Mixing It Up: Inducing Empathy and Politeness using Multiple Behaviour-aware Generators for Conversational Systems

Mauajama Firdaus[1*] and **Priyanshu Priya**[2*] and **Asif Ekbal**[2]
[1]University of Alberta, Canada
[2]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[1]mauzama.03@gmail.com,{[2]priyanshu_2021cs26,asif}@iitp.ac.in

Abstract

Politeness is a key component that can assist in building a strong customer-agent relationship. With the ongoing increase in customer-care systems, it is crucial to have healthy relations with the users providing satisfaction and a better customer experience. In this regard, it is significant to model the different polite behaviors in an agent to help the user in reaching the intended objectives. In our current work, we propose the task of polite behavior-aware generation considering the affective state of the user and the conversational context. We design a Transformer based encoder-decoder framework with three major components i.e., Affective tracker, Behaviour-aware generators, and Polite generator. The *affective tracker* is a context encoder that captures the contextual information along with the affective information in the utterances; the *behavior-aware generators* independently attends to the context information to compute behavior-aware polite representations and finally, *polite generator* generates the final polite response considering the representations from different generators. Experimental results on the CYCCD dataset prove that our approach generates contextually correct and relevant responses compared to the state-of-the-art approaches and the baselines.

1 Introduction

The technological advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have led to the proliferation of conversational systems (chatbots and personal assistants like Amazon's Alexa, Google's Home, and Apple's Siri) in our daily lives. These conversational systems that aim to assist users in numerous ways like making reservations, booking flights, scheduling appointments and many more are prevalent in several areas such as hospitality, education, and health care, to name a few. The success of these systems significantly depends on their ability to effectively communicate with the users. Thus, the research in recent times has been inclined towards modulating biases, styles, and control in text generation to improve these interactions.

Customer care is a typical application area for automated task-oriented conversational systems. These systems provide potentially cost-effective and reliable solutions for customer care. Nowadays, companies rely heavily on customer care service to successfully guide and assist customers and establish long-term and healthy relationships with them. Customers generally interact with conversational agents with diverse expectations and attitudes (Følstad and Skjuve, 2019). In such situation, if the agent fails to satisfy their expectations or fix their issue as desired, they may get frustrated or furious quickly (Jain et al., 2018). To avoid this, it is imperative for customer care agents to mimic human-like (Fink, 2012) behaviour during the conversation.

Politeness and empathy, which encompass social belongingness are a few fundamental sociolinguistic cues of humans (Brown et al., 1987; Maslow, 1943). The use of social language by the agents facilitates building rapport and emotional connection with the customers. Politeness has been investigated recently (Golchha et al., 2019; Firdaus et al., 2022a; Priya et al., 2023a; Mishra et al., 2023a, 2022a; Firdaus et al., 2022b; Mishra et al., 2022b, 2023c,b; Priya et al., 2023b) to ensure user satisfaction and to build strong user relations. Politeness could have different behavioral aspects linguistically such as one could be apologetic, appreciative, empathetic, or could be respectful by always greeting and assuring. Such behavioral patterns could lead to different responses that are contextually plausible and interactive.

Emotion and sentiment of the user are essential for meeting the user's wants appropriately and facilitating the creation of smooth and amicable conver-

*The authors are jointly first authors.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 336–347

Dialog Context	Customer Sentiment	Customer Emotion	Generic Response	Polite Response	Polite behaviour
I need a software update urgently, but I am unable to connect to the network.	Negative	Frustrated	What is happening with your internet?	Please don't worry, we can help! Kindly tell what is happening with your internet?	*Assurance*
Hey, i got stomach ache from your inflight meal on monday.	Negative	Anger	Send us a dm.	That's really sad to hear. We are sorry, please send us a dm.	*Apology*
Hi! I want help.	Neutral	Hopeful	What are you looking for?	Hey good morning! Good to have you with us, please let us know what are you looking for?	*Greet*
Dear this new update is awesome, got great new apps!	Positive	Joy	The update has many features.	Thank you very much, please checkout the exciting features in the update.	*Appreciation*
I lost my bag.	Negative	Sad	We will look into the matter.	That's really disappointing to hear. Please have patience until we look into the matter.	*Empathy*

Table 1: Examples of variation in polite behaviour in accordance with user's affective state (emotion and sentiment)

sations (Beale and Creed, 2009; Shi and Yu, 2018; Firdaus et al., 2020a, 2022c, 2023; Singh et al., 2022; Samad et al., 2022; Madasu et al., 2022; Firdaus et al., 2022c,d). The usage of user feedback in the form of emotion and sentiment is essential for generating contextually coherent polite responses reflecting relevant polite behaviour in the responses as depicted in Table 1. For the first example in Table 1, it can be seen that the user's emotion and sentiment are negative. The appropriate response in such a scenario could either exhibit empathy or apologize or provide assurance. For this particular example, the polite behaviour is assurance which could help in smoother conversations with the user.

