Polite vs. Direct Email Reply Generation

Alexander Motyka

Abstract

In this project, I explore tone-controlled email reply generation, where a model is tasked with generating either a polite or direct response to a given email. Using the Enron email dataset, I fine-tune three variants of the T5 transformer model (T5-small, T5-base, and T5-large) to conditionally generate replies in the desired tone. The final models were evaluated on ROUGE, BLEU, and perplexity to assess content quality, and on ConvoKit's normalized politeness score to evaluate stylistic fidelity. T5-large achieved the best performance, with a ROUGE-1 of 0.436, BLEU of 0.203, and perplexity of 3.312. More importantly, all models showed clear tone separation: T5-small generated polite replies with an average politeness score of 0.666 versus 0.424 for direct replies, while T5-large achieved an even wider separation (0.716 polite vs. 0.327 direct). These results demonstrate that conditioning large language models with tone-specific prompts enables meaningful control over sociolinguistic attributes like politeness while preserving semantic content.

Introduction

This project aims to build a text generation system that can produce both a polite/formal and a direct/informal reply, each retaining the same semantic meaning, to a given professional email. This task is situated within the broader field of style transfer in NLP, which focuses on rewriting text to match a target style without altering its core content. While much of the existing literature focuses on more overt stylistic features such as sentiment, politeness involves subtler and more context-sensitive linguistic signals. These include use of words such as “please” and “thank you”, hedging (“softening” a statement), gratitude, indirectness, and use of honorifics. Tone control in generated emails has clear real-world applications. Users of productivity tools, such as Gmail's Smart Reply, may wish to adjust the tone of automatic replies depending on the recipient, the context, or company norms. A system that can shift tone without distorting meaning could greatly increase the utility and adaptability of these tools. This project explores the feasibility of building such a system using the T5 transformer family.

Background & Related Work

1. Danescu-Niculescu-Mizil et al. (2013) introduced a classifier for detecting politeness using syntactic and lexical cues. Similar to this project, it focuses on politeness detection, though it uses feature engineering rather than transfer models.
2. Rao and Tetreault (2018) demonstrated that deep learning models outperform traditional classifiers on politeness detection. While their project also classifies tone, it also relies on parallel data, which this project does not.
3. Jin et al. (2021) surveyed text style transfer techniques. This paper will serve as a guide once I reach the evaluation stage.
4. Niu and Bansal (2018) proposed methods for polite dialogue generation using weak supervision, including reinforcement learning. Their work aligns closely with this project in goal and methodology, though I use a classification-based pipeline instead of RL.
5. Fu et al. (2017) used adversarial networks to perform non-parallel style transfer and introduced transfer strength and content preservation metrics. While this project doesn't concern adversarial learning, these are valuable evaluation metrics.

Approach

The goal of this project was to create a system capable of generating tone-controlled email replies- specifically producing either polite or direct responses that retain the same semantic meaning as the original. To achieve this, I fine-tuned transformer-based language models from the T5 family (T5-small- ~60 million parameters, T5-base- ~220 million, T5-large- ~780 million) on email prompt-reply pairs that had been labeled with politeness. The model input was formatted as a prompt with an explicit tone token: [TONE=polite] <prompt text>, followed by an instruction: “Generate a [TONE] reply.” This design enabled the model to learn a conditional generation task.

The main dataset for this project is the Enron Email Dataset, a large public archive of approximately 500,000 real corporate emails. This dataset is ideal for tone-conditioned email generation because, as noted by Cohen (2015), “it is the only substantial collection of ‘real’ email that is public.” However, the dataset poses a major preprocessing challenge: each email thread, including the subject line, headers, timestamps, sender/recipient metadata, and the entire message history, is stored in a single unformatted string within one column. Initially, I attempted to extract clean (prompt, reply) pairs using hard-coded logic, but this approach proved brittle and error-prone due to the variability in thread structure and inconsistencies in header formatting.

