Multilingual pre-training with Language and Task Adaptation for Multilingual Text Style Transfer

Anonymous ACL submission

Abstract

We exploit the pre-trained seq2seq model mBART for multilingual text style transfer. Using machine translated data as well as gold aligned English sentences yields state-of-the-art results in the three target languages we consider. Besides, in view of the general scarcity of parallel data, we propose a modular approach for multilingual formality transfer, which consists of two training strategies that target adaptation to both language and task. Our approach achieves competitive performance without monolingual task-specific parallel data and can be applied to other style transfer tasks as well as to other languages.

1 Introduction

Text style transfer (TST) is a text generation task where a given sentence must get rewritten changing its style while preserving its meaning. Traditionally, tasks such as swapping the polarity of a sentence (e.g. "This restaurant is getting worse and worse."↔"This restaurant is getting better and better.") as well as changing the formality of a text (e.g. "it all depends on when ur ready."↔"It all depends on when you are ready."), are considered as instances of TST. We focus here on the latter case only, i.e. *formality transfer*, because (i) recent work has shown that polarity swap is less of a style transfer task, since meaning is altered in the transformation (Lai et al., 2021a), and (ii) data in multiple languages has recently become available for formality transfer (Briakou et al., 2021b).

Indeed, mostly due to the availability of parallel training and evaluation data[1], almost all prior TST work focuses on monolingual (English) text (Rao and Tetreault, 2018; Li et al., 2018; Prabhumoye et al., 2018; Cao et al., 2020). As a first step towards multilingual style transfer, Briakou et al. (2021b) have released XFORMAL, a benchmark of multiple formal reformulations of informal text in Brazilian Portuguese (BR-PT), French (FR), and Italian (IT). For these languages the authors have manually created evaluation datasets. On these, they test several monolingual TST baseline models developed without any gold parallel training data, and several neural models trained from scratch on language-specific pairs obtained by machine translating GYAFC, the reference corpus for formality transfer in English (Rao and Tetreault, 2018). Briakou et al. (2021b) find that the models trained on translated parallel data do not outperform a simple rule-based system based on handcrafted transformations, especially on content preservation, and conclude that formality transfer on languages other than English is particularly challenging.

One reason for the poor performance could be the low quality (observed upon our own manual inspection) of the pseudo-parallel data, especially the informal side. Since machine translation systems are usually trained with formal texts like news (Zhang et al., 2020), informal texts are harder to translate, or might end up more formal when translated. But most importantly, the neural models developed by Briakou et al. (2021b) do not take advantage of two recent findings: (i) pre-trained models, especially the sequence-to-sequence model BART (Lewis et al., 2020), have proved to help substantially with content preservation in style transfer (Lai et al., 2021b); (ii) Multilingual Neural Machine Translation (Johnson et al., 2017; Aharoni et al., 2019; Liu et al., 2020) and Multilingual Text Summarization (Hasan et al., 2021) have achieved impressive results leveraging multilingual models which allow for cross-lingual knowledge transfer.

In this work we use the multilingual large model mBART (Liu et al., 2020) to model style transfer in a multilingual fashion exploiting available parallel data of one language (English) to transfer the task and domain knowledge to other target languages. To address real-occurring situations, in

[1]"Parallel data" in this paper refers to sentence pairs in the same language, with the same content but different formality.

our experiments we also simulate complete lack of parallel data for a target language (even machine translated), and lack of style-related data at all (though availability of out-of-domain data). Language specificities are addressed through adapter-based strategies (Pfeiffer et al., 2020; Üstün et al., 2020, 2021). We obtain state-of-the-art results in all three target languages we consider, and propose a modular methodology that can be applied to other style transfer tasks as well as to other languages.

2 Approach and Data

As a base experiment aimed at exploring the contribution of mBART (Liu et al., 2020; Tang et al., 2020) for multilingual style transfer, we fine-tune this model with parallel data specifically developed for style transfer in English (original) and three other languages (machine translated).

Next, in view of the common situation where parallel data for a target language is not available, we propose a two-step adaptation training approach on mBART that enables modular multilingual TST. We avoid iterative back-translation (IBT) (Hoang et al., 2018), often used in previous TST work (Lample et al., 2019; Prabhumoye et al., 2018; Luo et al., 2019; Yi et al., 2020; Lai et al., 2021a), since it has been shown to be computationally costly (Üstün et al., 2021; Stickland et al., 2021). We still run comparison models that use it.

