Parallel Data Augmentation for Formality Style Transfer

Yi Zhang[1]*, **Tao Ge**[2], **Xu Sun**[1]
[1]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Microsoft Research Asia, Beijing, China
{zhangyi16,xusun}@pku.edu.cn
tage@microsoft.com

Abstract

The main barrier to progress in the task of Formality Style Transfer is the inadequacy of training data. In this paper, we study how to augment parallel data and propose novel and simple data augmentation methods for this task to obtain useful sentence pairs with easily accessible models and systems. Experiments demonstrate that our augmented parallel data largely helps improve formality style transfer when it is used to pre-train the model, leading to the state-of-the-art results in the GYAFC benchmark dataset[1].

1 Introduction

Formality style transfer (FST) is defined as the task of automatically transforming a piece of text in one particular formality style into another (Rao and Tetreault, 2018). For example, given an informal sentence, FST aims to preserve the style-independent content and output a formal sentence.

Previous work tends to leverage neural networks (Xu et al., 2019; Niu et al., 2018; Wang et al., 2019) such as seq2seq models to address this challenge due to their powerful capability and large improvement over the traditional rule-based approaches (Rao and Tetreault, 2018). However, the performance of the neural network approaches is still limited by the inadequacy of training data: the public parallel corpus for FST training – GYAFC (Rao and Tetreault, 2018) – contains only approximately 100K sentence pairs, which can hardly satiate the neural models with millions of parameters.

To tackle the data sparsity problem for FST, we propose to augment parallel data with three specific data augmentation methods to help improve the model's generalization ability and reduce the overfitting risk. Besides applying the widely used back translation (BT) method (Sennrich et al., 2016a) in Machine Translation (MT) to FST, our data augmentation methods include formality discrimination (F-Dis) and multi-task transfer (M-Task). They are both novel and effective in generating parallel data that introduces additional formality transfer knowledge that cannot be derived from the original training data. Specifically, F-Dis identifies useful pairs from the paraphrased pairs generated by cross-lingual MT; while M-task leverages the training data of Grammatical Error Correction (GEC) task to improve formality, as shown in Figure 1.

FST (test instance)	Input (informal)	*I dunno, even if she like you, and then she 'll prob.*
	Reference (formal)	*I don't know. She probably will if she likes you.*
F-Dis	Source	*I dunno... good luck.*
	MT French	*Je ne sais pas... bonne chance.*
	MT Target	*I don't know ... Good luck.*
M-Task	Source	*I think she like cat too.*
	Target	*I think she likes cat too.*

Figure 1: An example that **F**ormality **S**tyle **T**ransfer (FST) benefits from data augmented via **f**ormality **dis**crimination (**F-Dis**) and **m**ulti-**task** transfer (**M-Task**). The mapping knowledge indicated by the color (blue→pink) in FST test instance occur in the pairs augmented by F-Dis and M-Task. F-Dis identifies useful sentence pairs from paraphrased sentence pairs generated by cross-lingual MT, while M-Task utilizes training data from GEC to help formality improvement.

Experimental results show that our proposed data augmentation methods can harvest large amounts of augmented parallel data for FST. The augmented parallel data proves helpful and significantly helps improve formality style transfer when it is used to pre-train the model, allowing the model to achieve the state-of-the-art results in the GYAFC benchmark dataset.

[*] Work done during the internship at Microsoft Research.

[1] Our augmented data is available at https://github.com/lancopku/Augmented_Data_for_FST

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3221–3228
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

2 Approach

2.1 Data Augmentation for Formality Style Transfer

We study three data augmentation methods for formality style transfer: back translation, formality discrimination, and multi-task transfer. We focus on informal→formal style transfer since it is more practical in real application scenarios.

2.1.1 Back translation

The original idea of back translation (BT) (Sennrich et al., 2016a) is to train a target-to-source seq2seq (Sutskever et al., 2014; Cho et al., 2014) model and use the model to generate source language sentences from target monolingual sentences, establishing synthetic parallel sentences. We generalize it as our basic data augmentation method and use the original parallel data to train a seq2seq model in the formal-to-informal direction. Then, we can feed formal sentences to this model that is supposed to be capable of generating their informal counterparts. The formal input and the informal output sentences can be paired to establish augmented parallel data.

2.1.2 Formality discrimination

According to the observation that an informal sentence tends to become a formal sentence after a round-trip translation by MT models that are mainly trained with formal text like news, we propose a novel method called formality discrimination to generate formal rewrites of informal source sentences by means of cross-lingual MT models. A typical example is shown in Figure 2.