To address this, I implemented a parsing system using the OpenAI API. I created few-shot examples of properly parsed emails, which guided GPT-3.5-turbo in converting raw threads into structured (prompt, reply) pairs. The model was prompted to identify the last message in the thread as the reply, and one immediately preceding it as the prompt. Once I had extracted the structured training pairs, I labelled each reply with a politeness tone using a two-stage strategy. My original plan was to use the HuggingFace gljj/politeness-model classifier; however, it proved unsuitable, as its sentence-level predictions were inconsistent, difficult to interpret, and the scores were confined to a narrow range (0.1–0.35), making them hard to threshold meaningfully. Instead, I used the ConvoKit politenessStrategies framework by Danescu-Niculescu-Mizil et al., which computes interpretable politeness scores based on linguistic strategies such as hedging, apologizing, and directness, and sets binary flags for each of these features, making it very clear which words and grammatical structures contribute to the politeness of the statement as a whole. I categorized each of these features as beneficial or detrimental to total politeness, and calculated a score for each instance as the sum of beneficial flags minus the sum of detrimental ones. I normalized these scores to the [0, 1] range and labeled replies in the top quartile as polite and those in the bottom quartile as direct, dropping the middle two quartiles so as to only show the model "extreme" data. This produced more stable and semantically meaningful tone labels for training.

For model training, I used HuggingFace's Trainer API to fine-tune each T5 variant. Training was performed for 12 epochs with a batch size of 4 and a learning rate of 1e-5. The inputs and outputs were tokenized using T5's tokenizer, with maximum lengths of 512 tokens for input and 128 for output. The same preprocessing and training setup was used across T5-small, T5-base, and T5-large to allow for fair comparisons. At this step, I had to drastically scale back the amount of data used to train the model. When using a hard-coded prompt:reply extractor, I was able to train the model on ~10,000 instances. However, when transitioning to the OpenAI API approach, I was handicapped by daily rate limits. Even after making the maximum possible number of requests in a day (10,000) to build the training set, I was left with only ~1,500 training instances after dropping single or empty messages, and those inside the middle two quartiles of the politeness distribution.

Experiments & Results

My experiments aimed to answer the following core questions: (1) Can T5 models learn to generate email replies that reflect a requested tone? (2) How does model scale impact tone fidelity and content quality? (3) What are the typical failure cases? To answer these, I trained and evaluated T5-small, T5-base, and T5-large on the same set of gpt-3.5-turbo-extracted, ConvoKit-tone-labeled Enron prompt:reply pairs. Each model was evaluated on the same held-out test set. The following table summarizes the key metrics:

Model	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity	BLEU	Politeness (Direct)	Politeness (Polite)
T5-Small	0.387	0.285	0.218	4.265	0.153	0.424	0.666
T5-Base	0.419	0.308	0.233	3.858	0.178	0.397	0.683
T5-Large	0.436	0.317	0.241	3.312	0.203	0.327	0.716

We observe clear trends. First, as model size increases, both semantic metrics (ROUGE, BLEU, perplexity) and stylistic separation improve. The BLEU score rises from 0.153 for T5-small to 0.203 for T5-large, indicating improved precision in capturing n-grams from reference replies. Perplexity, a measure of generation fluency, drops from 4.265 to 3.312, showing that larger models are more confident and fluent in their outputs. In terms of tone control, all models produced replies with noticeably different average politeness scores by tone. For example, T5-small's generated polite replies averaged a score of 0.666 vs. 0.424 for direct replies. This gap widened slightly with model scale- T5-large achieved a difference of nearly 0.39 between the two tones (0.716 vs. 0.327). This suggests that larger models internalize the tone conditioning more effectively, producing clearer distinctions between politeness levels.