In our current work, we take a step ahead and address the task of behaviour-aware polite response generation. To the best of our knowledge, this is one of the very first attempts that addresses the linguistically driven different behavioural patterns of the polite generation. We build an end-to-end transformer-based encoder-decoder architecture for the task of behaviour-aware polite response generation. Our proposed architecture captures the complete affective information in the form of emotion and sentiment from the conversational context and uses this knowledge for generating behaviour-aware polite customer care responses.

In summary, the key contributions of our work are three-fold. We first introduce the task of behaviour-guided polite response generation for customer-care systems. Second, we design a transformer-based encoder-decoder network having three crucial components, i.e., an affective tracker, behaviour-aware generators, and a final polite generator to capture the emotional quotient from the context and generate behaviour-wise response representations and finally generate the polite response using the weighted-sum approach. Lastly, experimental analysis on the CYCCD dataset shows that our approach performs better than all the existing baselines and generates more informative and varying responses.

2 Related Work

Natural language generation (NLG) module has been gaining prominence in a variety of applications, including dialogue systems (Shen et al., 2018; Zhang et al., 2018), question answering systems (Indurthi et al., 2017; Duan et al., 2017), and various other natural language interfaces. Users' feelings in the form of sentiments and emotions have been exploited in (Acosta, 2009; Pittermann et al., 2010; Shi and Yu, 2018; Firdaus et al., 2021b; Dias et al., 2022; Firdaus et al., 2021a, 2020b,d) to give humanly essence to the system, thereby facilitating better user experience. Emotion recognition and sentiment analysis in customer support system is essential for understanding and to provide better customer support (Herzig et al., 2016; Wang et al., 2020a).

Politeness in customer support systems is important for attaining customer satisfaction and retention (Bickmore and Picard, 2005; Bickmore et al., 2009; Liao et al., 2016; Wang et al., 2020b). In the past, (Gupta et al., 2007) explored building more affective and socially intelligent dialogue systems by incorporating different politeness strategies in the responses. Lately, (Niu and Bansal, 2018a) proposed a reinforcement learning (RL)-based model to induce politeness in chit-chat conversations without parallel data. Golchha et al. (2019); Firdaus et al. (2020c) presented a method for increasing user satisfaction by instilling courteous nature in the customer-care responses by exploiting reinforced pointer networks. Madaan et al. (2020) devised a tag and generate framework for transforming non-polite sentences into polite ones. Firdaus et al. (2022a) transformed generic customer-care responses into polite ones based on user's sentiment and conversational history using an RL-based approach. Firdaus et al. (2022b) designed a reinforced deliberation network based framework to inculcate politeness in the responses according to the personalized user information (age and gender) and dialog context.

Our present work is different in the sense that we aim to incorporate different polite behaviour in the generated responses according to emotion and sentiment of the user. The inclusion of user feedback in the form of emotion and sentiment provides complete affective information, which in turn, enhances the quality of generation and makes the responses contextually coherent with the dialog.

3 Methodology

In this section, we present the problem definition of our current task and provide a detailed description of our proposed approach. Our approach focuses on generating responses based on the speaker's emotional state from the conversational context constrained by the different polite behaviour such as *Appreciation, Empathy, Apology, Greeting, and Assurance*.

3.1 Problem Definition

Our current work aims at generating contextually appropriate responses based on different polite behaviour. Precisely, the conversation has an alternating set of utterances from the customer (U^i) and the customer care agent (A^i). We represent the conversational context as $C = (\mathcal{U}^1, \mathcal{A}^1, \mathcal{U}^2, \mathcal{A}^2, \cdots, \mathcal{U}^i)$ and each utterance is a sequence of words $w_1, w_2, \ldots, w_N$.

Similar to (Golchha et al., 2019), we use the output distribution from DeepMoji (Felbo et al., 2017) (pre-trained on the emoji prediction task) to get the emotional embeddings (E^{emo}) associated with both the customer and agent utterances. In addition, we also use the sentiment information (E^{sen}) along with the emotional embeddings to get complete affective information. This information is added with every utterance to assist in generating the final behaviour-aware polite response.

