Applying Natural Language Processing for Politeness Style Transfer

Ang Jun Ray[1] and Huang Haoyang[1]

Mentor(s): Chieu Hai Leong[2]

[1] Raffles Institution, One Raffles Institution Lane, Singapore 575954

[2] DSO National Laboratories, 12 Science Park Dr, Singapore 118225

ABSTRACT

Our research attempts to improve existing methods to politeness transfer, by harnessing Natural Language Processing in the process. The goal of the model is to transform text from an impolite or neutral style to one of a polite style. We experiment with improving the model's accuracy by experimenting with different datasets, and making use of different technologies. Through various metrics, which we will introduce later in the paper, we can then quantitatively and qualitatively measure our model's performance. Our approach is built on current technology such as the *Tag & Generate Approach* (Madaan, Setlur, Parekh. 2020), and th*e Prompt & Rerank Method* (Suzgun, Melas-Kyriazi, Jurafsky. 2022). Throughout this paper, we shall explore the challenges and aspects of the politeness transfer task in procuring successful politeness transfer.

INTRODUCTION

Social media is widely used for communication, with 61% of the world's population currently using it on a daily basis. However, social media can also be a source of impoliteness, with many people acting less polite online than they would in real life, as supported by a recent survey by VitalSmarts revealing that 88% have felt that people are less polite online than in person. This can lead to conflicts and negative interactions on social media, with some users unfriending, unsubscribing, or blocking others or services due to impoliteness. In a business setting, it is paramount to ensure that messages are polite and formal in order to maintain professionalism. Specifically, in Singapore, there is a common use of colloquial and informal language known as Singlish in daily conversations, text, and social media. Therefore, it may be useful to have a secondary checker to ensure that messages are appropriate and polite even in these contexts.

LITERATURE REVIEW

We begin our approach by introducing the two independent approaches, Tag & Generate and Prompt & Rerank. We attempt to assess both models and improve on them further.

Tag & Generate Approach (Madaan, Setlur, Parekh. 2020)

In our attempt at the politeness transfer task, the first model we will be harnessing is the tag-and-generate approach, which contains 2 critical components, the tagger and the generator. The tagger differentiates these non-parallel samples in a text based on two different styles, which we will refer to as S1 and S2 for simplicity. The tagger is a model to infer a sentence from an initial style structure and sample set, so that this inferred sentence is agnostic to the original style. The generator essentially generates samples of a new set based on the initial set of samples X1, conditioned with respect to a different style, S2. In short, in the case of politeness transfer, the tagger attempts to generate samples based on the "polite" style, while still ensuring that the samples adheres to the other non-politeness related styles that the original samples have. It does this by replacing key words in samples with tokens that will be replaced with words of the "polite" style by the generator.

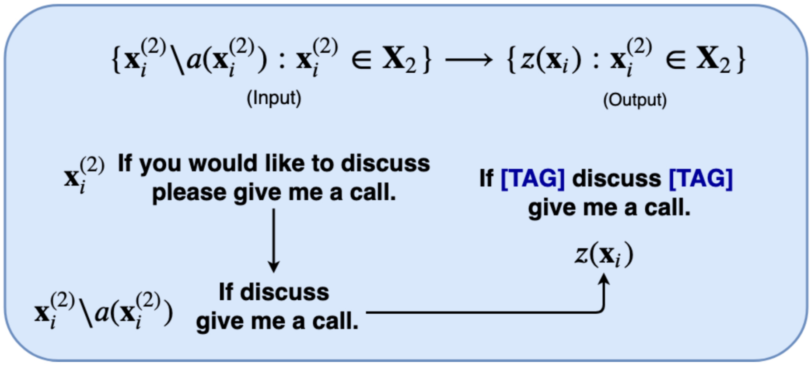
$\{\mathbf{x}_i^{(2)} \backslash a(\mathbf{x}_i^{(2)}) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\} \longrightarrow \{z(\mathbf{x}_i) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$
(Input)
(Output)
$\mathbf{x}_i^{(2)}$
If you would like to discuss please give me a call.
If [TAG] discuss [TAG] give me a call.
$z(\mathbf{x}_i)$
$\mathbf{x}_i^{(2)} \backslash a(\mathbf{x}_i^{(2)})$
If discuss give me a call.