While the models show clear improvement in both semantic and stylistic performance as scale increases, qualitative outputs reveal several limitations. In low-information prompts, such as the trade data in the fourth example of the following section, the model often fails to generate a coherent response, as there is little context for the model to pick up on. T5-small also frequently echoes the input verbatim, suggesting it struggles to generalize or rephrase unless explicitly guided by richer context. This parroting behavior diminishes as model size increases: T5-base begins to introduce phrasal variation and softening ("Thank you for your help"), while T5-large demonstrates tone-conscious rewrites, such as adding polite hedging or clarifying questions in polite responses, or offering concise confirmations in direct ones. Another strength observed across polite generations is the emergence of politeness markers such as "sorry," "thank you," and "please," especially in T5-base and T5-large. These cues suggest that the model has successfully learned certain stylistic features from the politeness-labeled training data and is able to deploy them appropriately at generation time.

Below are examples of actual outputs from each model on the same inputs, illustrating differences in tone realization.

Tone: Polite

- Prompt: “Wes - Are we still anticipating that Bridgeline will be allowed mark-to-market status at some point in the future? If so, what criteria must Bridgeline meet and what is the expected timeline on it? If this is going to happen soon, I would suggest that we avoid getting Commercial involved and developing a process to track our 40% portion of Bridgeline.”
- Reference: “Sorry I am so late in replying! It would seem that the MTM issue is Bridgeline's decision and would solely be based on whether or not they are, in fact, trading as defined in EITF 98-10. There are about 4 pages of criteria set forth in that document. That may be a question you want to ask Trevor Mihalik, but I do know that initially they did not meet 98-10 criteria. Their business may well have changed by now. I am confused by your suggestion about whether or not to involve commercial.”
- Generated (t5-small): “Wes, Are you still anticipating that Bridgeline will be allowed mark-to-market status at some point in the future? If so, what criteria must Bridgeline meet and what is the expected timeline on it?”
- Generated (t5-base): “I'm not sure if this is going to happen soon, but we will probably have to wait until the end of the year. Wes”
- Generated (t5-large): “I'm not sure if this is going to happen anytime soon. If so, what criteria must Bridgeline meet and what is the expected timeline for it?”

Tone: Polite

- Prompt: “I'm sorry my reply is so delayed - I was gone for the New Year holiday. This deal should have been entered on 12/19. It looks like it's still in there as 12/20; but I wanted to check with you before changing it. Let me know.”
- Reference: “O.K. you can change the date.”
- Generated (t5-small): “I'm sorry my reply is so delayed - I was gone for the New Year holiday. This deal should have been entered on 12/19.”
- Generated (t5-base): “Thank you for your help. I'm sorry to hear that this deal is in there as 12/20. It should be entered on 12/19.”
- Generated (t5-large): “Thank you for your help. I'm sorry to hear that this deal is in there as 12/20. It should be entered on 12/19.”

Tone: Direct

- Prompt: “Hi. The Market Risk Meeting has been scheduled for today at 5:15 p.m. Houston Time. The dial in information is as follows: Dial In Number: 203-719-0070 Freephone within the U.S.: 877-315-5218 Pin Number: 808787 Mercy, please reserve a conference room for those participants in Houston. Thanks. Jm”
- Reference: “i am waiting on the conference call and no one is on it yet?”
- Generated (t5-small): “I'm not sure if you have any questions or concerns about the market risk meeting will be held today at 5:15 p.m.”
- Generated (t5-base): “Thanks for the call. I'll be in Houston on Monday, May 15th at 5:15 PM. Mercy”
- Generated (t5-large): “Hi. The Market Risk Meeting is scheduled for today at 5:15 p.m. Houston Time. The dial in information is as follows: Dial In Number: 203-719-0070 Freephone within the U.S.: 877-315-5218 Pin Number: 808787”