In the first adaptation step, we address the problem of some languages being not well represented in mBART[2], which preliminary experiments have shown to hurt our downstream task. We conduct a language adaptation denoising training using unlabelled data for the target language. In the second step, we address the task at hand through fine-tuning cross-attention with auxiliary gold parallel English data adapting the model to the TST task.

For TST fine-tuning, we use parallel training data, namely formal/informal aligned sentences (both manually produced for English and machine translated for three other languages). For the adaptation strategies, we also collect formality and generic non-parallel data. Details follow.

English formality data GYAFC (Rao and Tetreault, 2018) is an English dataset of aligned formal and informal sentences. Gold parallel pairs are provided for training, validation, and test.

Multilingual formality data XFORMAL (Briakou et al., 2021b) is a benchmark for multilingual formality style transfer, which provides an evaluation set that consists of four formal rewrites of informal sentences in BR-PT, FR, and IT. This dataset contains pseudo-parallel corpora in each language, obtained via machine translating the English GYAFC pairs.

Language-specific formality non-parallel data Following Rao and Tetreault (2018) and Briakou et al. (2021b), we crawl the domain data in target language from Yahoo Answers[3]. We then use the style regressor from Briakou et al. (2021a) to predict formality score σ[4] of the sentence to automatically select sentences in each style direction.

Language-specific generic non-parallel data 5 M sentences containing 5 to 30 words for each language randomly selected from News Crawl[5].

3 Adaptation Training

To adapt mBART to multilingual TST, we employ two adaptation training strategies that target language and task respectively.

3.1 Language Adaptation

As shown in Figure 1(a), we introduce a module for language adaptation. Inspired by previous work (Houlsby et al., 2019; Bapna and Firat, 2019), we use an adapter (ADAPT; ~50M parameters), which is inserted into each layer of the Transformer encoder and decoder, after the feed-forward block.

Following Bapna and Firat (2019) and Üstün et al. (2021), the ADAPT module A_i at layer i consists of a layer-normalization LN of the input $x_i \in \mathbb{R}^h$ followed by a down-projection $W_{down} \in \mathbb{R}^{h \times h}$, a non-linearity and a up projection $W_{up} \in \mathbb{R}^{h \times h}$ combined with a residual connection with the input x_i:

$$A(x_i) = W_{up}\text{RELU}(W_{down}\text{LN}(x_i)) + x_i \quad (1)$$

Language adaptation training Following mBART's pretraining, we conduct the language adaptation training on a denoising task, which aims to reconstruct text from a version corrupted with a noise function:

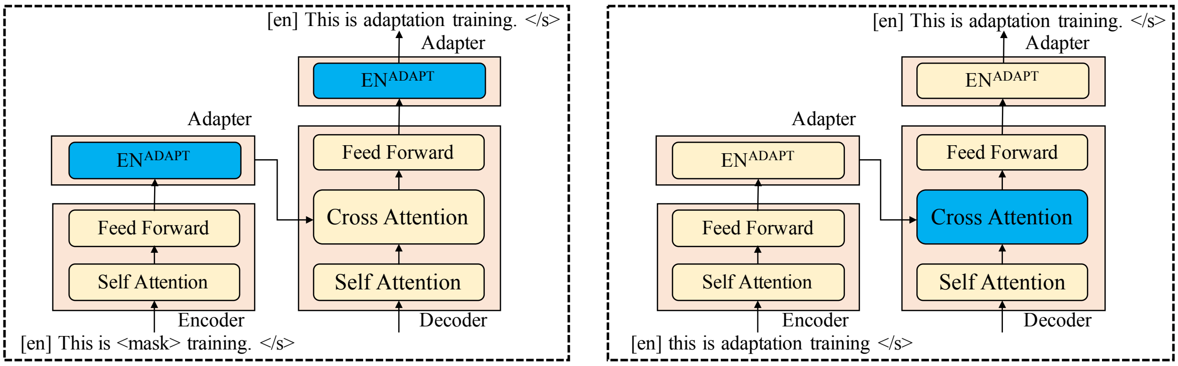
$$L_{\phi_{\mathbf{A}}} = -\Sigma log(T|g(T); \phi_{\mathbf{A}}) \quad (2)$$

[2]The number of monolingual sentences used in mBART-50's pre-training is only 49,446 for Portuguese, for example, versus 36,797,950 for French and 226,457 for Italian.