3.2 Approach

In this section, we describe each component of our proposed model in detail. Overall, our model is composed of three components: a *Affective tracker*, *Behaviour-aware generators* and a final *Polite generator* as shown in Figure 1. The *Affective tracker* is basically the context encoder that provides the representation of the context and computes a distribution over the possible sentiment and emotion categories for the context. The *Behaviour-aware generators* independently attend to this distribution to compute their own representation. Finally, the *Polite generator* takes the weighted sum of representation from the different generators and generates the final polite response. We also incorporate speaker information as embedding in the input with every utterance. This is used to facilitate the encoder to differentiate between the customer and agent utterances.

Affective Tracker: The context encoder uses a standard transformer encoder (Vaswani et al., 2017) for the sentiment and emotion (Affective) tracker. To capture the hierarchical nature of the conversation, we use a hierarchical transformer encoder to capture the utterance-level and dialogue-level representations. We use the utterance transformer to get the encoded utterance representation. To learn the representation of each utterance U^i, $U^i = w_1^i, w_2^i, \ldots, w_N^i$ is first mapped into continuous space

$$T_u = (t_1^i, t_2^i, \ldots, t_{|U^i|}^i); where [t_j^i = e(w_j^i) + p_j] \quad (1)$$

where, $e(w_j^i)$ and p_j are the word and positional embedding of every word w_j^i in an utterance, respectively. For words, we use Glove embeddings and adopt the sine-cosine positional embedding (Vaswani et al., 2017) as it performs better and does not introduce additional trainable parameters.

The utterance encoder (a Transformer) converts T_u into a list of hidden representations $h_1^i, h_2^i, \ldots, h_{|U^i|}^i$. We use the last hidden representation $h_{|U^i|}^i$ (i.e. the representation at the EOS token) as the textual representation of the utterance U^i. Specifically, the final utterance representation E^U is the sum of the final textual representation $h_{|U^i|}^i$, the positional embedding E^p (Vaswani et al., 2017), the speaker embedding E^s, the emotion embedding E^{emo} and the sentiment embedding E^{sen} of the corresponding utterance.

$$E^U(C) = h_{|U^i|}^i + E^p(C) + E^s(C) + E^{emo} + E^{sen} \quad (2)$$

In a similar fashion, we encode the agent utterances as well.

The final utterance representation is used as input for the dialogue transformer that encodes the dialog context into a context representation. In order to compute the weighted sum of the output tensor, we add a query token QRY at the start of each input sequence, similar to BERT (Devlin et al., 2018). If a transformer-context encoder is designated as TRS_{enc}, then the relevant context

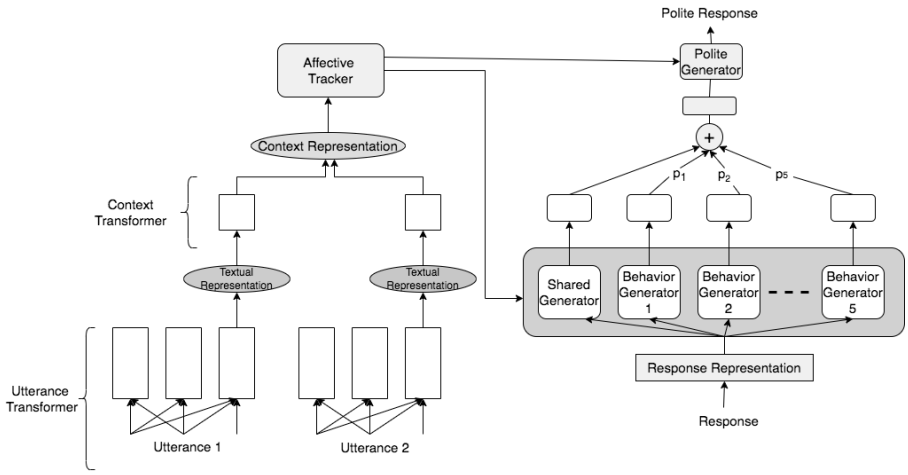

Figure 1: Architectural diagram of our proposed behaviour-aware polite generator

representation is:

$$H^C = TRS_{enc}(E^U([QRY; C])) \quad (3)$$

where [;] denotes concatenation, H $\in R^{L_{model}}$ where L is the dialogue length. Then, we determine the final representation of the token QRY as $q = H_0$ where $q \in R^{d_{model}}$, which is then used as the query for generating the sentiment and emotion distribution.

Behaviour-aware Generators: The behaviour-aware generators mainly consist of (1) a shared behaviour generator that learns shared information for all the emotions and their corresponding polite behaviour and (2) n independently parameterized Transformer decoders (Vaswani et al., 2017) that learns how to appropriately react in a polite-manner given a particular affective state.

