therefore, there is a need to control certain attributes (Schioppa et al., 2021) of the text generated in a target language such as politeness (Sennrich et al., 2016a; Feely et al., 2019) or formality (Niu et al., 2017, 2018; Viswanathan et al., 2020).

The lack of gold translation with alternate formality for supervised training and evaluation has lead researchers to rely on synthetic supervision training and manual evaluation in past work (Niu and Carpuat, 2020). Fortunately, the IWSLT formality control task now provides a new benchmark[1] (Nădejde et al., 2022; Agarwal et al., 2023) by contributing high-quality training datasets and test datasets for multiple language pairs.

This paper presents HW-TSC's submission on the IWSLT 2023 formality control task. How formality distinctions are expressed grammatically and lexically can vary widely by language. Thus, we participate in the formality control task of all these four language pairs to investigate a general formality control method that can be applied to different language pair. In addition, we also investigate the difference in formality control between constrained and unconstrained conditions by introducing the mBART model (Liu et al., 2020) under unconstrained condition.

2 Data

2.1 Pre-training Data

We use the CCMatrix[2] and OpenSubtitles[3] bilingual data given by the organizers to train a NMT model from scratch or fine-tune the mBART model as the general basic model. The bilingual data size of each language pair is shown in Table 1:

Language pair	CCMatrix	OpenSubtitles
EN-KO	19.4M	1.4M
EN-VI	50.1M	3.5M
EN-PT	173.7M	33.2M
EN-RU	139.9M	25.9M

Table 1: The bilingual data size of each language pair.

In order to achieve a better training effect, we also use some data pre-processing methods to clean bilingual data, such as: remove duplicate data, use

[1] https://github.com/amazon-science/contrastive-controlled-mt
[2] https://opus.nlpl.eu/CCMatrix.php
[3] https://opus.nlpl.eu/OpenSubtitles-v2018.php

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 180–186
July 13-14, 2023

Moses[4] to normalize punctuation, filter extremely long sentences, use langid[5] (Lui and Baldwin, 2011, 2012) to filter sentences that do not meet the language requirements, use fast-align[6] (Dyer et al., 2013) to filter unaligned sentence pairs.

2.2 Formality-annotated Data

The formality-annotated data is provided by the organizers, and the data size of each language pair is shown in Table 2:

Setting	Language pair	Train	Test
Supervised	EN-KO	400	597
Supervised	EN-VI	400	598
Zero-shot	EN-PT	0	599
Zero-shot	EN-RU	0	600

Table 2: The formality-annotated data size of each language pair.

For supervised language pairs, we split the formality-annotated train data into a train set and a dev set with a ratio of 3:1, and use the formality-annotated train set and a small amount of bilingual data for formality control training, while for zero-shot language pairs, we use formality-annotated train set from the other two supervised language pairs for formality control training.

3 Model

3.1 Constrained Model

Transformer (Vaswani et al., 2017) is the state-of-the-art model in recent machine translation evaluations. There are two parts of research to improve this kind: the first part uses wide networks (eg: Transformer-Big (Vaswani et al., 2017)), and the other part uses deeper language representations (eg: Deep Transformer (Wang et al., 2019; Wu et al., 2022; Wei et al., 2022)). Under the constrained conditions, we combine these two improvements, adopt the Deep Transformer-Big model structure, and train a one-to-many multilingual NMT model (Johnson et al., 2017; Zhang et al., 2020) from scratch using bilingual data of four language pairs provided by the organizers. The main structure of Deep Transformer-Big is that it features pre-layer-normalization and 25-layer encoder, 6-layer decoder, 16-head self-attention, 1024-dimensional embedding and 4096-dimensional FFN embedding.

3.2 Unconstrained Model

Recently, multilingual denoising pre-training method (Liu et al., 2020; Tang et al., 2021) produces significant performance gains across a wide variety of machine translation tasks. As the earliest sequence-to-sequence model using multilingual denoising pre-training method, mBART (Liu et al., 2020) has also achieved good results in various machine translation-related tasks. Under unconstrained conditions, we use the mBART50 1n model[7] as the initial model of the unconstrained formality control task. The mBART50 1n model adopts Transformer structure, which features 12-layer encoder, 12-layer decoder, 16-head self-attention, 1024-dimensional embedding and 4096-dimensional FFN embedding, and an additional layer-normalization layer (Xu et al., 2019) on top of both the encoder and decoder.