Figure 1: A visualisation of the steps the tagger (Madaan, Setlur, Parekh. 2020)

Words identified to be of undesired sentiment are marked with tokens, and the generator works mainly by generating samples into the target style, by taking the tokens labelled by the tagger and replacing them with stylistically relevant words. These words are inferred from the target style, which in the case of politeness transfer, is the "polite" style. The generator is expected to transform a style agnostic sentence into a style targeted one, through the meaningful replacement of words.

It is also noteworthy that the training process of the tagger and the generator was separated, since the generator is primarily concerned with generating text of the target style given a style agnostic representation. Since the positioning of the generated tags from the tagger greatly affects the style attributes that are used to fill in the new token positionings, multiple tokens are also generated by the tagger. This ensures that both tagger and generator can strategically distinguish different distinctions of style attributes in multiple positions of a sentence.

Prompt & Rerank Method (Suzgun, Melas-Kyriazi, Jurafsky. 2022)

The Prompt-and-Rerank model is a method for transferring the stylistic aspect of politeness in text without changing its main semantic content or meaning. It is useful for considering relevant methods for politeness transfer in natural language processing (NLP).

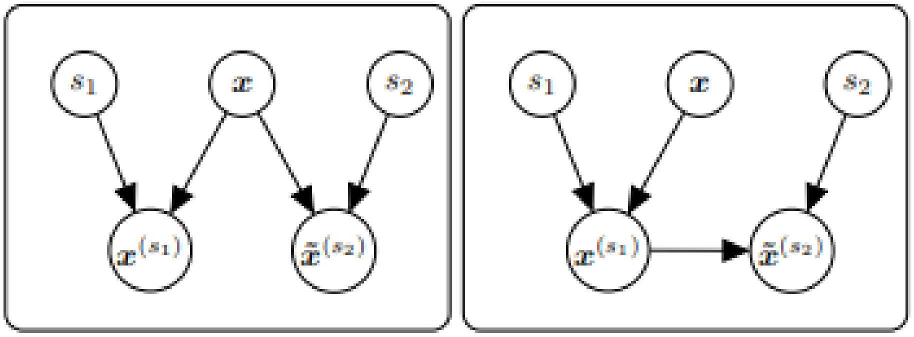

Figure 2: A visualisation of formulating TST. (Suzgun, Melas-Kyriazi, Jurafsky. 2022)

The model involves Textual Style Transfer (TST), which involves the transformation of a text from one style to another. In this case, the goal is to transfer the stylistic aspect of politeness without changing the main semantic content or meaning of the text. In our case, the right model in Figure 2 was used.

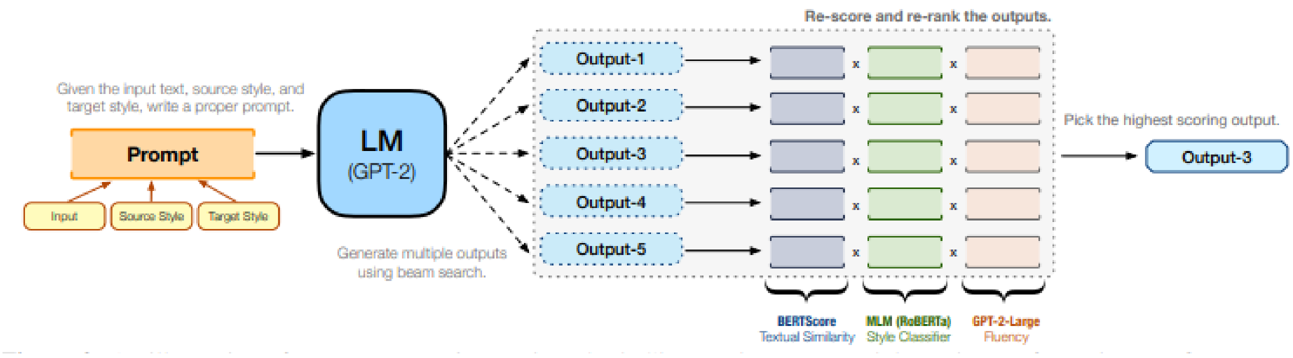
Re-score and re-rank the outputs.
Given the input text, source style, and target style, write a proper prompt.
Prompt
Input
Source Style
Target Style
LM
(GPT-2)
Generate multiple outputs using beam search.
Output-1
Output-2
Output-3
Output-4
Output-5
x
Pick the highest scoring output.
Output-3
BERTScore
Textual Similarity
MLM (RoBERTa)
Style Classifier
GPT-2-Large
Fluency