A standard transformer decoder layer block, abbreviated as TRS_{dec}, models all of the generators and is composed of three sub-components: a position-wise fully connected feed-forward network, multi-head attention over the output of the Affective tracker, and a multi-head self-attention over the response input embedding.

As a result, we specify the generator's set as $G = [TRS^0_{dec}, ..., TRS^n_{dec}]$. Each generator constructs its own polite response representation P_i using the target sequence altered by one $r_{0:t-1}$.

$$P_i = TRS^i_{dec}(H^C, E^R(r_{0:t-1})) \quad (4)$$

where TRS^i_{dec} refers to the i^{th} generator, including the shared one. From a conceptual standpoint, we anticipate that the shared generator, TRS^0_{dec}, will produce a general representation that will aid the model in capturing the discourse context. On the other hand, we anticipate that each behaviour-aware generator will develop its ability to react in a designated polite behaviour based upon a certain emotion and sentiment. In order to simulate this phenomenon, we give each behaviour-aware generator a varied weight based on the user's emotional distribution, while giving the shared listener a set weight of 1.

To illustrate, we create a Key-Value Memory Network (Miller et al., 2016) and represent each memory slot as a pair of vectors (k_i, P_i), where $k_i \in R^{d_{model}}$ signifies the key vector and P_i is from Equation 4. The key vectors k are then addressed using the encoder-informed query q by performing a dot product and then a Softmax function as follows:

$$v_i = \frac{e^{q^T k_i}}{\sum_{j=1}^{n} e^{q^T k_i}} \quad (5)$$

The weight of each listener is determined by using each v_i as the score given to P_i. Given the speaker's affective state e_t, we oversee each weight v_i throughout training by using a cross-entropy loss function to maximize the likelihood of the emotion state e_t:

$$\mathcal{L}_1 = -logp_{e_t} \quad (6)$$

Finally, the weighted sum of the shared generator output P_0 and the memory values P_i is used to calculate the combined output representation.

$$P_M = P_0 + \sum_{i=1}^{n} v_i P_i \quad (7)$$

Polite Generator: The polite generator is then implemented using a second transformer decoder layer, which further transforms the behaviour-aware generator's representation and produces the

final response. On the basis of emotion and sentiment, each behaviour-aware generator is thought to specialize in a certain polite behaviour, and the polite generator compiles the ideas from different generators to create the final response.

Therefore, we define another TRS_{dec}^{final}, and an affine transformation $W \in R^{d_{model} \times |P|}$ to compute:

$$P_{fin} = TRS_{dec}^{final}(H, P_M) \quad (8)$$

$$S_t = p(r_{1:t}|C, r_{0:t-1}) = softmax(P_{fin}^T W) \quad (9)$$

where $P_{fin} \in R^{d_{model} \times t}$ is the output of the polite generator and $p(r_{1:t}|C, r_{0:t-1})$ is the distribution over the vocabulary for the next tokens. The response prediction is then optimized using a standard maximum likelihood estimator (MLE):

$$\mathcal{L}_2 = -logp(S_t|C) \quad (10)$$

Finally, by minimizing the weighted-sum of two losses, all the parameters are jointly trained end-to-end to optimize behaviour selection and response generation:

$$\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 \quad (11)$$

where α and β are trainable hyperparameters to balance the two loss functions.

4 Dataset and Experiments

In this section, we present the dataset used for our experiments followed by training details, baselines, and evaluation metrics.

4.1 Dataset

For our current work, we use the CYCCD dataset (Golchha et al., 2019)[1] having interactions between customers and professional customer care agents of companies on their Twitter handles. The CYCCD Twitter data was taken from the dataset made available on Kaggle by Thought vector. We use the generic and polite annotated version of the CYCCD dataset in a similar manner as (Golchha et al., 2019). The dataset consists of 140k, 20k and 40k conversations in training, validation and test set respectively.

As the CYCCD dataset was not annotated for different polite behaviour, therefore we do the polite-behaviour annotations for the dataset. To annotate the CYCCD dataset with behavioural information, we employ crowd-workers from Amazon Mechanical Turk (AMT) that label every utterance with the provided set of labels (i.e., assurance, empathy, apology, greeting, appreciation). For labelling the utterances, the workers were asked to follow the instructions and guidelines provided for annotation.

