Modified Iterative Matching and Translation Approach for Formality Style Transfer in a Low-Resource Setting

Kenneth Uriel Loquinte, Charibeth Cheng
De La Salle University Manila
Taft Ave., Malate, 1004 Manila, Philippines
{kenneth_uriel_loquinte, charibeth.cheng}@dlsu.edu.ph

Abstract

There is limited research on text style transfer (TST) for non-English languages due to the scarcity of essential resources like parallel corpora. This paper explores a non-parallel approach to creating pseudo-parallel corpora for training a formality style classifier in Filipino. Specifically, it adapts the Iterative Matching and Translation (IMaT) method. This involves aligning texts from different corpora, training a model for style transfer, and refining the dataset iteratively. Modifications include using margin-based similarity scoring, training a pre-trained multilingual model, and applying data augmentation. Results show these modifications enhance formality style transfer performance compared to the original IMaT implementation. However, further improvements to the matching algorithm and dataset refinement are necessary for broader applicability and generalization.

1 Introduction

In natural language generation (NLG) applications, there is a necessity to make systems more user-centered; this entails that these systems should understand and communicate with nuances of language. With that said, it is important to consider style when modeling a language (Jin et al., 2022). This is what the field of text style transfer (TST) tackles, wherein the style of a given text is modified while its content is preserved.

The choice of approach in TST heavily relies on data accessibility. Supervised learning is common when parallel datasets are available (Rao and Tetreault, 2018; Zhang et al., 2020), but these are not always available, and creating one for each possible TST subtask is unsustainable. Therefore, many works only assume access to non-parallel corpora and apply techniques such as disentanglement (John et al., 2019), prototype editing (Li et al., 2018; Madaan et al., 2020), and pseudo-parallel corpus construction (Jin et al., 2019).

The exploration of TST in non-English languages is limited due to the lack of resources (Briakou et al., 2021b). This work specifically tackles the formality style transfer (FST) subtask in Filipino, a low-resource language. Although there exists work on adjacent NLP tasks in Filipino such as grammar correction (Go et al., 2017) and spell checking (Octaviano and Borra, 2017), no work has specifically explored FST techniques.

This work adapts the Iterative Matching and Translation (IMaT) approach (Jin et al., 2019) and explores its applicability in a low-resource setting. With that said, we make the following contributions:

- We explore using margin-based similarity scoring, training a multilingual language model, and applying data augmentation in building pseudo-parallel pairs via IMaT. We assess their benefits in a low-resource setting using the three common TST metrics: style accuracy, meaning preservation, and fluency.
- We provide a baseline work for Filipino FST, which can be used as a reference for future efforts in Filipino. We also contribute to the limited TST work in non-English languages.

Results show that these modifications are helpful in improving FST performance, although further improvements are necessary to make a more general solution in terms of both language and style. Like ours, works in non-English FST generally face the same issue of resource availability, but these efforts are important in making more robust conclusions about the current state of TST techniques.

2 Related Work

FST is one of the TST subtasks that has gained considerable attention, and it benefits from the availability of parallel data such as Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao

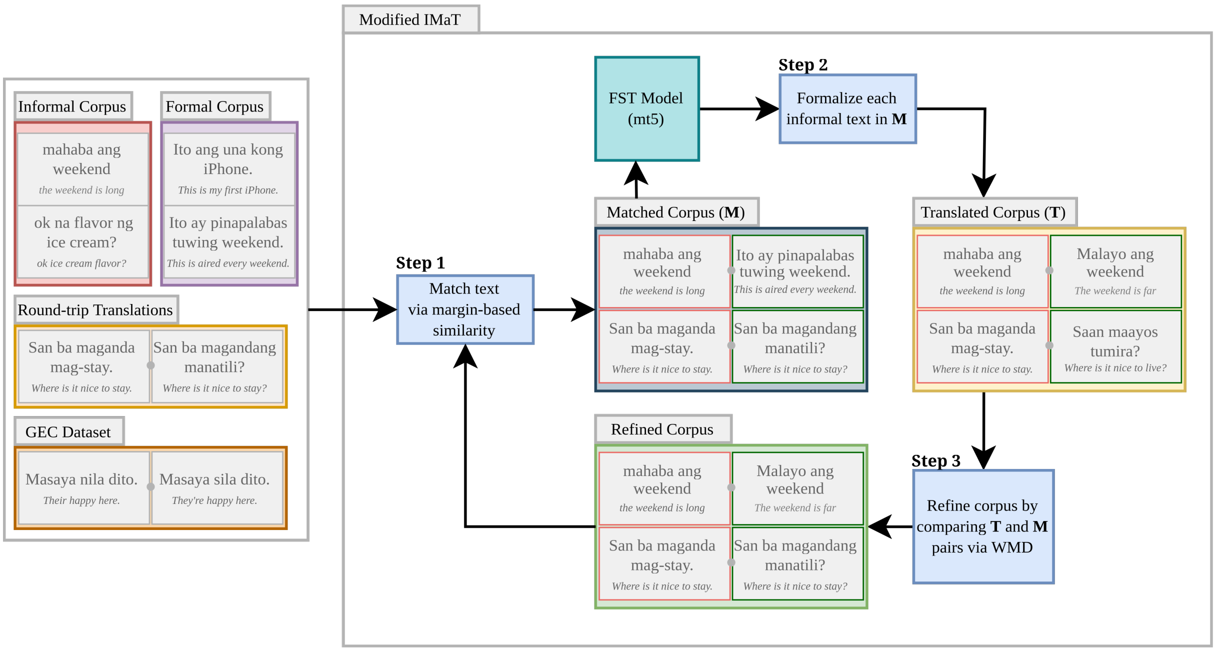

Figure 1: The input to the modified IMaT pipeline are two distinct informal and formal style corpora which are augmented with grammar error correction (GEC) and round-trip translation datasets. The steps are described as follows: (1) the first step is to match informal and formal text using margin-based similarity (Artetxe and Schwenk, 2019) resulting in a pseudo-parallel corpus. An mt5 model is trained to perform FST using the matched corpus (**M**). Then, (2) the next step is to formalize each informal text in **M** using the trained model, resulting in a new corpus **T** whose target consist of generated formal text. Finally, (3) the last step is to refine the dataset by comparing the generated targets in **T** to the previous targets in **M** using WMD (Kusner et al., 2015). The IMaT pipeline is iterative and the refined corpus can be used as an input to the next iteration.

and Tetreault, 2018). Several works have utilized GYAFC for experimentation such as Zhang et al. (2020) which used data augmentation techniques for improved FST performance, and Wang et al. (2019) which proposed a rule-based formalization approach on a neural-network-based system. This highlights the importance of a resource like GYAFC in TST.

Even though a formality dataset exists, there are still works that seek non-parallel approaches to FST due to their cost-effectiveness. One method is by constructing a pseudo-parallel corpus such as by using synonym as word replacements (Jain et al., 2019), or by aligning pairs from distinct corpora (Jin et al., 2019). Another technique is a two-stage approach which involves using a model to neutralize style attributes, and then a style-trained model paraphrases the neutral text to the desired style (Krishna et al., 2020).

There have been efforts to study FST in low-resource settings such as building a multilingual formality dataset (Briakou et al., 2021b), and using few-shot translation techniques (Krishna et al., 2022). Our work utilizes the pseudo-parallel corpus construction approach and data augmentation techniques, which are both useful in low-resource settings as it allows us to use available resources in Filipino.

3 Approach

3.1 Iterative Matching and Translation (IMaT)

This work adapts the IMaT algorithm proposed by (Jin et al., 2019), which pairs sentences from two separate style corpora by using a similarity metric. While IMaT has been primarily applied in English, it can be applied in low-resource settings as it is not reliant on language-specific qualities. We localize the pipeline by using Filipino-based embeddings.

The stages of the pipeline are discussed below with a focus on FST, but it is worth noting that the pipeline is style-agnostic. Figure 1 shows an illustration of the system.

Matching Each text from both informal and formal datasets are represented as sentence embeddings using Paraphrase-Fil-MPNet[1], which was

[1] The model was sourced from: meedan/paraphrase-

trained on data from OPUS using the student-teacher approach by Reimers and Gurevych (2020). Informal-formal pairs are created by matching the embeddings of text between the two datasets using pairwise cosine similarity. However, we found that using cosine similarity creates less diverse pairs because the same formal sentence would be paired with multiple informal sentences. Instead, we use margin-based scoring (Artetxe and Schwenk, 2019) which uses the margin between a given sentence's similarity and its nearest neighbors to mitigate the effects of the varied similarity scales.

