Improving Machine Translation Formality Control with Weakly-Labelled Data Augmentation and Post Editing Strategies

Daniel Zhang*, Jiang Yu*, Pragati Verma*, Ashwinkumar Ganesan* & Sarah Campbell
Alexa AI, Amazon
{dyz, janyu, vpragati, gashwink, srh}@amazon.com

Abstract

This paper describes Amazon Alexa AI's implementation for the IWSLT 2022 shared task on formality control. We focus on the unconstrained and supervised task for en→hi (Hindi) and en→ja (Japanese) pairs where very limited formality annotated data is available. We propose three simple yet effective post editing strategies namely, T-V conversion, utilizing a verb conjugator and seq2seq models in order to rewrite the translated phrases into formal or informal language. Considering nuances for formality and informality in different languages, our analysis shows that a language-specific post editing strategy achieves the best performance. To address the unique challenge of limited formality annotations, we further develop a formality classifier to perform *weakly-labelled* data augmentation which automatically generates synthetic formality labels from large parallel corpus. Empirical results on the IWSLT formality testset have shown that proposed system achieved significant improvements in terms of formality accuracy while retaining BLEU score on-par with baseline.

1 Introduction

Although neural machine translation (NMT) models have achieved state-of-the-art results with high BLEU scores[1], given a language pair, they are trained on generic parallel corpora that are extracted from various open source datasets such as the Europarl corpus (Koehn; Iranzo-Sánchez et al., 2019). These datasets make an implicit assumption that there is a single translation in the target language to a sentence from the source language. But the style of the language generated, through which meaning is conveyed, is also important (Heylighen et al., 1999). Thus, there is a need to control certain attributes of the text generated in a target language such as politeness or formality.

In this paper, we present our system for the IWSLT 2022 formality control task for machine translation.[2] We focus on the unconstrained and supervised scenario for en→hi and en→ja language pairs. In the proposed system, we explore post editing strategies that correct or alter textual formality once the translation has been completed. Post editing strategies can be language specific or language agnostic. We propose three strategies, T-V conversion (deterministically converting the informal or T-form of a pronoun to its corresponding formal or V-form) , verb conjugation, and a seq2seq model that learns to transform input text to be of a formal or informal nature. The T-V conversion and verb conjugation are language-specific strategies that are applied to en→hi, and en→ja pairs respectively. These two methods are compared against an alternative seq2seq model (Enarvi et al., 2020) that is language agnostic. We show that compared to a baseline translation model provided in task, a finetuned mBART model (Liu et al., 2020) with language-specific rule-based post editing significantly improved the baseline model performance and achieved the best formality control accuracy and BLEU score.

A unique challenge in this IWSLT Formality shared task is data sparsity - only few hundred formality annotated samples are available for finetuning the formality controlled NMT model. Therefore, we further devise a data augmentation method, utilizing linguistic cues to automatically annotate a small seed set of target (i.e., Hindi and Japanese) texts with formality labels. Then the seed set is utilized to train a multilingual text formality classifier that can further mine massive parallel corpus to find extra formality annotated data. We found such weakly-labeled data augmentation strategy significantly improved en→ja performance.

The paper is organized into the following sec-

*Equal contribution.

[1] http://nlpprogress.com/english/machine_translation.html

[2] https://iwslt.org/2022/formality

T-form (Informal)	V-form (Formal)	Translation
तुम	आप	you
तुम्हारा	आपका	your
तुम्हें	आपको	to you

Table 1: Examples of T-V distinction in Hindi.

tions: §2 describes each method, §3 shows the performance of each method and language it is applied to and §4 discusses the prior work on formality.

2 System Design

2.1 Task Definition

In this submission, we focus on unconstrained and supervised formality control machine translation task. Formally, given a source segment $\boldsymbol{X} = \{x_1, x_2, ..., x_m\}$, and a formality level $l \in \{\text{formal, informal}\}$, the goal is to find the model characterized by parameters $\boldsymbol{\Theta}$ that generates the most likely translation $\boldsymbol{Y} = \{y_1, y_2, ..., y_n\}$ corresponding to the formality level:

$$\boldsymbol{Y} = \arg\max_{Y_l} P(\boldsymbol{X}, l; \boldsymbol{\Theta}) \quad (1)$$

The overall architecture and workflow of the proposed system is described in Figure 1. We present the design of each component below.