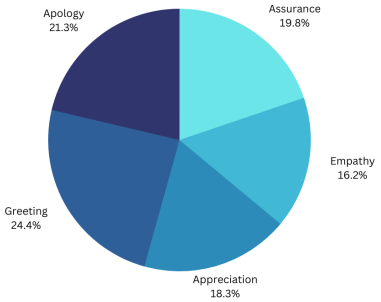

Figure 2: Polite-behaviour distribution in the CYCCD dataset

Some of the significant guidelines for annotation were as follows: (i). Every utterance of a given dialogue was to be marked with the provided labels; (ii) In addition, the workers were asked to provide the overall behaviour label for every sentence in an utterance as well. For cases where we found different annotations in polite behaviour for a particular sentence, we remove them from the dataset, and we also drop the entire conversation to maintain coherence among the utterances. A majority voting scheme was used for selecting the final label for every sentence. We observe a multi-rater Kappa (McHugh, 2012) agreement ratio of approximately 75%, which can be considered as reliable. The polite-behaviour distribution of the CYCCD dataset is provided in Figure 2.

4.2 Training Details

We used 300-dimensional word embedding and 300 hidden sizes in each and every experiment. We use word embedding initialized with Glove (Pennington et al., 2014) embedding pre-trained on Twitter and share it across the encoder and decoder. The rest of the parameters are randomly initialized. We employ two self-attention layers, each with two attention heads and a 40-element embedding dimension. We substitute a 1D convolution with 50 filters of width 3 for the Positionwise Feedforward sub-layer. We use batch sizes of 1 while testing while using batches of 16 during training.

In order to train our model, we used the AMSGrad (Reddi et al., 2019) as the optimizer to mitigate the slow convergence issues and changed the learning rate following (Vaswani et al., 2017). For

[1]https://github.com/Mauajama/Courteously-Yours

Model Description		PPL	BLEU-4	Rouge-L	PA
Existing Approaches	*Seq2Seq* (Sutskever et al., 2014)	1.112	0.145	0.278	0.38
	HRED (Serban et al., 2015)	1.085	0.198	0.308	0.45
	Polite-RL (Niu and Bansal, 2018b)	1.028	0.224	0.321	0.69
	PT-TGA (Madaan et al., 2020)	1.032	0.251	0.332	0.68
	PG-RL (Golchha et al., 2019)	1.018	0.264	0.339	0.73
	HT + RL + SE (Firdaus et al., 2022a)	1.004	0.275	0.352	0.77
Proposed Approach	**AT + BG + PG**	**0.987**	**0.310**	**0.382**	**0.81**
Ablation Study	*HT*	1.015	0.269	0.343	0.70
	HT + AT + PG	1.002	0.286	0.365	0.78
	HT + BG + PG	0.995	0.288	0.371	0.79

Table 2: Automatic evaluation results. Here, PPL: Perplexity, PA: Politeness accuracy, HT: Hierarchical transformer, AT: Affective Tracker, BG: Behaviour-aware generator, PG: Polite Generator

simplicity, we set the weight of both losses α and β to 1. *Affective Tracker* may assign weights to the *behaviour-generators* at random during the initial training phase and may send noisy gradient flow back to the incorrect *generators*, which can hinder model convergence.

4.3 Baselines

To demonstrate the effectiveness of our proposed model, we compare it with the previous state-of-the-art (SoTA) models:

Seq2Seq: It is the standard encoder-decoder framework with an attention mechanism that has been widely used in the generation, machine translation, etc. (Sutskever et al., 2014).

HRED: It is a hierarchical encoder-decoder model proposed for text-based dialogue systems (Serban et al., 2015).

Polite-RL: We implement the Polite-RL framework to induce politeness in responses in a similar manner as (Niu and Bansal, 2018b).

PT-TGA: We implement the politeness transfer framework presented in (Madaan et al., 2020) that uses a tag and generate approach to incorporate politeness.

PG-RL: We also take the reinforced pointer generator network employed in (Golchha et al., 2019) as one of the baselines to infuse politeness in generic responses.

HT+RL+SE: We also compare with the reinforced transformer network having sentiment information (Firdaus et al., 2022a) as one of the baselines to generate polite responses.

4.4 Evaluation Metrics

In this section, we describe both the automatic and manual evaluation metrics used to evaluate the performance of our proposed model.

Automatic Evaluation Metrics: To evaluate the model at the relevance and grammatical level, we report the results using the standard metrics like Perplexity (Chen et al., 1998), Rouge-L (Lin, 2004) and BLEU-4 (Papineni et al., 2002). We also report the Politeness Accuracy as a metric to measure the degree of politeness in the responses. We compute the politeness score using a pre-trained classifier, BERT (Devlin et al., 2018)[2] for measuring the degree of politeness in the generated responses similar to (Niu and Bansal, 2018b). The classifier takes as input the generated response and generates a probability value giving us the politeness accuracy of the generated response.