[3]https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11

[4]Sentences with σ < -0.5 are considered informal while > 1.0 are formal in our experiments.

[5]http://data.statmt.org/news-crawl/



(a) Language adaptation training with monolingual data

(b) Task adaptation training with English parallel data

Figure 1: Overview of adaptation training. In 1(a), the feed-forward network of each transformer layer or the inserted adapter layer is trained with monolingual data to adapt to the target language. In 1(b), the cross-attention of mBART is trained with auxiliary English parallel data to adapt to the TST task.

where $\phi_{\boldsymbol{D}}$ are the parameters of adaptation module A, T is a sentence in target language and g is the noise function that masks 30% of the words in the sentence. Each language has its own separate adaptation module. During language adaptation training, the parameters of the adaptation module are updated while the other parameters stay frozen.

3.2 Task Adaptation

As shown in Figure 1(b), after training the language adaptation module we fine-tune the model on the auxiliary English parallel data with the aim of making the model adapt to the specific domain of formality and style transfer task. Following Cooper Stickland et al. (2021), we only update the parameters of the decoder's cross-attention (i.e. task adaptation module) while the other parameters are fixed, thus limiting computational cost and catastrophic forgetting.

Multilingual TST process For the language adaptation modules we have two settings: (i) adaptation modules $\mathbf{A}_s^E$ on the encoder come from the model trained with source style texts, and modules $\mathbf{A}_t^D$ on the decoder come from the model trained with target style texts (M2.X, Table 1); (ii) both $\mathbf{A}^E$ and $\mathbf{A}^D$ are from a model trained with generic texts (M3.X), so there are no source and target styles for the adaptation modules. For the task adaptation modules, we also have two settings: (i) the module is from the English model (X + EN model's cross-attn); (ii) fine-tuning the model of the target language with English parallel data (X + EN data).

4 Experiments

All experiments are implemented atop Transformers (Wolf et al., 2020) using mBART-large-50 (Tang et al., 2020). We train the model using the Adam optimiser (Kingma and Ba, 2015) with learning rate $1e^{-5}$ for all experiments. We train the language adaptation modules with generic texts separately for each language for 200k training steps with batch size 32, accumulating gradients over 8 update steps, and set it to 1 for other training.[6]

Evaluation Following previous work (Luo et al., 2019; Sancheti et al., 2020; Lai et al., 2021a), we assess style strength and content preservation. We fine-tune mBERT (Devlin et al., 2019) with Briakou et al. (2021b)'s pseudo-parallel corpora to evaluate the style accuracy of the outputs. We also use a style regressor from Briakou et al. (2021a), which is based on XLM-R (Conneau et al., 2020) and is shown to correlate well with human judgments[7]. We calculate BLEU[8] and COMET (Rei et al., 2020) to assess content preservation. As overall score, following previous work, we compute the harmonic mean (HM) of style accuracy and BLEU.

Systems Based on our data (see Section 2), we have four settings for our systems. **D1**: pseudo-parallel data in the target language via machine translating the English resource; **D2**: non-parallel style data in the target language; **D3**: no style data in the target language; **D4**: no parallel data at all.

[6] Parameter values are based on previous work as well as preliminary experimental evidence.

[7] Detailed results for different classifiers/regressor on the test set are in Appendix A.2.

[8] We use multi-bleu.perl with default settings.