2.2 NMT & Formality Finetuning

We took a two-step process to finetune the formality controlled NMT model. First, we pretrain a generic NMT model using a large-scale parallel corpus. We chose two model architectures for building the NMT model - 1) the provided Transformer-based pretrained model implemented using Sockeye[3], and 2) a mBART model implemented using fairseq.[4] We described the datasets used and finetuning details of the NMT models in §3.1.

2.3 Post Editing

We explore three post editing strategies that rewrite the hypotheses generated for the formal/informal translations from the formality controlled NMT models.

T-V Conversion

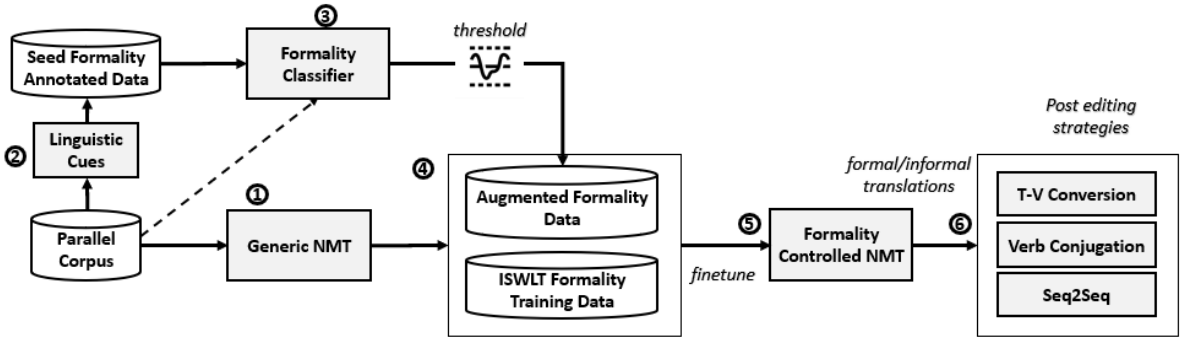
Many languages use honorifics to convey varying levels of politeness, social distance, courtesy, differences in age, etc. between addressor and addressee in a conversation. Even though the use of honorifics is not the only way to convey register (Wardhaugh, 1986), it is a way to ascertain register in sentences where pronouns are explicitly mentioned. The T-V distinction (Brown and Gilman, 1960) is a convention followed by many languages wherein different pronouns are used to convey familiarity or formality. In languages following this T-V distinction, it is applied to most pronouns of *address*, along with their verb conjugations. For sentences explicitly having pronouns of address, it is possible to write a simple, albeit noisy regex-based classifier to deterministically recognize the form (T-form or informal form; V-form or formal form) of the pronoun and thus output the grammatical register of the sentence in question. Examples of such T-V classification for Hindi is shown in Table 6.

For post editing using the T-V distinction in Hindi, we use a deterministic map of pronouns of address in T-form and their corresponding V-form in Hindi. For Hindi, this mapping is almost one-to-one, i.e. the map can be flipped along the horizontal axis to map V-form keys to T-form values without any loss in fidelity. This map can simply be looked up in the correct direction, and the values substituted for the keys in order to do a post-edit. We note that this method can be somewhat noisy as it only takes the pronouns of address into account and not the corresponding verb agreement. However, in our experiments this method has worked well in situations where some noise can be tolerated, such as post editing mistakes made by a predictive model, use in data augmentation, etc. The rules for T-V conversion and vice-versa are given in Appendix A.