4 Method

In our implementation, we first use a multi-stage pre-training method to train a general NMT model with relatively high translation quality. Then, we use domain adaptation method to fine-tune the NMT model so that the model can have basic formality control capability. Finally, we use the reranking-based transductive learning (RTL) method to further improve the formality control capability of the model.

4.1 Multi-stage Pre-training

There are four different types of formality control tasks, which are constrained supervised task, constrained zero-shot task, unconstrained supervised task, and unconstrained zero-shot task. For these four different tasks, we formulate different pre-training strategies and collectively refer to these strategies as multi-stage pre-training method.

Under the constrained condition, we adopt the Deep Transformer-Big model structure and use bilingual data of all four language pairs to train a one-to-many multilingual NMT model from scratch, which is used as the basic model for constrained zero-shot task. For constrained supervised task, we use the bilingual data of this task to further

[4] https://github.com/moses-smt/mosesdecoder
[5] https://github.com/saffsd/langid.py
[6] https://github.com/clab/fast_align

[7] https://dl.fbaipublicfiles.com/fairseq/models/mbart50/mbart50.ft.1n.tar.gz

pre-train the multilingual NMT model to obtain a bilingual NMT model as the basic model.

While under the unconstrained condition, we further pre-train the mBART50 1n model using bilingual data from all these four language pairs as the basic model for unconstrained zero-shot task. For unconstrained supervised task, we use the bilingual data of this task to further pre-train the pre-trained model, and use the final pre-trained bilingual model as the basic model.

4.2 Domain Adaptation for Formality Control

With the pre-trained basic model, we use domain adaptation method (Chu et al., 2017) to achieve basic formality control. First, we treat formal formality and informal formality as two special domains, and control the formality of the model's translation results using a tagging method (Chu et al., 2017; Nădejde et al., 2022), which attaches a formality-indicating tag to the source input. Then, in order to affect the general translation quality as little as possible, we use a mix fine-tuning method (Chu et al., 2017; Nădejde et al., 2022). Our specific implementation is to upsample the formality-annotated train set by 5 times, and mix it with the same amount of randomly sampled general bilingual data to fine-tune the pre-trained basic model.

As mentioned in Section 2.2, for the zero-shot task, due to the lack of formality-annotated data, we have to use the formality-annotated data of the two other supervised language pair, which is why we set the basic model of zero-shot task to a multilingual NMT model. After using domain adaptation method, the cross-lingual transfer learning capability of multilingual model can help zero-shot language pair achieve basic formality control.

4.3 Reranking-based Transductive Learning

After using domain adaptation method, we can enable the model to have the basic formality control capability. Inspired by the idea of transductive learning (Shi et al., 2018; Lee et al., 2021), we propose a RTL method, which can further improve the formality control capability of NMT model. Our method is mainly divided into two steps:

In the first step, we adopt beam search based decoding method (Sennrich et al., 2016b) for the formality control model, and then select the final translation result that meets the specified formality requirements from the top100 decoding results based on reranking idea (Dou et al., 2019). For supervised task, we use a reference-free formality classifier and the formality phrases from formality-annotated training data for reranking. The implementation details are shown in Algorithm 1. For zero-shot task, due to the lack of formality-annotated training data, we just use a reference-free formality classifier for reranking. Among them, the formality classifier under the constrained condition comes from self-training (Axelrod et al., 2011), while the formality classifier under the unconstrained condition comes from the organizer[8] (Briakou et al., 2021).

Algorithm 1: Reranking by reference-free formality classifier and formality phrases

Input: source sentence x, reference-free formality classifier C, formality control model M, formal and informal formality phrases $W_F = \{w_j^F\}_{j=1}^{|W_F|}$, $W_I = \{w_j^I\}_{j=1}^{|W_I|}$

Output: the formality translation y_F and y_I

translate x by M, the top 100 formality translations are respectively defined as: $D_F = \{y_i^F\}_{i=1}^{100}$, $D_I = \{y_i^I\}_{i=1}^{100}$

$y_F = y_0^F$

for y_i^F *in* D_F **do**

    $F_{flag} = False$

    for w_j^F *in* W_F **do**

        if w_j^F *in* y_i^F **then**

            $F_{flag} = True$

            break

        end

    end

    calculate the formality by C: C(y_i^F)

    if F_{flag} *and C(y_i^F)=="formal"* **then**

        $y_F = y_i^F$

        break

    end

end

pick y_I from D_I in a similar way to y_F

return y_F, y_I

In the second step, we add the source text of test set and the reranked formality translation results to the training data used for domain adaptation, and then use the adjusted training data to further fine-tune the formality control model.