Figure 3: An illustration of the Prompt-and-Rerank method, given an input text and the style transformation. (Suzgun, Melas-Kyriazi, Jurafsky. 2022)

The Prompt-and-Rerank algorithm can address the task of TST through three main components: prompt construction, candidate output generation, and evaluation or re-ranking of the candidate outputs as mapped in Figure 3.

In the prompt construction phase, a prompt is created based on the input text, source style, and target style. Various text boundary markers, such as delimiter pairs, are also taken into consideration. To create the prompt, a manually-written template is used as a discrete prompt. In the candidate output generation phase, the prompt is fed through a pre-trained language model (LM), such as GPT-2, and k different outputs are generated. These outputs are taken to be candidates for re-ranking. The outputs are generated independently without updating any parameters of the model. In the evaluation or re-ranking phase, the candidate outputs are scored based on three components: textual similarity, target style strength, and fluency. To evaluate the textual similarity term, BERTScore (a measure of the similarity between two texts) is used. To evaluate the target style strength, a masked language model (LM) such as RoBERTa is turned into a style classifier and is used to predict the masked token in a "fill-in-the-blank" cloze template. To evaluate fluency, GPT-2 is used to determine the overall likelihood of each candidate text. The scores are computed by multiplying the results of evaluating each component, and the candidates with the highest scores are selected as the final output.

This approach is model-agnostic, meaning that it can be used with any pre-trained language model. Overall, the Prompt-and-Rerank model is a great method for transferring the stylistic aspect of politeness in text while preserving the main semantic content and meaning.

MATERIALS AND METHODS

Phase 1: Data Preparation

To design our model of the politeness transfer system, we start by cloning the GitHub repository for the Tag & Generate Approach. In this phase, we aim to reproduce results obtained by the authors of the method, thus we get the same training data from the Enron Email corpus (Klimt, Yang. 2004). When preparing parallel data for training, we are able to decide between toggling on and off the unimodal setting. If toggled on, the unimodal setting will focus on one style/mode to prepare the parallel data on, otherwise it could focus on multi styles/modes. Thus, we decided to evaluate the performance of the model when unimodal is specified to be true or false, in order to infer if politeness can be transferred better as an independent style or as multiple styles. Past work was done with unimodal specified to be true, as the dataset has only one stylistic information as is with the politeness transfer task. Moreover, this also ensures that the parallel data is created as per the unimodal style setting. While in theory, setting unimodal to true is what we expect to be ideal, we decided to independently verify this by training the Enron corpus with unimodal set to false as well.

We trained the model after processing the data through BPE (Byte-Pair Encoding). After the training over 50 epochs, the best models are then filtered out according to validation perplexity. We proceeded to extract the test data from the Enron corpus as well and run it through the tagger and generator to generate their outputs.

Phase 2: Adapting Data Types to Fit Desired Context

Phase 2 represented our attempt at improving the current performance of the tag-and-generate models. As our project aimed to improve the model's performance in a social media context, but to maintain its performance in other contexts, we added to the original training and testing dataset. The dataset added is the Reddit Small Corpus with 297,132 utterances (The ConvoKit Developers) as well as a Singlish dataset (Wang, Yang, Zhang. 2010) with 84,177 utterances. While the datasets are small in comparison to the 1,397,119 utterances that the Enron Corpus has, Phase 2 acts as a pilot run before we expand the dataset with the Reddit Full Corpus (with a total of 948,169 subreddits, with each subreddit having utterances) for future work. This dataset is pruned via pre-processing, de-duplicating, and pruning. Filters invalidate certain conversations that are either too short/long (less than 5 characters or more than 2048 characters) or have spurious characters. This ensures the smooth subsequent classification of the data. After this, we classified the dataset using the BERT classifier, and categorised each utterance into their own politeness category (from 0 to 9). These categories are represented with politeness tags, which are in the form of "P_" and the category it is in. The results are then combined with that of the Enron Corpus.