Verb Conjugation

Apart from pronoun-based T-V form distinction, formality distinctions can be further encoded with verb morphology. For example, the word "to write" in Japanese 書く (kaku) can be transformed into its formal/polite form as 書きます (kakimasu). One complexity is that the conjugation of each verb depends on the class of the verb as well as its syntactic context in the sentence. For example, the verb "write!" 書け (kake) has the same stem "書" as 書く, yet its formal form is 書いてください (kaite kudasai). To address this issue, we first apply morphological analyzer that jointly identifies the verb and its corresponding verb class, as well as it Part-of-Speech Tag. Then dictionary rules adopted from (Feely et al., 2019a) are applied

[3] https://github.com/awslabs/sockeye
[4] https://github.com/pytorch/fairseq



Workflow Description. ① Parallel NMT corpus is used to train a generic NMT model. ② We leverage linguistic cues (dictionaries of formality indicators) to extract formal/informal target segments in the parallel corpus, and use then as seed formality annotated training data. ③ The seed training data is used to train a multilingual formality classifier which then during inference time, automatically labels the formality in the unannotated parallel corpus. ④ The segments that have prediction confidence >95%, together with the seed formality annotated data is selected as augmented formality data. ⑤ The augmented formality data and the provided IWSLT formality training data together finetune the NMT model for the formality control task. ⑥ Finally, the translation output of the formality controlled NMT model is further processed by one of three post editing strategies.

Figure 1: System Architecture Overview

to convert the verb into its formal/informal counterparts. In the proposed system, we applied verb conjugation for en→ja, and used Kytea[5] as the morphological analyzer.

Using Sequence-to-Sequence Model

Similar to neural machine translations architectures, post editing can be performed by a sequence-to-sequence model where the input is informal or formal while the output is the opposite. In our work, we experiment with transformer based pointer network from Enarvi et al. (2020).[6] The architecture, originally used for text summarizing, modifies the NMT transformer architecture from Vaswani et al. (2017) with a copy attention mechanism. In tasks where the input and output dictionary are highly similar such grammatical error correction or formality, copy attention allows the model to replicate parts of the input while autoregressing the output sequence (See et al., 2017). The main benefit of using such a post editing model is that it can be consistently applied across languages i.e. it is **language agnostic** and does not need any language specific editing methods compared to prior approaches.

In our implementation, we use the transformer pointer network that is part of the fairseq package and additionally finetune a pretrained mBART (Liu et al., 2020) with the formal-informal parallel corpus provided in this task and monolingual data from the standard translation corpus. For the monolingual data, the source and target sequences are the same (we copy the source text to the target), allowing the model to be trained as an auto-encoder (pre-training the copy attention mechanism). We add two tokens i.e. __F__ at the end of formal sentences and __IF__ at the end of informal sentences to provide a signal to the model of the formality change intent similar to Niu et al. (2018). These tokens are added only to the training data from the formality control corpus provided in this task while the monolingual data remains unchanged. The model is trained in two phases. The first phase pretrains the model as an auto-encoder. The second phase finetunes the model to perform the formality change.

For en→hi, we use the target language corpus from Kunchukuttan et al. (2018) while for en→ja, we reuse the corpus from Morishita et al. (2020). A subset of 20,000 Hindi or Japanese sequences are randomly sampled from the dataset.

2.4 Augment Weakly-Labeled Data

We further explore data augmentation technique to tackle the very limited access to formality annotated data. We propose to build a formality classifier that automatically labels an unannotated text as "formal" or "informal". The formality classifier can be trained using a set of seed training data with rule-based automatic annotations. In particular, we apply the T-V distinction technique for en→hi to automatically annotate Hindi texts in the en→hi parallel corpus as "formal" or "informal". Note that not all Hindi texts have T-V in-

[5] http://www.phontron.com/kytea/

[6] https://github.com/pytorch/fairseq/tree/main/examples/pointer_generator

dicators, therefore, only a small subset from the parallel corpus are labelled. Similarly, for en→ja, we follow the technique in Feely et al. (2019b), where we search for Japanese sentences that have more than one verb that indicates formality, and annotate these sentences accordingly. Tables 6-8 in Appendix summarize the T-V rule for en→hi and formality-indicating verbs for en→ja that were used to generate seed training data.