We can also repeat the previous two steps until the formality control capability of the model on test set is no longer improved. We refer to this iterative

[8] https://github.com/amazon-science/contrastive-controlled-mt/releases/tag/classifier-v1.0.0

EN-VI	To Formal				To Informal				Flores	
	M-Acc	C-F	BLEU	COMET	M-Acc	C-F	BLEU	COMET	BLEU	COMET
AWS-baseline	99.40%	99.16%	43.2	0.6189	98.10%	98.49%	41.5	0.6021	-	-
Multilingual pre-training	10.86%	1.67%	25.6	0.2023	89.14%	98.33%	30.0	0.2873	42.3	0.6653
+ Bilingual pre-training	8.80%	3.01%	24.8	0.1782	91.20%	96.99%	28.9	0.2630	42.4	0.6706
+ Domain adaptation	98.17%	97.83%	49.1	0.7248	99.37%	99.83%	48.0	0.6952	41.3	0.6576
+ RTL	99.59%	**100.00%**	49.5	0.7296	99.38%	**100.00%**	48.1	0.7034	41.7	0.6614
+ Iterative RTL	**100.00%**	99.83%	**51.3**	**0.7522**	**100.00%**	**100.00%**	**49.8**	**0.7209**	**41.8**	**0.6730**
UMD-baseline	96.00%	99.67%	26.7	0.3629	96.00%	98.16%	25.3	0.3452	-	-
mBART50 1n	3.82%	1.51%	26.7	0.3516	96.18%	98.49%	31.0	0.4426	34.7	0.6040
+ Multilingual pre-training	9.44%	1.84%	25.4	0.2089	90.56%	98.16%	29.9	0.2975	42.2	0.6673
+ Bilingual pre-training	12.20%	2.51%	25.2	0.1579	87.80%	97.49%	29.4	0.2445	42.4	0.6698
+ Domain adaptation	99.02%	99.50%	47.8	0.7181	99.36%	**100.00%**	47.4	0.6930	43.2	0.6916
+ RTL	99.22%	**100.00%**	47.7	0.7190	99.16%	**100.00%**	47.8	0.7053	**43.4**	**0.7033**
+ Iterative RTL	**100.00%**	**100.00%**	**48.2**	**0.7214**	**100.00%**	**100.00%**	**48.3**	**0.7102**	**43.4**	0.6983

Table 3: The overall translation quality and formality control accuracy of EN-VI models.

EN-KO	To Formal				To Informal				Flores	
	M-Acc	C-F	BLEU	COMET	M-Acc	C-F	BLEU	COMET	BLEU	COMET
AWS-baseline	28.50%	54.61%	11.1	0.5044	80.40%	57.62%	11.1	0.5125	-	-
Multilingual pre-training	**100.00%**	69.85%	5.0	0.2408	0.00%	30.15%	4.5	0.2288	12.9	0.6497
+ Bilingual pre-training	**100.00%**	65.33%	5.5	0.2189	0.00%	34.67%	4.7	0.2105	13.8	0.6610
+ Domain adaptation	**100.00%**	97.49%	24.5	0.7234	**100.00%**	96.31%	25.1	0.7194	12.6	0.6528
+ RTL	**100.00%**	97.65%	**25.8**	0.7337	**100.00%**	98.51%	26.5	0.7337	13.0	**0.6828**
+ Iterative RTL	**100.00%**	**99.83%**	25.0	**0.7434**	**100.00%**	**99.66%**	**27.0**	**0.7495**	**13.2**	0.6729
UMD-baseline	78.30%	98.60%	4.9	0.2110	97.60%	99.50%	4.9	0.1697	-	-
mBART50 1n	**100.00%**	98.49%	4.1	0.4468	0.00%	1.51%	3.2	0.3670	9.5	0.5854
+ Multilingual pre-training	**100.00%**	65.66%	5.0	0.2501	0.00%	34.34%	4.3	0.2338	13.3	0.6605
+ Bilingual pre-training	**100.00%**	64.66%	5.2	0.2240	0.00%	35.34%	4.6	0.2114	14.2	0.6734
+ Domain adaptation	**100.00%**	99.33%	24.9	0.7297	**100.00%**	99.66%	25.5	0.7379	12.8	0.6666
+ RTL	**100.00%**	99.66%	**25.5**	**0.7393**	**100.00%**	**100.00%**	26.2	**0.7340**	13.8	0.6845
+ Iterative RTL	**100.00%**	**100.00%**	24.2	0.7254	**100.00%**	**100.00%**	**26.7**	0.7311	**14.0**	**0.6882**