To this end, we collect a number of potentially informal English sentences (e.g., from online forums). Formally, we denote the collected sentences as $\mathcal{S} = \{\boldsymbol{s}_i\}_{i=1}^{|\mathcal{S}|}$ where $\boldsymbol{s}_i$ represents the i-th sentence. We first translate[2] them into a pivot language (e.g., French) and then translate them back into English, as Figure 2 shows. In this way, we obtain a rewritten sentence $\boldsymbol{s}'_i$ for each sentence $\boldsymbol{s}_i \in \mathcal{S}$.

To verify whether $\boldsymbol{s}'_i$ improves the formality compared to $\boldsymbol{s}_i$, we introduce a formality discriminator which in our case is a Convolutional Neural Network (CNN) to quantify the formality level of a sentence. We trained the formality discriminator with the sentences and their formality labels in the FST corpus (e.g., GYAFC). The pairs $(\boldsymbol{s}_i, \boldsymbol{s}'_i)$ where $\boldsymbol{s}'_i$ largely improves the formality of $\boldsymbol{s}_i$ will be selected as the augmented data. The resulting data set $\mathcal{T}_{aug}$ is such a set of pairs:

$$\mathcal{T}_{aug} = \{(\boldsymbol{s}_i, \boldsymbol{s}'_i) | P_+(\boldsymbol{s}'_i) - P_+(\boldsymbol{s}_i) \geq \sigma\} \quad (1)$$

where $P_+(\boldsymbol{x})$ is the probability of sentence $\boldsymbol{x}$ being formal, predicted by the discriminator, and σ is the threshold[3] for augmented data selection. In this way, we can obtain much helpful parallel data with valuable rewriting knowledge that is not covered by the original parallel data.

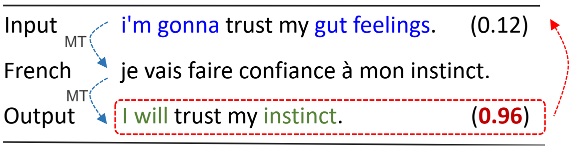
Input
MT
i'm gonna trust my gut feelings.
(0.12)
French
je vais faire confiance à mon instinct.
MT
Output
I will trust my instinct.
(0.96)

Figure 2: Formality discrimination for FST. The numbers following the sentences are formality scores predicted by a formality discriminator. The pair (connected by the red dashed arrow) that obtains significant formality improvement will be kept as augmented data.

2.1.3 Multi-task transfer

In addition to back translation and formality discrimination that use artificially generated sentence pairs for data augmentation, we introduce multi-task transfer that uses annotated sentence pairs from other seq2seq tasks. We observe that informal texts are usually ungrammatical while formal texts are almost grammatically correct. Therefore, a desirable FST model should possess the ability to detect and rewrite ungrammatical texts, which has been verified by the previous empirical study (Ge et al., 2019) showing that using a state-of-the-art grammatical error correction (GEC) model to post-process the outputs of an FST model can improve the result. Inspired by this observation, we propose to transfer the knowledge from GEC to FST by leveraging the GEC training data as the augmented parallel data to help improve formality. An example is illustrated in Figure 1 in which the annotated data for GEC provides knowledge to help the model rewrite the ungrammatical informal sentence.

2.2 Pre-training with Augmented Data

In general, massive augmented parallel data can help a seq2seq model to learn contextualized representations, sentence generation and source-target alignments better. When the augmented parallel

[2] https://translate.google.com/

[3] $\sigma = 0.6$ in our experiments.

data is available, previous studies (Sennrich et al., 2016a; Edunov et al., 2018; Karakanta et al., 2018; Wang et al., 2018) for seq2seq tasks are inclined to train a seq2seq model with original training data and augmented data simultaneously. However, augmented data is usually noisier and less valuable than original training data. In simultaneous training, the massive augmented data tends to overwhelm the original data and introduce unnecessary and even erroneous editing knowledge, which is undesirable for our task.

To better exploit the augmented data, we propose to first pre-train the model with augmented parallel data and then fine-tune the model with the original training data. In our pre-training & fine-tuning (PT&FT) approach, the augmented data is not treated equally to the original data; instead it only serves as prior knowledge that can be updated and even overwritten during the fine-tuning phase. In this way, the model can better learn from the original data without being overwhelmed or distracted by the augmented data. Moreover, separating the augmented and original data into different training phases makes the model become more tolerant to noise in augmented data, which reduces the quality requirement for the augmented data and enables the model to use noisier augmented data and even training data from other tasks.

3 Experiments

In this section, we present the experimental settings and related experimental results. We focus on informal→formal style transfer since it is more practical in real application scenarios.