The pairs that exceed a similarity threshold constitute the pseudo-parallel corpus, which is used to train a seq2seq model in the next stage. In cases where multiple candidate sentences exist, the formal sentence that achieves the highest similarity score was selected. The paired sentences create a matched corpus $M = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$.

For succeeding iterations $i > 0$, the matching process is conducted on the sentences of M_i instead of the distinct style corpora, resulting in a new set of matched pairs N_i. Pairs that get a similarity score less than the threshold are removed to retain the size of the dataset. We compute the Word Mover's Distance (WMD), a metric that quantifies distance based on word embeddings (Kusner et al., 2015) between pairs from M_i and N_i, and the lower scoring pairs become part of the corpus that was passed to the translation stage. We use Filipino FastText word embeddings (Velasco, 2021) to represent the words on each text. This operation is summarized as:

$$M_{ij} = \min(\mathrm{wmd}(x_j, m_j), \mathrm{wmd}(x_j, n_j)) \quad (1)$$

where $(x_j, m_j) \in M_i$ and $(x_j, n_j) \in N_i$.

Translation & Refinement A seq2seq model is trained using the created informal-formal pairs from the matching stage resulting in model F_i at iteration i, whose goal is to convert an informal text to a formal one. The original implementation uses an LSTM encoder-decoder model, but this work uses a pre-trained multilingual model, mt5 (Xue et al., 2021), which is useful in a low-resource setting.

The formality style transfer model F_i is used to generate a transferred sentence t for each informal sentence x from pair $(x, m) \in M_i$, creating a new pair (x, t). All generated pairs form a new pseudo-parallel corpus T_i.

The new pair (x, t) is compared with the existing pair (x, m) using WMD. The pair with a lower WMD score is incorporated into the refined corpus M_{i+1}, which is used in the next iteration. This operation is summarized by:

$$M_{i+1}j = \min(\mathrm{wmd}(x_j, m_j), \mathrm{wmd}(x_j, t_j)) \quad (2)$$

where $(x_j, m_j) \in M_i$ and $(x_j, t_j) \in T_i$.

The pipeline is iterative and the resulting refined dataset can be fed to the next iteration's matching stage. In this work, the end condition is based on update rate and number of iterations.

3.2 Data Augmentation

Using augmented data related to the formality task is useful in improving performance, especially when dealing with pseudo-parallel pairs whose quality cannot be ensured. We augment the original pairs with round-trip translation and grammar error correction (GEC) pairs.

Firstly, we use round-trip translation pairs which have been found useful by other works (Briakou et al., 2021b; Zhang et al., 2020) where the motivation is based on the observation that machine translation systems often produce formal text.

Secondly, we use GEC data following a multi-task transfer approach (Zhang et al., 2020). This approach proposes the use of available resources for tasks related to formality, resulting in the style transfer model also learning said task. GEC is suitable to FST because grammaticality is an important part of formality.

Zhang et al. (2020) found that pre-training the model with the augmented data achieves better performance than using for simultaneous fine-tuning with the original pairs. This occurs because the augmented pairs are usually noisier than the original pairs. However, in this work where the original pairs are pseudo-parallel pairs, this cannot be assumed, and doing so can be limiting because the augmented datasets can have better quality pairs.

Given that, simultaneous fine-tuning is done instead of pre-training. To maintain the priority on the original pairs, we set the number of augmented training pairs to around half the number of pseudo-parallel training pairs.

filipino-mpnet-base-v2

4 Experiments

4.1 Datasets

There are no datasets in Filipino that have been specifically curated for either formal or informal text. Hence, some assumptions were made to leverage existing resources. The datasets used to represent informal and formal styles are described below.

PEx-Conversations For the informal style, we use the PEx-Conversations dataset (Co et al., 2022), which comprises ~2.4M comments across ~45k threads from the Philippine Exchange online forum. This dataset was chosen based on the assumption that discussions on these kinds of platforms exhibit a more casual nature, resulting in a lower level of formality.

WikiText-TL-39 For the formal style, we use the WikiText-TL-39 dataset (Cruz and Cheng, 2019), which comprises ~2M lines of text from Tagalog Wikipedia articles. It is assumed that these articles were written with a certain level of formality, as they were written in accordance with a style guide.