Manual Evaluation Metrics: We recruit six annotators (in a similar manner as (Firdaus et al., 2022c; Tian et al., 2019)) from a third-party company, having high-level language skills. We sampled 250 responses per model for evaluation with the utterance and the conversational history provided for generation. First, we evaluate the quality of the response on two conventional criteria: *Fluency* and *Relevance*. These are rated on a five-scale, where 1, 3, and 5 indicate unacceptable, moderate, and excellent performance, respectively, while 2 and 4 are used for unsure.

Secondly, we evaluate the politeness quotient of a response in terms of *Politeness Appropriateness* metric that measures whether the politeness induced in the response is in accordance with the user's affective state (both emotion and sentiment) and the dialogue history. Here, 0 indicates irrelevant or contradictory, and 1 indicates consistent with the provided persona and dialogue context.

We compute Fleiss' kappa (Fleiss, 1971) to mea-

[2]The classifier is trained on the Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013) and achieves an accuracy of 92%.

Model Description		F	R	PA
Existing Approaches	*Seq2Seq* (Sutskever et al., 2014)	3.82	3.73	48%
	HRED (Serban et al., 2015)	3.86	3.78	52%
	Polite-RL (Niu and Bansal, 2018b)	3.91	3.79	61%
	PT-TGA (Madaan et al., 2020)	4.03	3.85	64%
	PG-RL (Golchha et al., 2019)	4.11	4.06	67%
	HT + RL + SE (Firdaus et al., 2022a)	4.23	4.17	75%
Proposed Approach	**AT + BG + PG**	**4.35**	**4.53**	**80%**
Ablation Study	*HT*	4.09	4.03	65%
	HT + AT + PG	4.25	4.19	76%
	HT + BG + PG	4.28	4.22	78%

Table 3: Human evaluation results. Here, F: Fluency, R: Relevance, PA: Politeness Appropriateness, HT: Hierarchical transformer, AT: Affective Tracker, BG: Behaviour-aware generator, PG: Polite Generator

sure inter-rater consistency. The Fleiss' kappa for fluency and relevance are 0.53 and 0.49, indicating moderate agreement. For politeness appropriateness, we obtain 0.65 as the kappa score indicating substantial agreement.

5 Results and Analysis

In this section, we present the results of both automatic and manual evaluation. In addition, we also provide a few examples of our generated responses to showcase the effectiveness of our proposed model followed by a brief error analysis of our approach and the baselines.

Automatic Evaluation Results: To illustrate the efficacy of our model we provide the automatic evaluation results in Table 2. The table shows the results of different existing approaches that generate polite responses followed by the results of our model and the ablation study for our approach. The perplexity of our approach is the lowest compared to the existing approaches, i.e., *0.987* indicating the responses are fluent and grammatically correct.

The BLEU-4 and Rouge-L metrics provide information regarding content preservation to avoid loss of information or inconsistent generation. From the table, we see that the *Seq2Seq* model has the lowest scores for both the metrics followed by the *HRED* approach. This indicates the inability of LSTM-based models to capture long-term information for a consistent and informative generation. In addition, the reinforced approach *Polite-RL* shows slight improvement compared to the basic models. The *PG-RL* framework gives better scores for both BLEU-4 and Rouge-L using pointer generator network that directly copies the words from the context. Our approach performs better for both the metrics with the ability of transformers to capture better representations from the dialogue context.

We also provide the politeness accuracy of the generated responses in the table. Our approach has an improvement of 4% and 8% compared to the *HT+RL+SE* and *PG-RL* frameworks, respectively. Even though the *HT+RL+SE* approach uses sentiment information for polite response generation, we outperform their approach with the help of affective tracker that captures the overall emotional content from the context for the consistent and relevant polite generation.

We also provide the ablation study of our approach in Table 2. The *HT* approach without the affective tracker and the behaviour-aware generators gives the lowest scores for all the metrics. By adding the affective tracker there is a gain in performance proving the significance of capturing the overall emotional quotient from the context. We also show the performance without the affective tracker, having only the behaviour-aware generators and the final polite generator. The performance is slightly better but still lower than the proposed approach indicating the importance of all the components in the proposed method.

Manual Evaluation Results: In Table 3, we provide the results of human evaluation on the CYCCD dataset. From the table, we see that the fluency scores of our proposed method is the highest compared to all the existing and the ablated models. The responses are fluent and grammatically correct making the responses complete.

In the case of the relevance metric, we gain an improvement of around 0.4% compared to the *HT+RL+SE* framework that uses sentiment information as feedback to generate consistent and relevant polite responses. We perform better as we employ the entire emotional content present in the context using the affective tracker for generating behaviour-aware polite responses that are in accordance with the context.