3.1 Experimental Settings

We use GYAFC benchmark dataset (Rao and Tetreault, 2018) for training and evaluation. GYAFC's training split contains a total of 110K annotated informal-formal parallel sentences, which are annotated via crowd-sourcing of two domains: *Entertainment & Music* (E&M) and *Family & Relationships* (F&R). In its test split, there are 1,146 and 1,332 informal sentences in E&M and F&R domain respectively and each informal sentence has 4 referential formal rewrites. We use all the three data augmentation methods we introduced and obtain a total of 4.9M augmented pairs. Among them, 1.6M are generated by back-translating (BT) formal sentences identified (as formal) by the formality discriminator in E&M and F&R domain on Yahoo Answers L6 corpus[4], 1.5M are derived by formality discrimination (F-Dis) by using French, German and Chinese as pivot languages, and 1.8M are from multi-task transfer (M-task) from the public GEC data (Lang-8 (Mizumoto et al., 2011; Tajiri et al., 2012) and NUCLE (Dahlmeier et al., 2013)). The informal sentences used in F-Dis strategy are also from Yahoo Answers L6 corpus.

Model	**E&M** *BLEU*	**F&R** *BLEU*
Original data	69.44	74.19
Augmented data	51.83	55.66
ST	59.93	63.16
ST (up-sampling)	68.43	73.04
ST (down-sampling)	68.54	73.69
PT&FT	**72.63**	**77.01**

Table 1: The comparison of simultaneous training (ST) and Pre-train & Fine-tuning (PT&FT). Down-sampling and up-sampling are for balancing the size of the augmented data and the original data. Specifically, down-sampling samples augmented data, while up-sampling increases the frequency of the original data.

We use the Transformer (base) (Vaswani et al., 2017) as the seq2seq model with a shared vocabulary of 20K BPE (Sennrich et al., 2016b) tokens. We adopt the Adam optimizer to pre-train the model with the augmented parallel data and then fine-tune it with the original parallel data. In pre-training, the dropout rate is set to 0.1 and the learning rate is set to 0.0005 with 8000 warmup steps and scheduled to an inverse square root decay after warmup; while during fine-tuning, the learning rate is set to 0.00025. We pre-train the model for 80k steps and fine-tune the model for a total of 15k steps. The CNN we use as the formality discriminator has filter sizes of 3, 4, 5 with 100 feature maps. The dropout rate is set to 0.5. It achieves an accuracy of 93.09% over the GYAFC test set.

3.2 Experimental Results

3.2.1 Effect of Proposed Approach

Table 1 compares the results of the models trained with simultaneous training (ST) and pre-training & fine-tuning (PT&FT). ST with the augmented and original data leads to a performance decline, because the noisy augmented data cannot achieve desirable performance by itself and may distract the model from exploiting the original data in simultaneous training. In contrast, PT&FT only uses

[4] https://webscope.sandbox.yahoo.com/catalog.php

Model	E&M $BLEU$	F&R $BLEU$
Original data	69.44	74.19
Pre-training & Fine-tuning		
+ BT	71.18	75.34
+ F-Dis	71.72	76.24
+ M-Task	71.91	76.21
+ BT + M-Task + F-Dis	**72.63**	**77.01**

Table 2: The comparison of different data augmentation methods for FST.

System	E&M $BLEU$	F&R $BLEU$
No-edit	50.28	51.67
Rule	60.37	66.40
PBMT	66.88	72.40
NMT	58.27	68.26
NMT-PBMT	67.51	73.78
NMT-MTL	71.29	74.51
NMT-MTL-Ensemble*	72.01	75.33
GPT-CAT	72.70	77.26
GPT-Ensemble*	69.86	76.32
Our Approach	72.63	77.01
Our Approach*	**74.24**	**77.97**

Table 3: The comparison of our approach to the state-of-the-art results. * denotes the ensemble results.

the augmented data in the pre-training phase and treats it as the prior knowledge supplementary to the original training data, reducing the negative effects of the augmented data and improving the results.

Table 2 compares the results of different data augmentation methods with PT&FT. Pre-training with augmented data generated by BT enhances the generalization ability of the model, thus we observe an improvement over the baseline. However, it does not introduce any new informal-to-formal transfer knowledge, leading to the least improvement among the three methods. In contrast, both F-Dis and M-Task introduce abundant transfer knowledge for FST. The augmented data of F-Dis includes various informal→formal rewrite knowledge derived from the MT models, allowing the model to better handle the test instances whose patterns are never seen in the original training data; while M-Task introduces GEC knowledge that helps improve formality in terms of grammar.