Training of the new datasets were set with unimodal to true due to the Phase 1 results obtained. All three datasets were used for testing.

Phase 3: Modifying Inference Methods

Phase 3 of our methodology involves optimising our model inference methods by increasing Beam Search Size, from 5 to 10. Beam Search is used to select and rank outputs from the different models. It does this by generating a specified number of unique outputs from the models, of which the best is used as the final result. Increasing the Beam Search size would increase the sampling range, but will also increase the inference time. Thus, we aim to find out if increasing it would provide a significant benefit to our models evaluations. The datasets used for training are similar to that of Phase 2, which is the Enron Corpus, Reddit Corpus and Singlish datasets with unimodal specified to true. Testing of the model is also conducted on all three datasets.

Improving the Prompt & Rerank Method

One major flaw identified when experimenting with the Prompt & Rerank method was the usage of GPT-2 Large as the Language Model (LM). GPT-2 Large was not optimised to be fine-tuned due to its size, and thus we experimented with GPT Neo instead. Due to time constraints, the dataset used also had to be scaled down so the training process could be done quicker. Fine-tuning is an essential step in adapting this method as it allows for the LM to cater specifically to our dataset, in this case the Enron Corpus, Reddit Corpus and Singlish datasets. Fine tuning was applied to GPT Neo over 2 epochs, after which testing was conducted on all 3 datasets.

The samples that were hand-written consisted of an impolite version, and a version that had its politeness transferred. The model was fed both the impolite and polite version, to demonstrate to the model how the politeness transfer was intended. Some of the examples also included passive-aggressive sentences, that was interpreted as impolite. 3 of these samples were selected and were fed before the model was prompted:

Impolite	**Polite**
your pants are so ugly.	your pants are nice but they are not for me.
it was overpriced and very useless.	it was priced on the high side and could work better.
no offence, but your manners are terrible.	I was slightly taken aback by your manners.
your customer service is terrible.	i could use more help for customer service.
i saw you do the dishes, i am surprised to say the least.	thank you for starting to do the dishes
you are too sensitive.	sorry, i didn\'t mean it that way.

Evaluating the Data

For the tag and generate method, we evaluated the results from all phases similarly using 2 types of evaluation metrics, which are accuracy and similarity. For the accuracy evaluation, we will evaluate the results by classifying it using the BERT classifier. We consider a score of more than or equal to 6.5, on a scale of 1 to 10 as polite, and represent the results as a percentage score. For similarity metrics, we use the BLEU-s (Papineni et al., 2002), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), and METEOR (Denkowski and Lavie, 2011) metrics to evaluate the similarity between the hypothesis and reference data. This helps to evaluate the context and meaning preservation. For the prompt-and-rerank method, we performed qualitative analysis of the data, as a quantitative evaluation of the data could not be done due to time constraints.

RESULTS AND DISCUSSION

Phase 1

Metric	**Unimodal = False 1 (%)**	**Unimodal = True 1 (%)**
Accuracy	68.38	73.00
Bleu_S	88.28	88.25
METEOR	58.98	57.74
ROUGE_L	91.20	92.84
CIDEr	82.03	83.12

Table 1: A comparison between the results of the unimodal setting toggled true and false

The main goal for phase 1 was to differentiate performance levels when the unimodal setting was toggled between true and false. We observe better performance with unimodal toggled to false only for the METEOR metrics, whereas for Accuracy, Bleu-s, ROUGE-L and CIDEr metrics, we observe better performance with unimodal toggled to true. The general trend is that the performance is better with unimodal set to **true**. We believe that there could be a few reasons for the differing performance between the two settings: firstly, with unimodal set to true, the algorithm focuses on politeness as a main style, while unimodal set to false could interpret politeness as “formality, sentiment (positive/negative)” which may not be as accurate, thus accounting for the observed poorer performance.