Data	Model	Informal→Formal									Formal→Informal								
		Italian			French			Portuguese			Italian			French			Portuguese		
		BLEU	ACC	HM	BLEU	ACC	HM	BLEU	ACC	HM	BLEU	ACC	HM	BLEU	ACC	HM	BLEU	ACC	HM
D1	Multi-Task Tag-Style (Briakou et al., 2021b)	0.426	0.727	0.537	0.480	0.742	0.583	0.550	0.782	0.645	-	-	-	-	-	-	-	-	-
	M1.1: pseudo-parallel data	0.459	**0.856**	**0.598**	**0.530**	0.829	0.647	0.524	**0.852**	0.649	0.177	0.311	0.226	**0.195**	0.377	0.257	0.225	0.306	**0.259**
	M1.2: M1.1 + EN parallel data	**0.461**	0.841	0.596	0.525	**0.863**	**0.653**	**0.553**	0.809	**0.657**	**0.178**	0.315	0.227	0.194	**0.458**	**0.273**	0.219	**0.313**	0.258
	M1.3: all data (one model)	**0.461**	0.850	**0.598**	0.515	0.851	0.642	0.537	0.803	0.644	0.175	**0.368**	**0.237**	0.191	0.439	0.266	**0.229**	0.292	0.257
D2	DLSM (Briakou et al., 2021b)	0.124	-0.223	0.159	0.180	0.152	0.165	0.185	0.191	0.188	-	-	-	-	-	-	-	-	-
	M2.1: IBT training + EN data	0.460	0.510	0.484	0.500	0.487	0.492	0.491	0.428	0.457	0.168	0.420	0.240	0.196	0.235	0.214	**0.237**	0.083	0.123
	M2.2: ADAPT + EN model's cross-attn	0.467	0.637	0.539	0.516	0.627	0.566	0.499	-0.365	0.422	0.175	0.672	0.278	**0.212**	**0.627**	**0.317**	**0.237**	0.471	**0.315**
	M2.3: ADAPT + EN data	**0.476**	**0.731**	**0.577**	**0.519**	**0.702**	**0.597**	**0.526**	**0.509**	**0.517**	**0.180**	**0.719**	**0.288**	0.209	0.567	0.305	0.169	**0.534**	0.257
D3	M3.1: EN data	**0.485**	0.670	**0.563**	**0.553**	0.727	0.628	0.039	**0.890**	0.074	**0.186**	**0.767**	**0.299**	**0.216**	**0.692**	**0.329**	0.020	0.403	0.038
	M3.2: ADAPT + EN model's cross-attn	0.480	0.672	0.560	0.545	**0.749**	**0.631**	**0.547**	0.559	**0.553**	0.179	0.421	0.251	0.209	0.685	0.320	0.175	**0.560**	0.267
	M3.3: ADAPT + EN data	0.423	**0.735**	0.537	0.547	0.722	0.622	0.423	0.508	0.462	0.169	0.733	0.275	0.205	0.584	0.303	**0.189**	0.505	**0.275**
D4	Rule-based (Briakou et al., 2021b)	**0.438**	**0.268**	**0.333**	**0.472**	**0.208**	**0.289**	**0.535**	**0.448**	**0.488**	-	-	-	-	-	-	-	-	-
	M4.1: original mBART	0.380	0.103	0.162	0.425	0.080	0.135	0.128	0.200	0.156	0.160	**0.146**	**0.153**	0.189	**0.189**	**0.189**	0.080	**0.657**	**0.143**
	M4.2: ADAPT (generic data)	0.401	0.092	0.150	0.444	0.075	0.128	0.463	0.223	0.301	**0.164**	0.130	0.145	**0.194**	0.170	0.181	**0.237**	0.082	0.122

Table 1: Results for multilingual formality transfer. Notes: (i) for formal-to-informal there are four different source sentences and a human reference only, so for each instance scores are averaged; (iii) bold numbers denote best systems for each block, and underlined denote the best score for each transfer direction for each language.

The first three settings all contain gold English parallel data.

Results Table 1[9] shows the results for both transfer directions for our models. We also include the models from Briakou et al. (2021b) for comparison (they only model the informal-to-formal direction).

Results in block **D1** show that fine-tuning mBART with pseudo-parallel data (M1.1) yields the best performance in the informal-to-formal direction. The formal-to-informal results, instead, are rather poor and on Italian even worse than IBT-based models (M2.2). This could be due to this direction being harder in general, since there is more variation in informal texts[10], but it could also be made worse by the bad quality of the informal counterpart in the translated pairs. Indeed, work in machine translation has shown that low-quality synthetic data is more problematic in the target side (the case of our formal-to-informal direction) than in the source side (informal-to-formal direction) (Bogoychev and Sennrich, 2019).

In **D2**, we see that our proposed adaptation approaches outperform IBT-based models on both transfer directions. The results of fine-tuning the target language's model with English parallel data are generally better than inserting the EN model's cross-attention module into the target language's model. This suggests that the former can better transfer task and domain knowledge.