Using the formality labeled texts, we train a multilingual text classifier using multilingual Bert implemented with SimpleTransformers.[7] Then given the text classifier, we automatically label each target segments in the unannotated parallel corpus as formal or informal, which will be used during formality control finetuning. To ensure the quality of the formality label, we only select the annotated sentences that have a prediction score higher than a predefined threshold of 0.95. During formality finetuning, we upsampled the formality training data to a 1:1 ratio compared to the automatically annotated data. We summarize the size of the augmented data as well as the formality classifier accuracy in Appendix C.

3 Experiments

3.1 Training Details

The NMT model is first finetuned using a large parallel corpus. For the en→hi pair, we use IIT Bombay English-Hindi parallel corpus (Kunchukuttan et al., 2017) that contains 1.6 Million segments for training. For en→ja, we use two parallel corpora - WikiMatrix (Schwenk et al., 2019) and JParaCrawl (Morishita et al., 2019). When finetuning the mBART models for both en→hi and en→ja formality tasks, we set the following hyperparameters: maximum tokens = 512, drop out = 0.3, learning rate is 3e-05 for en→ja and 3e-04 for en→hi, random seed = 222, attention-dropout = 0.1, weight-decay = 0.0. The model is trained for a total of $20,000$ updates for en→ja and $160,000$ updates for en→hi, and the first 500 updates are used as warmup steps. The model is trained using Adam Optimizer (Kingma and Ba, 2015) with β_1 = 0.9. β_2 = 0.98, and ϵ = 1e-06. For the alternative Transformer-based NMT architecture, we pretrained the model with the same dataset, using the same model architecture and setup as the WMT14 en-de Transformer model (Gehring et al., 2017).

We further finetune the NMT models using the IWSLT Formality dataset for 1,000 steps for both language pairs. We chose a small number of training steps for this finetuning step to avoid overfitting the model and maintain a balanced BLEU score on the generic NMT performance.

3.2 Evaluation Dataset & Metrics

We evaluate the proposed system using the novel *IWSLT Formality Dataset* from Nădejde et al. (2022), which is part of the shared IWSLT task. This dataset comprises of source segments paired with two contrastive reference translations, one for each formality level (informal and formal). Since the reference was not disclosed during submission, we used a random sample of 25% of the training set as validation data and another non-overlapping 25% of the training set as test data. We report the BLEU score (Post, 2018) for measuring machine translation quality. We also report the formality control accuracy leveraging phrase-level formality annotations.[8] We use training / test dataset from both domains, i.e., telephony and topical-chats (Gopalakrishnan et al., 2019).

3.3 Results & Findings

The performance of all candidates are presented in Table 2. We make the following observations. First, compared to the pretrained base model, finetuning strategies significantly improved both BLEU score and formality accuracy. Moreover, the rule-based post editing strategy significantly improves the formality accuracy as compared to the finetuned model without post editing, while maintaining on-par BLEU scores. In particular, the formal accuracy improved from 93.9% to 95.5%, whereas the informal accuracy improved from 98.1% to 100% for the en→ja pair. For en→hi, the formal accuracy already reached 100% accuracy without post editing. Therefore, post editing was only performed to improve the informal accuracy where we observe a huge improvement from 84.4% to 97.8%.

For the seq2seq model-based post editing strategy, we only change formal text to informal text. The hypothesis generated is assumed to be formal and then post editing is applied to make it informal when necessary. Hence, the performance of the model for formal translation is the same

[7] https://simpletransformers.ai/

[8] https://github.com/amazon-research/contrastive-controlled-mt/tree/main/IWSLT2022#evaluation