The sentences from the train, validation, and test splits were collated for each dataset. PEx-Conversations was balanced by downsampling the subforum categories based on the smallest category via random selection. We applied the following pre-processing steps to both datasets, which are partially based on Briakou et al. (2021b) and Rao and Tetreault (2018): (1) remove sentences with more than 25 words, or with less than 5 words, (2) normalize punctuations in the text[2], (3) remove non-Tagalog sentences[3], and (4) remove article titles for Wikitext-TL-39.

Next, we describe the datasets for data augmentation.

RT-Fil The first dataset for augmentation is RT-Fil, which consists of 20k round-trip translation pairs. The source texts were taken from PEx-Conversation texts that were filtered out by the IMaT matching stage (i.e., informal text from pairs with similarity scores that are below the set threshold). By doing so, we prevent duplicates with the pseudo-parallel pairs. The translation was done using the Google Translate API with English as the pivot language, and Filipino as the target language.

Balarila The other dataset for augmentation is the Balarila dataset (Ponce et al., 2023), which comprises ~906k pairs that cover grammatical errors (morphological and spelling errors). The Balarila dataset was downsampled to match the number of RT-Fil pairs by randomly sampling a uniform percentage from each transformation / error category from the dataset.

4.2 Implementation

We created a baseline model (**BASE**), which follows the original IMaT implementation, but with localized embedding representations. We used a cosine similarity threshold of 0.75 for filtering, and utilized a 2-layer LSTM encoder-decoder model with attention as the translation model. At each iteration, the model was trained for 10 epochs with a batch size of 16 and a learning rate of 1e-4.

Next, as discussed in Section 3.1, we used a margin-based similarity score for matching with a threshold of 1.05, and used mt5-small[4] as the translation model. We trained a model with PEx-Conversations and Wikitext-TL-39 pseudo-parallel pairs only (**MT5-PW**), then we trained another model with the same pairs augmented with RT-Fil and Balarila (**MT5-AUG-PW**). At each iteration, both models were trained for 5 epochs with a batch size of 64 and a constant learning of 1e-3, following the mt5 fine-tuning setup (Xue et al., 2021).

For generation sampling during the translation stage and testing, we used a top-k of 50 and a temperature of 0.7. Furthermore, for all experiments, the pipeline stopped after 5 iterations, or when the update rate during the refinement stage was less than 5%.

4.3 Evaluation Metrics

To evaluate the model, we employ widely-used automatic metrics in FST.

Formality This work follows the zero-shot approach that was detailed by Krishna et al. (2022). Specifically, we fine-tuned XLM-RoBERTa-Base model for a regression task using the PT16 dataset (Pavlick and Tetreault, 2016). The dataset contains sentences sourced from various text domains, where each sentence was manually annotated with a formality score on a Likert scale ranging from -3 to 3. The fine-tuned model was applied zero-shot

[2] Normalization was done using a Python wrapper of the Moses toolkit

[3] Language detection was done using the langdetect package

[4] From google/mt5-small at Huggingface

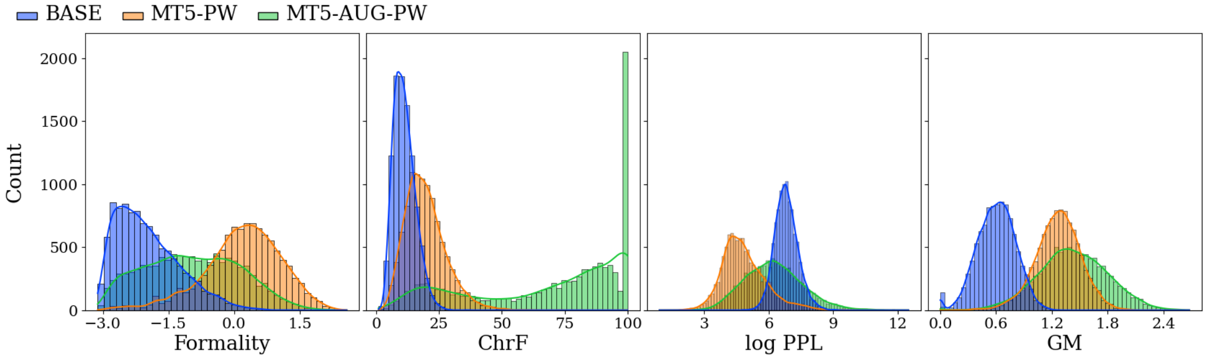

Figure 2: Distribution of scores of BASE, MT5-PW, and MT5-AUG-PW on each evaluation metric.