In **D3**, the large amounts of generic texts yield more improvement in informal-to-formal rather than formal-to-informal. This could be due to generic texts being more formal than informal. The performance improvement on Portuguese is particularly noticeable (compare M3.1 trained with EN data only with other M3.X models), and mostly due to this language being less represented than the others in mBART. Interestingly, the performance of task adaptation strategies is reversed compared to D2: it is here better to adapt cross attention in the English model rather than fine-tune the target language model directly. Future work will need to investigate how using different data sources for language adaptation (D2, style-specific vs D3, generic) interacts with task adaptation strategies.

Results for **D4** show that language adaptation training helps with content preservation, especially for Portuguese, confirming this curbs the problem of language underrepresentation in pre-training. However, low performance on style accuracy shows that task-specific data is necessary, even if it comes from a different language.

5 Conclusions

Fine-tuning a pre-trained multilingual model with machine translated training data yields state-of-the-art results for transferring informal to formal text. The results for the formal-to-informal direction are considerably worse—the task is more difficult, and the quality of translated informal text is lower.

We have also proposed two adaptation training strategies that can be applied in a cross-lingual transfer strategy or in complete absence of task-specific data, though in the latter case results are poor. These strategies target language and task adaptation, and can be combined to adapt mBART for multilingual formality transfer. The adaptation strategies with auxiliary parallel data from a different language are effective, yielding state-of-the-art results for informal-to-formal and outperforming more classic IBT-based approaches without task-specific parallel data in the target language.

[9] Complete results are in Appendix A.3.

[10] We have observed the same pattern even in optimal conditions for this task, i.e., gold aligned data (English) and a monolingual model (BART). Results are shown in Appendix A.1.

Ethics Statement

All work that automatically generates and/or alters natural text could unfortunately be used maliciously. While we cannot fully prevent such uses once our models are made public, we do hope that writing about risks explicitly and also raising awareness of this possibility in the general public are ways to contain the effects of potential harmful uses. We are open to any discussion and suggestions to minimise such risks.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021a. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021b. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021a. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021b. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. 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.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Sudha Rao and Joel 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, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560.

Asa Cooper Stickland, Alexandre Bérard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for nmt: Decoupling language and domain information with adapters. *arXiv preprint, arXiv: 2110.09574*.

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

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3801–3807.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

A Appendices for ACL Rolling Review:

This appendices include: 1) Results for BART and mBART on English data (A.1); 2) Results for style classifiers/regressor (A.2); 3) Detailed results for multilingual formality transfer (A.3).

.

A.1 Results for BART and mBART on English data

We fine-tune BART (Lewis et al., 2020) and mBART-50 (Tang et al., 2020) with English parallel data specifically developed for formality transfer in English (GYAFC). The performance of BART and English data can be seen as a sort of upperbound, as these are best conditions (monolingual model, and gold parallel data). The drop we see using mBART is rather small, suggesting mBART is a viable option. We also see that formal to informal is much harder than viceversa, probably due to high variability in informal formulations. Table A.1 shows the results for both models.

MODEL	DIRECTION	COMET	BLEU	REG.	ACC	HM
BART	Inf.→For.	0.544	0.795	-0.527	0.928	0.856
	For.→Inf.	0.170	0.436	-1.143	0.683	0.532
mBART	Inf.→For.	0.512	0.779	-0.531	0.916	0.842
	For.→Inf.	0.151	0.422	-1.031	0.591	0.492

Table A.1: Results of BART and mBART on English data. Note that REG. indicates the score of the style regressor (the higher is better in Inf.→For.(informal-to-formal), lower is better in For.→Inf. (formal-to-informal);.

A.2 Results for style classifiers/regressor

We compare three different style classifiers and one regressor: (i) TextCNN (Kim, 2014) trained with pseudo-parallel data; (ii) mBERT (Devlin et al., 2019) trained with pseudo-parallel data and English data respectively; and (iii) a XLM-R (Conneau et al., 2020) based style regressor from Briakou et al. (2021a), which is trained with formality rating data in English.