	Formal BLEU		Informal BLEU		Formal Accuracy		Informal Accuracy	
	en→hi	en→ja	en→hi	en→ja	en→hi	en→ja	en→hi	en→ja
Base$_{TRF}$	19.2	13.0	15.9	13.5	0.982	0.256	0.018	0.744
Base$_{mBART}$	22.0	19.4	20.3	16.9	0.857	0.585	0.143	0.415
Finetuned$_{TRF}$	21.8	23.1	17.5	20.7	**1.000**	0.763	0.844	0.854
Finetuned$_{mBART}$	**33.7**	27.8	32.7	23.6	**1.000**	0.939	0.973	0.981
Finetuned$_{TRF}$ + Augmentation	17.1	22.1	14.5	18.3	**1.000**	0.776	0.714	0.931
Finetuned$_{mBART}$ + Augmentation	29.6	**27.9**	25.4	23.7	**1.000**	**0.962**	**1.000**	**1.000**
Finetuned$_{TRF}$ + Rule-based Editing	21.8	23.2	17.4	20.7	**1.000**	0.789	0.978	0.935
Finetuned$_{mBART}$ + Rule-based Editing	**33.7**	27.7	**32.9**	23.9	**1.000**	0.955	0.987	**1.000**
Finetuned$_{TRF}$ + Model-Based Editing	21.8*	10.4	20.4	20.7*	**1.000***	0.594	0.972	0.854*
Finetuned$_{mBART}$ + Model-Based Editing	**33.7***	27.8*	30.9	**25.8**	**1.000***	0.939*	1.000	0.262

Table 2: **Summary of overall performance**. The **Base** model is the pretrained translation model available through sockeye (Domhan et al., 2020). The **Finetuned** model represents the model finetuned on the IWSLT dataset provided. We utilize two different types of encoder-decoder models. **TRF** is the Transformer-based translation model available from sockeye, while **mBART** is the multilingual BART model. We provide results with data augmentation and post editing strategies that include rule-base editing (T-V conversion or verb conjugation) and model-based editing (using mBART transformers from Enarvi et al. (2020)). * represents the type that is generated directly by the **Finetuned$_{mBART/TRF}$** model without post editing.

as Finetuned$_{mBART}$, while the informal accuracy and BLEU score changes. We observe that in case of Japanese, the model improves the BLEU score from 23.1 to 25.8 but the informal output's accuracy score is low at 26.2%. For Hindi, the BLEU score is 30.9 while informal accuracy is 1.00%. Analysis of generated informal sentences shows that the model arbitrarily creates copies of text segments (repetition), leading to a reduced BLEU score.

We also observe that the data augmentation strategy improves the en→ja pair significantly, resulting in formal accuracy increased from 93.9% to 96.2%, and informal accuracy increases from 98.1% to 100%. In contrast, the data augmentation causes degradation on the formality accuracy for en→hi and did not improve the BLEU score. This may be due to the noisy seed training data where we used single T-V pronoun matching heuristics for Hindi to select formal/informal seed data instead of using a more complete set of heuristics including verb conjugation matching together with T-V pronoun matching. For Japanese however, the annotations are more accurate as we only select seed data that contains *multiple* formality indicating verbs.

While applying post editing strategies, we made an observation that using different conversion directions lead to very different results as indicated in Table 3. In particular, we found that unidirectional conversions, including formal→formal (i.e., convert formal hypothesis to formal) and informal→informal perform much better than cross-directional conversions such as formal→informal (i.e.,convert formal hypothesis to informal) and informal→formal. This is expected due to the typically high precision but low recall of rule-based formality conversions (Feely et al., 2019a), meaning that it cannot capture all formality pairs during the conversion, causing degraded accuracy.

Direction	BLEU		Accuracy	
	en→hi	en→ja	en→hi	en→ja
Formal hypothesis	23.5	23.8	0.896	0.789
Formal → Formal	**24.2**	**23.7**	**0.982**	**0.810**
Informal → Formal	23.7	21.6	0.981	0.612
Informal hypothesis	21.4	20.4	0.353	0.935
Informal → Informal	**22.3**	**20.5**	**0.902**	**1.000**
Formal → Informal	22.3	18.8	0.775	0.581

Table 3: Rule-based Post Editing Effect w.r.t. Conversion Directions. → represents the direction in which post editing happens.

	Testset	BLEU	COMET
en→hi	newstest2014	38.9	0.8741
en→ja	newstest2020	19.4	0.3783

Table 4: Generic NMT performance.