Model	**Formality**	**ChrF**	**PPL**	**GM**
BASE	-2.11 (-1.98)	10.65 (11.2)	887.81 (1068.25)	0.61 (0.60)
MT5-PW	0.26 (0.19)	18.55 (19.6)	111.04 (262.62)	1.27 (1.27)
MT5-AUG-PW	-0.98 (-0.96)	78.46 (67.98)	481.32 (1503.26)	1.42 (1.42)
Dataset				
INITIAL	0.23 (0.03)	20.89 (37.85)	159.06 (817.80)	1.30 (1.40)
REFINED	-0.95 (-0.90)	83.10 (73.85)	431.03 (1636.94)	1.49 (1.53)

Table 1: The median (and mean in parentheses) evaluation scores on a test set of ~10k samples for BASE, MT5-PW, and MT5-AUG-PW. The median is highlighted instead of the mean due to outliers, especially for perplexity. Formality ranges from -3 to 3, while chrF ranges from 0 to 100. PPL is perplexity, and GM is geometric mean. In the bottom, the same evaluation scores are shown for the training dataset of MT5-AUG-PW before and after pipeline refinement.

on the generated Filipino sentences to determine their formality scores.

Meaning Preservation This work uses the character n-gram F-score (chrF) (Popović, 2015) between source text and output text to compute meaning preservation. Although the BLEU score is a popular metric that has been used in several TST works (Rao and Tetreault, 2018; Li et al., 2018; Jin et al., 2019; Briakou et al., 2021b), chrF has been shown to correlate better with human evaluations (Briakou et al., 2021a). Furthermore, it is beneficial for a morphologically-rich language like Filipino because it assesses similarity on a character level.

Fluency We measure fluency by calculating the perplexity (PPL) of each output sentence through GPT2-Tagalog (Cruz et al., 2020). The model was trained on Philippine news articles and Wikitext-TL-39, both of which contain formal text. Aside from the model being one of the few available causal language models in Filipino, the expected output of the FST model (formal text) fits the domain of GPT2-Tagalog's training corpus.

Overall Score We use geometric mean (GM) as a measure of performance across the three TST metrics following previous works (Krishna et al., 2020; Yi et al., 2021). We compute the geometric mean of formality, chrF, and $\frac{1}{logPPL}$. The formality score is shifted from a $(-3, 3) \rightarrow (0, 1)$ range to be compatible with GM.

4.4 Results

Table 1 shows that using margin-based similarity scoring for matching and mt5 for translation (MT5-PW and MT5-AUG-PW) lead to better performance on all TST aspects compared to the original IMaT implementation (BASE). This indicates that an improved matching algorithm and the use of a multilingual model can boost style transfer performance when working with the pseudo-parallel corpus approach.

Furthermore, the increase in chrF from MT5-PW to MT5-AUG-PW (18.55 → 78.46) shows that augmenting the original pairs with RT-Fil and Balarila pairs greatly improves meaning preservation in the text. This is not surprising because both of these datasets contain pairs whose source and target texts have similar words, especially with Balarila whose corrections are very localized. However, this also means that the model has a tendency to output an

INFORMAL	Si johnskie manlilibre sakin ng xmen hahahaha
	johnskie will buy me xmen hahahaha
BASE	"...ileasisiraan ka na manakit ka ka, hindi ako!.. na nanalo..........!..
	"...ileasisiraan you to hurt you you, not me!.. who won..........!..
MT5-PW	Si John Lennon ay isang mang-aawit.
	John Lennon is a singer.
MT5-AUG-PW	Si Johnski manlilibre sa akin ng Xmen hahahaha
	Johnski will buy me Xmen hahahaha
INFORMAL	That is tumataginting ng 28k + +.
	That is a glowing 28k + +.
BASE	Si Ana ay maaaring tumukoy sa:. sumusunod:. sa katapusan at mga.......... ng mga
	Ana may refer to:. following:. at the end and those.......... of those
MT5-PW	Ang halaga ng pangungupahan ng isang sasaeng halaga ay nagkakahalagang P30.00.
	The renting cost of a saseang amount is P30.00.
MT5-AUG-PW	Ang mga ito ay tinatayang 28k.
	They are estimated at 28k.
INFORMAL	Di ko lang alam kung anung zodiac nila.
	I just don't know what their zodiac is.
BASE	Sa isang paraan ng balat, ang mga tao, dapat na.. sa ng isang ng.........
	In a way of skin, the people, should be.. in of an of.........
MT5-PW	Hindi alam kung paano ito nakilala.
	It is not known how this was identified.
MT5-AUG-PW	Hindi ko alam kung ano ang zodiac nila?
	I don't know what their zodiac is?