MODEL	TRAINING DATA	ITALIAN				FRENCH				PORTUGUESE			
		ACC	Precision	Recall	F1	ACC	Precision	Recall	F1	ACC	Precision	Recall	F1
TextCNN	Pseudo data	0.865	0.885	0.839	0.861	0.838	0.876	0.787	.829	0.799	0.793	0.809	0.801
mBERT	Pseudo data	**0.898**	0.905	0.890	**0.897**	0.879	**0.918**	0.831	0.872	**0.851**	0.806	0.924	**0.861**
mBERT	English data	0.889	0.856	**0.934**	0.893	**0.896**	0.856	**0.951**	**0.901**	0.839	0.771	**0.964**	0.857
mBERT	All data	0.891	**0.906**	0.872	0.888	0.882	0.911	0.846	0.877	**0.851**	**0.815**	0.909	0.859
XLM-R	Formality ratings	Informal: -1.672		Formal: 0.108		Informal: -1.701		Formal:0.050		Informal:-1.438		Formal: 0.065	

Table A.2: Results for style classifiers/regressor on test set. The data used for evaluation are 1000 sentences from the test set and the corresponding 1000 human references. For informal sentences, the smaller the XLM-R score is better, higher is better for formal sentences.

A.3 Detailed results for multilingual formality transfer

Data	Model	Italian COMET	Italian BLEU	Italian REG.	Italian ACC	Italian HM	French COMET	French BLEU	French REG.	French ACC	French HM	Portuguese COMET	Portuguese BLEU	Portuguese REG.	Portuguese ACC	Portuguese HM
	Transfer Direction: Informal→Formal															
D1	Translate Train Tag (Briakou et al., 2021b)	-0.059	0.426	-0.705	0.735	0.539	-0.164	0.451	-0.586	0.696	0.547	0.194	0.524	-0.636	0.755	0.619
D1	+ Back-Tranlated Data (Briakou et al., 2021b)	0.026	0.430	-0.933	0.556	0.485	0.004	0.491	-0.898	0.485	0.488	0.301	0.546	-0.875	0.627	0.584
D1	Multi-Task Tag-Style (Briakou et al., 2021b)	-0.021	0.426	-0.698	0.727	0.537	-0.062	0.480	-0.501	0.742	0.583	0.266	0.550	-0.578	0.782	0.645
D1	M1.1: pseudo-parallel data	0.143	0.459	-0.426	**0.856**	**0.598**	0.124	**0.530**	-0.305	0.829	0.647	0.297	0.524	**-0.334**	**0.852**	0.649
D1	M1.2: M1.1 + EN parallel data	**0.147**	**0.461**	-0.442	0.841	0.596	**0.130**	0.525	-0.275	**0.863**	**0.653**	**0.331**	**0.553**	-0.395	0.809	**0.657**
D1	M1.3: all data (one model)	0.137	**0.461**	**-0.409**	0.850	**0.598**	0.127	0.515	**-0.267**	0.851	0.642	0.309	0.537	-0.367	0.803	0.644
D2	DLSM (Briakou et al., 2021b)	-1.332	0.124	-2.141	0.223	0.159	-1.267	0.180	-2.021	0.152	0.165	-1.131	0.185	-2.078	0.191	0.188
D2	M2.1: IBT training	0.057	0.420	-1.351	0.240	0.305	-0.019	0.465	-1.303	0.219	0.298	0.233	0.487	-1.074	0.411	0.446
D2	M2.2: M2.1 + EN data	0.105	0.460	-0.867	0.510	0.484	0.036	0.500	-0.814	0.487	0.492	0.236	0.491	-1.040	0.428	0.457
D2	M2.3: ADAPT + EN model's cross-attn	**0.139**	0.467	-0.684	0.637	0.539	0.066	0.516	-0.613	0.627	0.566	0.288	0.499	-1.143	0.365	0.422
D2	M2.4: ADAPT + EN data	0.131	**0.476**	**-0.537**	**0.731**	**0.577**	**0.074**	**0.519**	**-0.572**	**0.702**	**0.597**	**0.291**	**0.526**	**-0.922**	**0.509**	**0.517**
D3	M3.1: EN data	**0.134**	**0.485**	-0.590	0.670	**0.563**	**0.102**	**0.553**	-0.591	0.727	0.628	-1.673	0.039	**-0.550**	**0.890**	0.074