Finally, we report the performance of our submitted system on generic NMT test set, and blind IWSLT test set in Table 4 and Table 5 as required by the task. For en→hi, our submitted system employed finetuned mBART + data augmentation strategy which demonstrated the best performance on the development set. For en→ja, the submitted system employs finetuned mBART + data augmentation + post editing (verb conjugation). We have observed that the formality accuracy improvements are consistent with the observation in

	Formal BLEU		Informal BLEU		Formal Accuracy		Informal Accuracy	
	en→hi	**en→ja**	**en→hi**	**en→ja**	**en→hi**	**en→ja**	**en→hi**	**en→ja**
Finetuned$_{\text{mBART}}$	30.3	27.1	29.3	24.6	0.989	0.858	0.919	0.949
Our System	27.7	28.9	22.6	25.1	0.998	0.888	0.993	0.988

Table 5: Formality control performance on blind submission.

Table 2. Specifically, compared to the finetuned mBART candidate system, we observed 0.09% formal and 7.4% informal absolute accuracy improvements for en→hi. For en→ja, we observed 3.0% formal and and 3.9% informal absolute accuracy improvements. These results indicate the effectiveness of the proposed post editing and data augmentation strategies. We observed en→ja improved BLEU score as well. Interestingly, we observed that the proposed system for en→hi had worse BLEU score compared to the finetuned mBART model. One potential cause of this is that the formality augmented data for en→hi came from a different domain than the test set which is conversational in nature. We can potentially improve the BLEU score by augmenting the training data with more conversational data or up-sampling the IWSLT formality data during training. We leave these directions for future improvement.

4 Background

The task of controlling formality in the output of machine translation has drawn much attention in recent MT architectures. Earlier approaches are rule-based systems where non-linguistic information such as speaker profile and gender information is used to personalized MT with gender/speaker-specific data (Rabinovich et al., 2016; Michel and Neubig, 2018). More recently, Niu et al. (2017) coined the term Formality Sensitive Machine Translation (FSMT), and proposed lexical formality models to control the level of formality of MT output by selecting phrases of that are most similar to a desired formality level from the k-best list during decoding. Alternatively, a popular formality control approach is by leveraging side constraints in NMT where a style tag (e.g., <Formal>/<Informal>) is attached to the beginning of each source example, and the NMT model is forced to "pay attention to" these style tags during translation (Sennrich et al., 2016; Niu and Carpuat, 2020).

Formality control for machine translation is closely related to formality transfer (FT), which is the task of automatically transforming text in one formality style (e.g., "informal") into another (e.g., polite) (Niu et al., 2018). The FT task usually takes a seq2seq-like approach (Zhang et al., 2020) given parallel corpus such as Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018). These FT models are often applied as a rewriting mechanism after the MT outputs are generated. Recently, Niu et al. (2018) proposed a novel multi-task model that jointly perform FT and FSMT. Honorifics based post editing approaches have also been widely deployed for formality control tasks. A widespread instance of using honorifics to determine register is the grammatical T-V distinction (Brown and Gilman, 1960), distinguishing between the informal (Latin *Tu*) and the formal (Latin *Vos*). Alternatively, verb conjugation combined with syntactic parsing has been used to alter the inflection of the main verb of the sentence to achieve multiple levels of formality (Feely et al., 2019a).

5 Conclusion

In this paper, we target improving the machine translation formality control performance given limited formality annotated training data. We explored three different strategies including rule-based post editing, seq2seq point networks, and formality classifier-based augmentation. We found that data augmentation using formality classifier significantly improved formality accuracy on en→ja pair. We also found that post editing strategies on top of finetuned mBART models are simple and effective ways to improve the formality control performance. Results on the IWSLT testset have indicated performance improvements in terms of formality accuracy in both en→hi and en→ja pairs while retaining on-par BLEU score.

References

R. Brown and A. Gilman. 1960. The pronouns of power and solidarity. In T. A. Sebeok, editor, *Style in*

Language, pages 253–276. MIT Press, Cambridge, Mass.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Weston Feely, Eva Hasler, and Adrià de Gispert. 2019a. Controlling japanese honorifics in english-to-japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53.