Table 2: Example outputs of BASE, MT5-PW, and MT5-AUG-PW on informal texts from the test set. The accompanying text, rendered in gray in each row, are the translations of the preceding Filipino text.

INFORMAL	Nastranded sa edsa kagabi / kahapon dahil sa lakas ng ulan?
(RT-Fil)	*Got stranded at edsa last night / yesterday because of the rain's intensity?*
INITIAL	Stranded sa Edsa kagabi / kahapon dahil sa malakas na ulan?
	Stranded at Edsa last night / yesterday because of the intense rain?
REFINED	Nastranded sa edsa kagabi / kahapon dahil sa lakas ng ulan?
	Got stranded at edsa last night / yesterday because of the rain's intensity?
INFORMAL	Wala ako plano mag migrate sa netherlands, hehe.
(PEx-WikiTL)	*I don't have a plan to migrate to netherlands, hehe.*
INITIAL	Sa huli, siya ay hindi ligtas sa Netherlands.
	In the end, they are not safe in the Netherlands.
REFINED	Wala akong planong mag-migrae sa Netherlands, hehe.
	I don't have a plan to migrae to the Netherlands, hehe.
INFORMAL	Dahil mahirap ang pamilya ni Ralph, hikahos din nila makakain.
(Balarila)	*Because Ralph's family is poor, their also eating poorly.*
INITIAL	Dahil mahirap ang pamilya ni Ralph, hikahos din sila makakain.
	Because Ralph's family is poor, they're also eating poorly.
REFINED	Dahil mahirap ang pamilya ni Ralph, hikahos din nila makakain.
	Because Ralph's family is poor, their also eating poorly.

Table 3: Example of formal target refinements for the training set of MT5-AUG-PW. The sources of the initial informal-formal pairs are indicated in parentheses. The accompanying text, rendered in gray in each row, are the translations of the preceding Filipino text.

exact copy of the informal text and not do any style transfer at all. Figure 2 shows that more than 2,000 (~20%) outputs from MT5-AUG-PW have perfect chrF values, which means that the model generates exact copies frequently.

With that said, better meaning preservation led to worse style accuracy (0.26 → -0.98) and higher perplexity scores (111.04 → 481.32). Since the target outputs become alike to the informal text, the attributes carry over including low formality and high perplexity. This also highlights the limitation of measuring fluency using perplexity — the informal text would appear fluent to a native speaker, but not to GPT2-Tagalog which was only trained on a formal text domain. Therefore, the language model for calculating perplexity should be able to handle either informal and formal text. Unfortunately, such a model is not always available.

Although MT5-AUG-PW is inferior to MT5-PW in two of the three evaluation metrics, the overall score indicates that the latter has better quality style transfer outputs when viewed holistically. The same trend can be seen in Figure 2. Arguably, a translation can only be a proper translation if the actual message is preserved; Table 2 shows that although MT5-PW can retain keywords or a semblance of the topic, only MT5-AUG-PW is able to convey the actual meaning of the informal text.

Nonetheless, the outputs from both models are a stark contrast to that of BASE. The model generates incomprehensible text, often with repeating tokens; this behavior complements its poor evaluation scores in Table 1. Training an LSTM-based model from scratch at each iteration means that the model is learning the language and the FST task at the same time. In this case where we are working with a pseudo-parallel corpus whose pairs have varying quality, performance issues such as what is displayed by BASE can occur. Therefore, in low-resource settings, pre-trained multilingual models offer better starting points.

As much as meaning preservation is a foundation of a good style transfer, it can also be detrimental to the refinement process. Table 3 shows that there is a tendency for the algorithm to "refine" the dataset with lesser-quality targets. It is expected to occur frequently when training with datasets that encourage copying such as RT-Fil and Balarila because the model is likely to generate candidate targets that are equal to the informal text; hence, the algorithm would select the equal-copy target and would ignore a previous target that might have correct formal changes. To illustrate, the initial target for the Balarila example correctly fixes the wrong use of *nila* (their) to *sila* (they), but the FST model generated a copy of the informal text, thus the algorithm selects that as the refined target due to a perfect WMD score, even though it is a worse target text.