D3	M3.2: ADAPT + EN model's cross-attn	0.130	0.480	-0.588	0.672	0.560	0.091	0.545	**-0.446**	**0.749**	**0.631**	**0.302**	**0.547**	-0.859	0.559	**0.553**
D3	M3.3: ADAPT + EN data	-0.107	0.423	**-0.579**	**0.735**	0.537	0.101	0.547	-0.488	0.722	0.622	-0.260	0.423	-1.112	0.508	0.462
D4	Round-trip MT (Briakou et al., 2021b)	-0.053	0.346	**-1.026**	**0.354**	**0.350**	-0.065	0.416	**-0.748**	**0.406**	**0.411**	0.213	0.430	**-0.661**	**0.601**	**0.501**
D4	Rule-based (Briakou et al., 2021b)	**0.071**	**0.438**	-1.167	0.268	0.333	**-0.013**	**0.472**	-1.236	0.208	0.289	**0.291**	**0.535**	-1.081	0.448	0.488
D4	M4.1: original mBART	-0.067	0.380	-1.672	0.103	0.162	-0.106	0.425	-1.709	0.080	0.135	-1.444	0.128	-1.870	0.200	0.156
D4	M4.3: ADAPT (generic data)	0.033	0.401	-1.675	0.092	0.150	-0.033	0.444	-1.700	0.075	0.128	0.230	0.463	-1.438	0.223	0.301
	Transfer Direction: Formal→Informal															
D1	M1.1: pseudo-parallel data	**0.298**	0.177	-0.225	0.311	0.226	**0.239**	**0.195**	-0.188	0.377	0.257	0.388	0.225	-0.273	0.306	**0.259**
D1	M1.2: M1.1 + EN parallel data	0.278	**0.178**	-0.228	0.315	0.227	0.215	0.194	**-0.304**	**0.458**	**0.273**	0.373	0.219	**-0.282**	**0.313**	0.258
D1	M1.3: all data (one model)	0.283	0.175	**-0.287**	**0.368**	**0.237**	0.207	0.191	-0.301	0.439	0.266	**0.407**	**0.229**	-0.241	0.292	0.257
D2	M2.1: IBT training	0.335	0.166	-0.082	0.338	0.223	0.272	0.195	0.037	0.194	0.194	0.467	**0.237**	0.042	0.084	0.124
D2	M2.2: M2.1 + EN data	**0.337**	0.168	-0.174	0.420	0.240	**0.274**	0.196	-0.016	0.235	0.214	**0.471**	**0.237**	0.045	0.083	0.123
D2	M2.3: ADAPT + EN model's cross-attn	0.176	0.175	**-0.631**	0.672	0.278	0.226	**0.212**	**-0.464**	**0.627**	**0.317**	0.441	**0.237**	-0.343	0.471	**0.315**
D2	M2.4: ADAPT + EN data	0.279	**0.180**	-0.582	**0.719**	**0.288**	0.232	0.209	-0.444	0.567	0.305	-0.022	0.169	**-0.520**	**0.534**	0.257
D3	M3.1: EN data	0.289	**0.186**	-0.646	**0.767**	**0.299**	**0.244**	**0.216**	-0.566	**0.692**	**0.329**	-1.695	0.020	**-1.225**	0.403	0.038
D3	M3.2: ADAPT + EN model's cross-attn	**0.300**	0.179	-0.285	0.421	0.251	0.221	0.209	**-0.594**	0.685	0.320	**0.367**	0.175	-0.449	**0.560**	0.267
D3	M3.3: ADAPT + EN data	0.100	0.169	**-0.744**	0.733	0.275	0.220	0.205	-0.447	0.584	0.303	0.130	**0.189**	-0.586	0.505	**0.275**
D4	M4.1: original mBART	0.260	0.160	**0.076**	**0.146**	**0.153**	0.204	0.189	**0.031**	**0.189**	**0.189**	-1.363	0.080	**-1.406**	**0.657**	**0.143**
D4	M4.2: ADAPT (generic data)	**0.317**	**0.164**	0.084	0.130	0.145	**0.268**	**0.194**	0.052	0.170	0.181	**0.475**	**0.237**	0.047	0.082	0.122

Table A.3: Results for multilingual formality transfer. Notes: (i) REG. indicates the score of the style regressor (the higher is better in informal-to-formal, lower is better in formal-to-informal; (ii) for formal-to-informal there are four different source sentences and a human reference only, so for each instance scores are averaged; (iii) bold numbers denote best systems for each block, and underlined indicate the best score for each transfer direction.