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

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Francis Heylighen, Jean Marc Dewaele, and Léo Apostel. 1999. Formality of language: definition, measurement and behavioral determinants.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2019. Europarl-st: A multilingual corpus for speech translation of parliamentary debates.

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. Europarl: A parallel corpus for statistical machine translation.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. *CoRR*, abs/2001.08210.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2019. Jparacrawl: A large scale web-based english-japanese parallel corpus. *arXiv preprint arXiv:1911.10668*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 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.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. *arXiv preprint arXiv:1806.04357*.

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.

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.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

R Wardhaugh. 1986. *Introduction to Sociolinguistics* , 2nd edition. Wiley Series in Probability and Statistics. Cambridge: Blackwell.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:2005.07522*.

Appendix

A T-V conversion

Following tables 6 and 7, provide a list of rules applied to the dataset in order to change formality. Table 6 provides rules to change the language from informal to formal, while table 7 performs the inverse.

T-form (Informal)	V-form (Formal)
"तुम्हें"	"आपको"
"तुमको"	"आपको"
"तुम्हारे"	"आपके"
"तुम्हारा"	"आपका"
"तुम्हारी"	"आपकी"
"तुम"	"आप"
" हो "	" हैं "

Table 6: Rules for converting T-form to V-form for Hindi. The order of applying the rules is significant, along with the spaces within quotes, if present.

V-form (Formal)	T-form (Informal)
"आपको"	"तुम्हें"
"आपके"	"तुम्हारे"
"तुम्हारे"	"आपके"
"आपका"	"तुम्हारा"
"आपकी"	"तुम्हारी"
"आप "	"तुम "
" हैं "	" हो "

Table 7: Rules for converting V-form to T-form for Hindi. The order of applying the rules is significant, along with the spaces within quotes, if present.

B Formality-indicating verbs for Japanese

	Formality-indicating verbs
Formal	ございます, いらっしゃいます, おります, なさいます, 致します, ご覧になります, おいでになります, 伺います, 参ります, 存知します, 存じ上げます, 召し上がります, 頂く, 頂きます, 頂いて, 差しあげます, 下さいます, おっしゃいます, 申し上げます, 拝見します, お目に掛かります
Informal	だ, だった, じゃない, じゃなかった, だろう, だから, だけど, だって, だっけ, そうだ, ようだ

Table 8: Indicating verbs for generating seed training data for en→ja formality classifier.

C Formality Classifier Accuracy and Data Sizes

		Precision	Recall	F1
en→hi	Formal	0.802	0.757	0.779
	Informal	0.776	0.827	0.801
en→ja	Formal	0.885	0.817	0.850
	Informal	1.0	0.852	0.920

Table 9: Formality classifier accuracy using IWSLT formality testset as groundtruth.

	Seed	Unlabeled	Augmented
en→hi	142,900	1,667,803	142,900*
en→ja	9,856	13,956,005	26,294

Table 10: Weakly labeled data sizes. *Due to the relatively poor performance of the formality classifier for en→hi, only the seed training data was used for data augmentation.

D Post Editing Seq2seq Model

Following are details about the post editing model utilized to perform formality change. We use a base model architecture from Enarvi et al. (2020). As described in §2.3, the transformer model is trained in two phases, viz., pretraining with monolingual language data and then finetuning the formality control dataset.

Following are the hyper-parameters with which the model is trained and later inference is performed:

Hyperparameter	**Value**
Tokenizer	Sacremoses
Pointer layers	-2
Pointer head	2
Pointer markers	1000
Label Smoothing	0.1
Weight Decay	0.0
Learning Rate	0.001
Batch Size	512
Total Number of Updates	20000

Table 11: **Hyperparameters of Post Editing model**. The table shows values of hyperparameters that are manually set. All other parameters are set to their default value in the package. *Pointer layers* are the attention layers being pointed to and *Pointer head* denotes the number of attention heads used.