Finally, the politeness appropriateness scores are highest with an increase of 5% and 13% from the *HT+RL+SE* and *PG-RL* frameworks. This could be contributed to the behaviour-aware generators that assist in generating correct polite responses in accordance with the ongoing dialogue. This also ensures that the different polite behaviours are important to generate the correct and appropriate responses. The performance of the proposed approach is significantly better than all the existing approaches and the ablated models.

Case Study and Discussion: In Table 4, we

Context	*Hey, i got food poisoning from your inflight meal on Sunday*
BG$_{apology}$	We are sorry for the inconvenience, we would look into this.
BG$_{empathy}$	That's sad to hear, we would get you a refund for this.
BG$_{assurance}$	Don't worry, we will get in touch with the staff and help you with this.
Context	*Order 2 zinger box meals n got free popcorn chicken, yayyyy*
BG$_{greeting}$	Hello, good morning!!! Have a nice meal.
BG$_{appreciation}$	That's nice, hope you have a great meal ahead.
BG$_{assurance}$	We hope to surprise you again in the future, keep ordering.
Context	*I need the software update urgently, the battery lasts literally half a day*
BG$_{greeting}$	Good evening!!! Please DM us the information.
BG$_{apology}$	We apologize, but we will help you with the updates at the earliest.
BG$_{assurance}$	We are here to help! Tell us what you need.
Context	*Had a flight delay yesterday due to less ground staff, got credit points for the delay and free lunch.*
BG$_{greeting}$	Good evening!!! Have a good flight next time you fly with us.
BG$_{apology}$	We regret the inconvenience, by the flight delay, we will take better care of you in future.
BG$_{empathy}$	We are happy you got free food. Let us know if you need to book flights next time.

Table 4: Examples of polite responses in accordance to the polite behaviour

present a few examples of the generated response. From the table, we see that for a given context there are multiple plausible polite responses in accordance with the behaviour and nature of politeness. Also, the responses are consistent and relevant.

For the first example, "Hey, I got food poisoning from your inflight..." could be responded to in one of the three ways such as the agent could be apologetic about the situation, empathize with the customer or assure the user for a better outcome. In either of the above-mentioned behaviours, politeness is a key component and is well represented in the generated response showcasing the effectiveness of our proposed method. For the last example, "I need the software update urgently, the battery lasts.. ", the agent shows variations in polite behaviour and helps the user with the said problem either by just greeting, apologizing, or assuring the user. From the generated responses, it is clear that polite behaviour is significant for generating relevant, interactive, and better responses.

In addition, from the last example in Table 4, we see that the generator is capable of generating responses capturing the information in the context. This concludes the fact that the proposed approach is not only capable of variating politeness in responses but also inculcates informative knowledge from the context. We also perform an error analysis for the proposed approach to properly evaluate the performance of our approach in comparison to the baselines. Some of the errors encountered by the models are:

Loss of information: For a few responses, we see that the generated polite response lacks some of the information presents in the context making the response inconsistent with the ongoing dialogue. For example, the ground-truth response is "*The order from KFC has been taken and will be delivered soon*" while the generated response is "*We will look into this and call the service center*."

Polite-behaviour inconsistency: In some cases, we see that the responses fail to capture the correct polite behaviour and generate responses that are inconsistent with the context. For the context "*The fries were soggy and the burger was stale*" the generated response is "*Thank you for ordering, enjoy your meal*." The generated response is incorrect as the agent should either be apologizing for the order or assuring the customer that the feedback will be taken care of in the future.

6 Conclusion and Future Work

In our current work, we focus on the variations in politeness based on the different behavioural pattern linguistically present in polite phrases such as *greeting*, *apology*, *appreciation*, *assurance* and *empathy*. We design a transformer-based encoder-decoder network with three major components an affective tracker, behaviour-aware generators, and a polite generator to incorporate politeness in responses according to the context. Experimental analysis of the CYCCD dataset shows that the proposed approach effectively infuses the behaviour-wise polite phrases in the responses.

In the future, we intend to use unsupervised techniques to incorporate politeness in responses to create robust end-to-end systems. Also, the usage of politeness is important for goal-oriented conversations, therefore we plan to apply politeness to different domains.