Table 4: The overall translation quality and formality control accuracy of EN-KO models.

process as iterative RTL method.

5 Experiments

5.1 Training Details

We use the Pytorch-based Fairseq framework[9] (Ott et al., 2019) to pre-train or fine-tune NMT model, and use Adam optimizer (Kingma and Ba, 2014) with parameters $\beta 1$=0.9 and $\beta 2$=0.98. During the multi-stage pre-training phase, each model uses 8 GPUs for training, warmup steps is 4000, batch size is 4096, learning rate is 5×10^{-4}, label smoothing rate (Szegedy et al., 2016) is 0.1, and dropout is 0.1. In the domain adaptation and RTL phases, each model only uses 1 GPU for training without warm-up, batch size is 1024, learning rate is 3×10^{-5}, label smoothing rate is 0.1, and dropout is 0.3.

5.2 Evaluation Metrics

We evaluate the translation results of formality control model from the following two dimensions:

- We use SacreBLEU v2.0.0 [10] (Papineni et al., 2002; Post, 2018) and COMET (eamt22-cometinho-da)[11] (Rei et al., 2022) to evaluate the overall translation quality of formality control model on the official formality test sets and FLORES-200 devtest sets[12] (Goyal et al., 2022).
- We also use the reference-based corpus-level automatic metric Matched-Accuracy (M-Acc) and the reference-free automatic metric (C-F) that uses a multilingual formality classifier provided by the organizer to evaluate the formality control accuracy of the model on the official formality test sets, respectively.

5.3 Evaluation Results

Based on the above evaluation metrics, we evaluate the formality control models trained at different phases for each language pair under constrained and unconstrained conditions, and compare with constrained baseline (AWS-baseline) (Nădejde et al., 2022) and unconstrained baseline

[9] https://github.com/facebookresearch/fairseq

[10] https://github.com/mjpost/sacrebleu

[11] https://github.com/Unbabel/COMET

[12] https://github.com/facebookresearch/flores/tree/main/flores200

EN-RU	To Formal				To Informal				Flores	
	M-Acc	C-F	BLEU	COMET	M-Acc	C-F	BLEU	COMET	BLEU	COMET
Multilingual pre-training	99.27%	67.83%	29.7	0.4265	0.73%	32.17%	23.7	0.3869	32.2	0.7790
+ Domain adaptation	99.71%	90.67%	33.8	0.5977	85.49%	70.67%	31.2	0.5333	27.8	0.7040
+ RTL	99.74%	**100.00%**	34.5	0.6155	97.14%	**100.00%**	33.4	0.6019	**29.4**	**0.7261**
+ Iterative RTL	**100.00%**	**100.00%**	**36.5**	**0.6472**	**100.00%**	**100.00%**	**35.6**	**0.6442**	29.0	0.7153
UMD-baseline	96.20%	92.00%	22.0	0.3492	84.10%	85.17%	21.6	0.3475	-	-
mBART50 1n	**100.00%**	91.67%	25.6	0.2916	0.00%	8.33%	19.3	0.2351	25.0	0.5950
+ Multilingual pre-training	98.15%	67.00%	28.9	0.4263	1.85%	33.00%	23.1	0.3904	32.1	0.7638
+ Domain adaptation	99.49%	98.17%	31.8	0.5336	99.73%	**99.83%**	30.8	0.5214	30.7	0.7386
+ RTL	98.76%	**100.00%**	32.3	0.5575	99.73%	**99.83%**	31.6	0.5363	30.9	0.7417
+ Iterative RTL	**100.00%**	**100.00%**	**33.7**	**0.5804**	**100.00%**	**99.83%**	**32.4**	**0.5558**	**31.0**	**0.7521**

Table 5: The overall translation quality and formality control accuracy of EN-RU models.