The proponents of IMaT make an assumption that the two corpora used to represent the involved styles already have text with good style accuracy and high fluency. However, as the results show, this does not always hold true for a low-resource setting where available text resources are not guaranteed to properly represent the styles and/or may not contain fluent text. Therefore, using only WMD to refine the pseudo-parallel corpus may be insufficient. As an unsupervised approach, the pipeline would greatly benefit from replacing WMD with a score that considers all three metrics, such as the geometric mean.

It is important to discuss that augmented datasets are not always available, and their suitability is still dependent on the language and style. For instance, a good neural machine translation model may not be available for certain low-resource languages, which hinders efforts in creating good-quality round-trip translation pairs. In the same light, there may not be available task-related data that can be applied to the chosen style — GEC pairs are relevant for formality, but not for other styles. Therefore, it remains necessary to improve the matching algorithm to find better pairs from distinct sources, because relying on augmented datasets is not sustainable.

5 Conclusion

This study demonstrates that adapting the IMaT approach with modifications — such as using a pre-trained multilingual language model, margin-based similarity scoring, and data augmentation — enhances the effectiveness of formality style transfer (FST) in Filipino. These findings emphasize the value of customizable, non-parallel techniques in low-resource settings, which allow for more effective utilization of existing resources.

While data augmentation can temporarily boost dataset quality, reliance on it is not a sustainable long-term solution. Therefore, refining the matching algorithm remains a critical avenue for improvement. Future research should explore semi-supervised learning approaches and incorporate data filtering mechanisms that evaluate style accuracy, meaning preservation, and fluency. This approach could improve the overall quality of pseudo-parallel pairs, making the IMaT framework more robust.

The current pipeline's assumptions about text fluency and style accuracy may not hold true in unsupervised, low-resource settings, where the quality of available text can vary. To address this, optimizing all three key TST metrics — style accuracy, meaning preservation, and fluency — simultaneously might provide a more reliable and comprehensive evaluation of TST quality. Investigating the use of an overall score instead of solely relying on the Word Mover's Distance (WMD) score could make the dataset refinement process more aligned with the objectives of TST.

Although the study followed established evaluation methods, direct comparison with other works is challenging due to variations in implementation, including the use of a Filipino-based model for perplexity. Incorporating human evaluations and analyzing their correlation with automatic metrics would enhance the reliability and validity of the findings.

References

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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.

Denzel Adrian Co, Schuyler Ng, Gabriel Louis Tan, Adrian Paule Ty, Jan Blaise Cruz, and Charibeth Cheng. 2022. Using Synthetic Data to Train a Conversational Response Generation Model in Low Resource Settings. In *2022 International Conference on Asian Language Processing (IALP)*, pages 306–311.

Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating Language Model Finetuning Techniques for Low-resource Languages.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of Fake News Detection via Multitask Transfer Learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2596–2604, Marseille, France. European Language Resources Association.

Matthew Phillip Go, Nicco Nocon, and Allan Borra. 2017. Gramatika: A grammar checker for the low-resourced Filipino language. In *TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 471–475.

Parag Jain, Abhijit Mishra, Amar Prakash Azad, and Karthik Sankaranarayanan. 2019. Unsupervised controllable text formalization. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pages 6554–6561, Honolulu, Hawaii, USA. AAAI Press.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised Text Attribute Transfer via Iterative Matching and Translation. In *EMNLP-IJCNLP 2019*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 957–966. PMLR.

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.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Manolito Octaviano and Allan Borra. 2017. A spell checker for a low-resourced and morphologically rich language. In *TENCON 2017 - 2017 IEEE Region 10 Conference*, pages 1853–1856.

Ellie Pavlick and Joel Tetreault. 2016. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Andre Dominic H. Ponce, Joshue Salvador A. Jadie, Paolo Edni Andryn Espiritu, and Charibeth Cheng. 2023. Balarila: Deep learning for semantic grammar error correction in low-resource settings. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 21–29, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Dan John Velasco. 2021. Filipino Word Embeddings. `https://github.com/danjohnvelasco/Filipino-Word-Embeddings`.

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 3573–3578, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2021. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

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.