Limitations

Our paper focuses on incorporating behaviour-aware politeness in responses according to the dialogue context for polite response generation. The primary limitation is the availability of labeled data for modeling the variations in politeness for the different user-related queries. Nevertheless, we employed crowd workers for the data annotation which even though is a time-consuming and costly process yet is the most reliable way of getting the data annotated. Due to the unavailability of the polite behaviour-annotated dataset, we conducted experiments using CYCCD dataset only. In future, we will attempt to extend the experiments to more task-oriented datasets. Also, because of limited computational resources in academia, we weren't able to conduct experiments using LLMs such as GPT2, GPT3, PaLM, LLaMa, etc.

Acknowledgements

Priyanshu Priya acknowledges the financial support provided by the Department of Science and Technology, Ministry of Science and Technology, Government of India, through the Innovation in Science Pursuit for Inspired Research (INSPIRE) Fellowship.

References

Jaime Cesar Acosta. 2009. Using emotion to gain rapport in a spoken dialog system.

Russell Beale and Chris Creed. 2009. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776.

Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1265–1274.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Stanley Chen, Douglas H. Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Isabel Dias, Ricardo Rei, Patrícia Pereira, and Luisa Coheur. 2022. Towards a sentiment-aware conversational agent. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–3.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 866–874.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Julia Fink. 2012. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International conference on social robotics*, pages 199–208. Springer.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020a. Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*, 13(3):1555–1566.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020b. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2021a. More the merrier: Towards multi-emotion and intensity controllable response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12821–12829.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020c. Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In

Proceedings of the 12th Language Resources and Evaluation Conference, pages 4172–4182.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022a. Polise: Reinforcing politeness using user sentiment for customer care response generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6165–6175.

Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. 2021b. Seprg: sentiment aware emotion controlled personalized response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 353–363.

Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2022b. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems*.

Mauajama Firdaus, Gopendra Vikram Singh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Affect-gcn: a multimodal graph convolutional network for multi-emotion with intensity recognition and sentiment analysis in dialogues. *Multimedia Tools and Applications*, pages 1–22.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2022c. Sentiment guided aspect conditioned dialogue generation in a multimodal system. In *European Conference on Information Retrieval*, pages 199–214. Springer.

Mauajama Firdaus, Naveen Thangavelu, Asif Ekba, and Pushpak Bhattacharyya. 2020d. Persona aware response generation with emotions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Mauajama Firdaus, Naveen Thangavelu, Asif Ekbal, and Pushpak Bhattacharyya. 2022d. I enjoy writing and playing, do you: A personalized and emotion grounded dialogue agent using generative adversarial network. *IEEE Transactions on Affective Computing*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*, pages 1–9.

Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Courteously yours: inducing courteous behavior in customer care responses using reinforced pointer generator network. 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 851–860.

Swati Gupta, Marilyn A Walker, and Daniela M Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *International Conference on Affective Computing and Intelligent Interaction*, pages 203–217. Springer.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *SIGDIAL Conference*, pages 64–73.

Sathish Reddy Indurthi, Dinesh Raghu, Mitesh M Khapra, and Sachindra Joshi. 2017. Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 376–385.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*, pages 895–906.

Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What can you do? studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 acm conference on designing interactive systems*, pages 264–275.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries.

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. *arXiv preprint arXiv:2004.14257*.

Avinash Madasu, Mauajama Firdaus, and Asif Eqbal. 2022. A unified framework for emotion identification and generation in dialogues. *arXiv preprint arXiv:2205.15513*.

Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological review*, 50(4):370.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022a. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing*, 494:242–254.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022b. Predicting politeness variations in goal-oriented conversations. *IEEE Transactions on Computational Social Systems*.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023a. Genpads: Reinforcing politeness in an end-to-end dialogue system. *Plos one*, 18(1):e0278323.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023b. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14408–14416.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023c. Pal to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271.

Tong Niu and Mohit Bansal. 2018a. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Tong Niu and Mohit Bansal. 2018b. Polite dialogue generation without parallel data. *TACL*, 6:373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. 2010. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, 13(1):49–60.

Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2023a. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Systems with Applications*, 224:120025.

Priyanshu Priya, Kshitij Mishra, Palak Totala, and Asif Ekbal. 2023b. Partner: A persuasive mental health and legal counselling dialogue system for women and children crime victims. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6183–6191. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8):434–441.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. *arXiv preprint arXiv:1804.10731*.

Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L Zhang. 2019. Learning to abstract for memory-augmented conversational response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3816–3825.

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

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020a. Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9177–9184.

Yi-Chia Wang, Alexandros Papangelis, Runze Wang, Zhaleh Feizollahi, Gokhan Tur, and Robert Kraut. 2020b. Can you be more social? injecting politeness and positivity into task-oriented conversational agents. *arXiv preprint arXiv:2012.14653*.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pages 4567–4573.