We then combine all these beneficial augmented data for pre-training. As expected, the combination strategy achieves further improvement as shown in Table 2 since the it enables the model to take advantage of all the data augmentation methods.

3.2.2 Comparison with State-of-the-Art Results

We compare our approach to the following previous approaches in the GYAFC benchmark:

- Rule, PBMT, NMT, PBMT-NMT: Rule-based, phrase-based MT, NMT, PBMT-NMT hybrid model (Rao and Tetreault, 2018).
- NMT-MTL: NMT model with multi-task learning (Niu et al., 2018).
- GPT-CAT, GPT-Ensemble: fine-tuned encoder-decoder models (Wang et al., 2019) initialized by GPT (Radford et al., 2019). Specifically, GPT-CAT concatenates the original input sentence and the input sentence preprocessed by rules as input, while GPT-Ensemble is the ensemble of two GPT-based encoder-decoder models: one takes the original input sentence as input, the other takes the preprocssed sentence as input.

Following Niu et al. (2018), we train 4 independent models with different initializations for ensemble decoding. According to Table 3, our single model performs comparably to the state-of-the-art GPT-based encoder-decoder models (more than 200M parameters) with only 54M parameters. Our ensemble model further advances the state-of-the-art result only with a comparable model size to the GPT-based single model (i.e., GPT-CAT).

We also conduct human evaluation. Following Rao and Tetreault (2018), we assess the model output on three criteria: *formality*, *fluency* and *meaning preservation*. We compare our baseline model trained with original data, our best performing model and the previous state-of-the-art models (NMT-MTL and GPT-CAT). We randomly sample 300 items and each item includes an input and four outputs that shuffled to anonymize model identities. Two annotators are asked to rate the outputs on a discrete scale of 0 to 2. More details can be found in the appendix. The results are shown in Table 4 which demonstrates that our model is consistently well rated in human evaluation.

3.2.3 Analysis of Pivot Languages in Feature Discrimination

We also conduct an exploratory study of the pivot languages used in formality discrimination. Among the three pivot languages (i.e. French, German and Chinese) in our experiments, it is interest-

Model	Formality	Fluency	Meaning
Original data	1.31	1.77	1.80
NMT-MTL	1.34	1.78	**1.92***
GPT-CAT	1.42	1.84*	1.90
Ours	**1.45***	**1.85**$*^\dagger$	**1.92***

Table 4: Results of human evaluation of FST. Scores marked with $*/^\dagger$ are significantly different from the scores of Original data / NMT-MTL ($p < 0.05$ in significance test).

French	German	Chinese
300k	530k	680k

Table 5: The sizes of augmented datasets generated by F-Dis based on different pivot languages.

ing to observe a significant difference in the sizes of the obtained parallel data given the same source sentences and filter threshold, as shown in Table 5. Using Chinese as the pivot language results in the most data, probably due to the fact that Chinese and English belong to different language systems. The formality of original informal English sentences may be lost during translation, which turns out to facilitate the MT system to translate Chinese back into formal English. In contrast, French and German have much in common with English, especially for French in terms of the lexicon (Baugh and Cable, 1993). The translated sentences are likely to maintain informal sense, which hinders the MT system from generating formal English translations.

We compare the performance with augmented data generated by three pivot languages separately in Table 6. Manual inspection reveals that a few pairs have the issue of meaning inconsistency in all the three sets, which mainly arises from the translation difficulties caused by omissions and poor grammaticality in informal sentences and the segmentation ambiguity in some pivot languages like Chinese. Among the three languages, the Chinese-based augmented data introduces more noise due to the additional segmentation ambiguity problem but brings fair improvement because of its largest size. In contrast, the German-based augmented data has relatively high quality and a moderate size, leading to the best result in our experiments.

4 Related Work

Data augmentation has been much explored for seq2seq tasks like Machine Translation (He et al., 2016; Fadaee et al., 2017; Zhang et al., 2018b; Poncelas et al., 2018; Edunov et al., 2018; Li et al., 2019) and Grammatical Error Correction (Kiyono et al., 2019; Grundkiewicz et al., 2019; Zhao et al., 2019; Zhou et al., 2019; Ge et al., 2018a,b; Xie et al., 2018; Yuan et al., 2016; Rei et al., 2017). For text style transfer, however, due to the lack of parallel data, many studies focus on unsupervised approaches (Luo et al., 2019; Wu et al., 2019; Zhang et al., 2018a) and there is little related work concerning data augmentation. As a result, most recent work (Jhamtani et al., 2017; Xu et al., 2012) that models text style transfer as MT suffers from a lack of parallel data for training, which seriously limits the performance of powerful models. To solve this pain point, we propose novel data augmentation methods and study the best way to utilize the augmented data, which not only achieves a success in formality style transfer, but also would be inspiring for other text style transfer tasks.