EN-PT	To Formal				To Informal				Flores	
	M-Acc	C-F	BLEU	COMET	M-Acc	C-F	BLEU	COMET	BLEU	COMET
Multilingual pre-training	84.23%	77.46%	34.5	0.4750	15.77%	22.54%	31.4	0.4488	51.3	0.9047
+ Domain adaptation	**100.00%**	99.67%	43.0	0.6689	96.68%	96.49%	43.7	0.6689	45.0	0.7995
+ RTL	99.47%	**100.00%**	43.1	0.6769	92.76%	**100.00%**	44.1	0.6949	**45.3**	**0.7994**
+ Iterative RTL	**100.00%**	**100.00%**	**47.4**	**0.7337**	**100.00%**	**100.00%**	**47.9**	**0.7442**	44.9	0.7926
UMD-baseline	96.30%	97.66%	27.3	0.4477	93.20%	90.82%	30.9	0.4161	-	-
mBART50 1n	86.81%	91.32%	32.2	0.5011	13.19%	8.68%	31.5	0.4955	33.8	0.6767
+ Multilingual pre-training	82.19%	77.96%	34.1	0.4872	17.81%	22.04%	31.4	0.4598	49.8	0.8753
+ Domain adaptation	**100.00%**	99.83%	39.9	0.7070	98.29%	90.32%	45.1	0.7170	46.7	0.8302
+ RTL	**100.00%**	**100.00%**	39.9	0.7165	94.97%	99.33%	45.0	0.7341	48.0	**0.8457**
+ Iterative RTL	**100.00%**	**100.00%**	**45.4**	**0.7737**	**100.00%**	**99.66%**	**49.1**	**0.7845**	**48.1**	**0.8457**

Table 6: The overall translation quality and formality control accuracy of EN-PT models.

(UMD-baseline) (Lin et al., 2022) provided by the organizers.

5.3.1 EN-VI & EN-KO

The formality control task for EN-VI and EN-KO language pairs is supervised, and we adopt the same training methods on these two language pairs. Table 3 and Table 4 are the evaluation results of the models trained at different phases for these two language pairs. From the experimental results, the multi-stage pre-training method can improve the translation quality of the model on the FLORES-200 devtest sets, while domain adaptation and RTL methods are effective in improving formality control capability of the model. Besides, domain adaptation and RTL methods have relatively little impact on the general translation quality of the model on the FLORES-200 devtest sets. Finally, we submit the Iterative RTL model as primary system.

5.3.2 EN-RU & EN-PT

The formality control tasks for the EN-RU and EN-PT language pairs are zero-shot, and we only use one-stage pre-training on these two tasks. Table 5 and Table 6 are the evaluation results of the models trained in different phases for these two language pairs. The experimental results show that domain adaptation and RTL methods are still effective in improving the zero-shot formality control capability of multilingual model. Finally, we still submit the Iterative RTL model as primary system.

6 Conclusions

This paper presents HW-TSC's submission on the IWSLT 2023 formality control task, in which we participate in both constrained and unconstrained tasks for all four language pairs. For the formality control task, we use a multi-stage pre-training method to improve the general translation quality of the basic model. We also adopt domain adaptation and RTL methods to improve the model's formality control capability. Experimental results show that these methods we have adopted are extremely effective, but how to improve general translation quality more effectively and achieve formality control with less training resources is still worthy of further research.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, 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, Xutail 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 Ojha Kr., 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)*. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 355–362.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336.

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.

Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. 2019. Domain differential adaptation for neural machine translation. *EMNLP-IJCNLP 2019*, page 59.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

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.

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.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. 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 7250–7264.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. 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.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8568–8575.

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.

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.

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. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, USA. 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*.

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.

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, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

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. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. 2018. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hwtsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation, Online. Association for Computational Linguistics*.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022. Hwtsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation, Online. Association for Computational Linguistics*.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. *Advances in Neural Information Processing Systems*, 32.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *2020 Annual Conference of the Association for Computational Linguistics*, pages 1628–1639. Association for Computational Linguistics (ACL).