Tone: Direct

- Prompt: “587415 not apb missing deals mark f buy srp 217.00 25 mw pv on peak holden sell williams 24-30 102.00 25 mw pv off peak holden sell williams 24-30 100.00 50 mw pv off peak tom a buy pacificorp 200.00 23rd 25 mw on peak tom a buy aep 283.00 25 mw may pv on peak”
- Reference: “587415 - changed to Bloomberg All the rest of these were just entered, except last deal Tom buys AEP May PV 283.00 Tom does not recognize this trade - is APB sure of the terms?”
- Generated (t5-small): “[TONE=direct] 587415 not apb missing deals mark f buy srp 217.00 25 mw pv on peak holden sell williams 24-30 102.00 25 Mw”
- Generated (t5-base): “587415 not apb missing deals mark f buy srp 217.00 25 mw pv on peak holden sell williams 24-30 102.00 25-mw-pv off peak tom a buy pacificorp 200.00 23rd 25 MW on peak”
- Generated (t5-large): “I'm not sure if there are any deals missing from the list, but it looks like they may be. I’ll get back to you. Thanks”

<u>Conclusion</u>

This project demonstrates that transformer-based language models can be effectively fine-tuned to generate email replies with controllable tone. The results show that tone conditioning is both learnable and scalable. Even with a relatively small fine-tuning dataset, all three T5 models were able to distinguish between polite and direct tones, as evidenced by consistent separation in politeness scores. Larger models, especially T5-large, not only produced more fluent and coherent responses (as reflected in lower perplexity and higher BLEU/ROUGE scores) but also demonstrated greater stylistic sensitivity, using hedging, gratitude, and honorifics more appropriately in polite responses.

At the same time, the project revealed important limitations. Smaller models tended to parrot the prompt, especially in sparse or low-context inputs. All models occasionally struggled to differentiate tone in low content messages (e.g., trade data or scheduling notifications). Still, the progression in generation quality from T5-small to T5-large suggests that future improvements in model architecture and training data scale will continue to yield more stylistically controllable outputs.

There are several promising directions for extending this work. One is to introduce sentence-level tone control within a single reply, allowing models to vary tone dynamically based on context (e.g., opening with formality and closing with brevity). Another is to support multi-turn email threads by conditioning on full conversational history, enabling more coherent and context-aware responses. Additionally, incorporating metadata such as recipient identity or role could allow the model to adapt tone appropriately- using more deferential language for executives, for example. Finally, exploring semi-supervised or reinforcement learning approaches could provide finer-grained control over tone, especially in scenarios where explicit style labels are limited or unavailable.

References

Cohen, William. *Enron Email Dataset*, CMU, 8 May 2015, www.cs.cmu.edu/~enron/.

Cukierski, Will. “The Enron Email Dataset.” *Kaggle*, CMU, 16 June 2016, www.kaggle.com/datasets/wcukierski/enron-email-dataset/data.

Danescu-Niculescu-Mizil, Cristian, et al. “A Computational Approach to Politeness with Application to Social Factors.” *ACL Anthology*, 2013, aclanthology.org/P13-1025/.

Fu, Zhenxin, et al. “Style Transfer in Text: Exploration and Evaluation.” *arXiv.Org*, 27 Nov. 2017, arxiv.org/abs/1711.06861.

“GLJJ/Politeness-Model · Hugging Face.” *Gljj/Politeness-Model · Hugging Face*, huggingface.co/gljj/politeness-model.

Jin, Di, et al. “Deep Learning for Text Style Transfer: A Survey.” *arXiv.Org*, 16 Dec. 2021, arxiv.org/abs/2011.00416.

Niu, Tong, and Mohit Bansal. “Polite Dialogue Generation without Parallel Data.” *ACL Anthology*, 2018, aclanthology.org/Q18-1027/.

“PolitenessStrategies.” *politenessStrategies - Convokit 3.1.0 Documentation*, Cornell, convokit.cornell.edu/documentation/politenessStrategies.html.

Raffel, Colin, et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” *arXiv.Org*, Google, 19 Sept. 2023, arxiv.org/abs/1910.10683.

Rao, Sudha, and Joel Tetreault. “Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer.” *arXiv.Org*, 16 Apr. 2018, arxiv.org/abs/1803.06535.

T5, HuggingFace, huggingface.co/docs/transformers/v4.13.0/en/model_doc/t5.