Model	E&M	F&R
	BLEU	*BLEU*
Original data	69.44	74.19
F-Dis (Fr)	70.09	74.52
F-Dis (De)	71.15	75.18
F-Dis (Zh)	70.51	74.79

Table 6: Performances of formality discrimination based on different pivot languages: French (Fr), German (De) and Chinese (Zh).

5 Conclusion

In this paper, we propose novel data augmentation methods for formality style transfer. Our proposed data augmentation methods can effectively generate diverse augmented data with various formality style transfer knowledge. The augmented data can significantly help improve the performance when it is used for pre-training the model and leads to the state-of-the-art results in the formality style transfer benchmark dataset.

Acknowledgements

We thank all the reviewers for providing the constructive suggestions. This work is partly supported by Beijing Academy of Artificial Intelligence. Xu Sun is the corresponding author of this paper.

References

Albert C Baugh and Thomas Cable. 1993. *A history of the English language*. Routledge.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 567–573.

Tao Ge, Furu Wei, and Ming Zhou. 2018a. Fluency boost learning and inference for neural grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1055–1065.

Tao Ge, Furu Wei, and Ming Zhou. 2018b. Reaching human-level performance in automatic grammatical error correction: An empirical study. *arXiv preprint arXiv:1807.01270*.

Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Automatic grammatical error correction for sequence-to-sequence text generation: An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6059–6064.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *CoRR*, abs/1707.01161.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1-2):167–189.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. *arXiv preprint arXiv:1909.00502*.

Rumeng Li, Xun Wang, and Hong Yu. 2019. Metamt, a metalearning method leveraging multiple domain data for low resource machine translation. *arXiv preprint arXiv:1912.05467*.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5116–5122. ijcai.org.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155.

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, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1008–1021.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140.

Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. 2017. Artificial error generation with machine translation and syntactic patterns. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 287–292.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. 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.

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*.

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.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for esl learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 198–202. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. *CoRR*, abs/1808.07512.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. 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 3571–3576, Hong Kong, China. Association for Computational Linguistics.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4873–4883. Association for Computational Linguistics.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 619–628.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *CoRR*, abs/1903.06353.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2899–2914.

Zheng Yuan, Ted Briscoe, and Mariano Felice. 2016. Candidate re-ranking for smt-based grammatical error correction. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 256–266.

Yi Zhang, Jingjing Xu, Pengcheng Yang, and Xu Sun. 2018a. Learning sentiment memories for sentiment modification without parallel data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1103–1108. Association for Computational Linguistics.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018b. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.

Wangchunshu Zhou, Tao Ge, Chang Mu, Ke Xu, Furu Wei, and Ming Zhou. 2019. Improving grammatical error correction with machine translation pairs. *arXiv preprint arXiv:1911.02825*.

A Details of Human Evaluation

We describe the grading standard of the three criteria we present in the main paper for FST: *formality*, *fluency* and *meaning preservation*. The outputs are rated on a discrete scale of 0 to 2. We hire two annotators who major in Linguistics and have received Bachelor degree.

Formality Given the informal source sentence and an output, the annotators are asked to rate the formality of a sentence according to the formality improvement level, regardless of fluency and meaning. If the output shows significant formality improvement over the input, it will be rated 2 points. If the output is just slightly more formal than the input, it will be rated 1 point. If the output shows no improvement in the formality or even decreases the formality, it will be rated 0 point.

Fluency Given the outputs, the annotators are asked to evaluate the fluency of each sentence in isolation. A sentence is considered to be *fluent* if *it makes sense and is grammatically correct*. The sentences satisfying the requirements will be rated 2 points. The sentences with minor errors will be rated 1 point. If the errors lead to confusing meaning, we give it 0 point.

Meaning preservation Given the output sentence and the corresponding source sentence, the annotators are asked to estimate how much information is preserved of the output compared to the input sentences. If the output sentence and the input exactly convey the same idea, the corresponding system of the output gets 2 points. If they are mostly equivalent but different in some trivial details, the corresponding system gets 1 point. If the output omits some important details that affect the sentence's meaning, the system will get no credit.

For inter-annotator agreement, we calculate the Pearson correlation coefficient of two annotators over the three criteria. The Pearson correlation over the formality criteria is 0.62. For fluency and meaning preservation, the correlation scores are 0.69 and 0.61, respectively.