Comparison of Phase 2 and 3

Metric	**Phase 2 (%)**	**Phase 3 (%)**
Accuracy	83.64	83.09
Bleu_S	88.14	88.93
METEOR	60.22	60.93

ROUGE_L	93.29	94.00
CIDEr	79.41	81.33

Table 2: A comparison of the results between Phase 2 and 3 adapted to a Singaporean context

Incorporating the social media context, specifically for Singaporeans, required us to adapt different data types and ensure the model is trained to handle such data. Evidently, we added the Singlish and Reddit Datasets into the training and test data, while keeping the Enron Corpus as a main portion of the dataset. Phase 3 of our methodology sees us further improving on our training methods. The results show that phase 3 performs better for the similarity metrics, but poorer for the accuracy metric. A possible reason for this could be the doubling of the beam search size which increased the sampling size, giving the model more possible outputs thus more likely to find one which stands out in terms of similarity. However, the expanded sampling size results in some of the outputs chosen to be less accurate, as the beam search might prioritise content and meaning preservation more than the actual politeness. Regardless, the difference in performance is relatively marginal for this change, thus there is little advantage when using a larger beam search size.

Prompt And Rerank - Qualitative Analysis

Text	**Phase**	**Generated Samples**
i will attempt to learn more about this process	**1**	you should go ahead and tell me if you can read this first!
		i will get to the point where i will give you more help, but i will not try to help you with this process.
		i may attempt more help for customer service.
	2	the customer is not there, but the customers are not there and they have not bought

		i will try to learn more about this process.
		i am looking for a great customer service person.

Table 3: An example of a set of input and corresponding generated samples from the prompt and rerank method

It is evident from the generated samples that neither Phase 1 nor Phase 2 performed exceptionally when trying to improve the text of neutral politeness style to one of polite style. Furthermore, there is a frequent appearance of "customer service" as a topic, even though it is not discussed in the original text at all. This could be due to the prompts that were given to the language models, one of which had "customer service" as a topic. Thus, fine tuning the model does not seem to have a large effect on the effectiveness of politeness transfer, and future work could be done to fine tune this process of prompt-and-rerank for politeness transfer.

CONCLUSION

There are many ways to carry out the task of politeness transfer: in our research, we worked on improving on existing models, specifically the Tag & Generate Approach as well as the Prompt & Rerank method. By incorporating our own dataset, and modifying individual aspects of the model, we were able to achieve meaningful improvements in the transfer of performance, such as preserving meaning and content accuracy. In conclusion, the tag and generate method performed better than the prompt-and-rerank method. Furthermore, the beam search size of 5 for the tag and generate method is sufficient to ensure decent results. While our results were an improvement, due to various constraints, the most significant being the lack of time, there are still a lot of possible improvements to be made, especially for the Prompt & Rerank method.

ACKNOWLEDGEMENTS

We would like to thank our mentor, Dr Chieu Hai Leong, for supervising us throughout the project and clearing any doubts or questions that we had so that our progress could run smoothly. In addition, huge thanks to our Teacher-In-Charge, Mr Tay Zhi Xiong, for conducting

meetings to check on our progress, giving us tips on how to improve our model, and encouraging us to continually strive for the best.

REFERENCES

Bryan Klimt and Yiming Yang. 2004. Introducing the Enron Corpus. In CEAS.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In-Text summarization branches out, pages 74–81.

Hongmin Wang, Jie Yang, Yue Zhang. 2010. From Genesis to Creole language: Transfer Learning for Singlish Universal Dependencies Parsing and POS Tagging

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.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Madaan, Setlur, Parekh. 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.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the sixth workshop on statistical machine translation, pages 85–91. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriaszi, Dan Jurafsky. 23 May 2022. Prompt-and-Rerank: A Method for Zero-Shot and Few-Shot Arbitrary Textual Style Transfer with Small Language Models.

Reddit Corpus (small) — convokit 2.5.3 documentation. (n.d.). https://convokit.cornell.edu/documentation/reddit-small.html

Shirley Anugrah Hayati, Dongyeop Kang, Lyle Ungar. Understanding Linguistic Styles through Lexica. University of Pennsylvania, Georgia Institute of Technology, University of Minnesota.

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.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In Advances in neural information processing systems, pages 6830–6841.

Tianyi Zhang, Varsha Kishore, Felix Wu, Lilian Q. Weinberger and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In the *International Conference on Learning Representations.*