Courteously Yours: Inducing courteous behavior in Customer Care responses using Reinforced Pointer Generator Network

Hitesh Golchha*, Mauajama Firdaus*, Asif Ekbal, Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna, India
(hitesh,mauajama.pcs16,asif,pb)@iitp.ac.in

Abstract

In this paper, we propose an effective deep learning framework for inducing courteous behavior in customer care responses. The interaction between a customer and the customer care representative contributes substantially to the overall customer experience. Thus, it is imperative for customer care agents and chatbots engaging with humans to be personal, cordial and emphatic to ensure customer satisfaction and retention. Our system aims at automatically transforming neutral customer care responses into courteous replies. Along with stylistic transfer (of courtesy), our system ensures that responses are coherent with the conversation history, and generates courteous expressions consistent with the emotional state of the customer. Our technique is based on a reinforced pointer-generator model for the sequence to sequence task. The model is also conditioned on a hierarchically encoded and emotionally aware conversational context. We use real interactions on Twitter between customer care professionals and aggrieved customers to create a large conversational dataset having both forms of agent responses: generic and courteous. We perform quantitative and qualitative analyses on established and task-specific metrics, both automatic and human evaluation based. Our evaluation shows that the proposed models can generate emotionally-appropriate courteous expressions while preserving the content. Experimental results also prove that our proposed approach performs better than the baseline models.

1 Introduction

With the advancement of artificial intelligence (AI) and natural language processing (NLP), automatic systems have made immense impact on human lives by assisting them in their everyday works. Human-computer interaction is pervasive in many applications such as chatbots, personal assistants and many more. Natural language generation (NLG) component of such systems is an important aspect of every human computer interaction. Thus research in recent years have been on modulating biases, styles and control in text generation to enhance these interactions.

Customer care is an essential tool used by companies to provide guidance, assistance and in building stable customer relations. The ease of access, ease in following-up and immediacy of social media has made it a strong platform for companies and applications to interact with their customers. In this platform, we see the usage of courteous and emphatic language, which is the center of our current study. For the growth of any company or application it is necessary for the customer care agents to be cordial and amicable to the customer. Thus along with handling queries, it is important for agents to provide customer satisfaction by greeting, empathizing, appreciating feedback, apologizing at the right time, and thus build a strong relation with the customer. In Table 1, we showcase different situations in which an agent can behave courteously, thereby providing a good customer experience.

In this work, we focus on proposing an effective deep learning framework to enhance the existing NLG systems by converting their replies to courteous ones, by staying conversationally grounded, and emotionally aware of the user. For any Natural Language Generation (NLG) module (generic or task oriented), courteous response can play an important role in keeping the user engaged with the system. Also, it will make the system more human-like while generating responses. Inducing courteous behavior in responses can be fused with any existing NLG system to give them humanly essence and simultaneously make users

* First two authors are jointly the first authors

Proceedings of NAACL-HLT 2019, pages 851–860
Minneapolis, Minnesota, June 2 - June 7, 2019.

Generic	Courteous	Behaviour
How can we help?	*Help has arrived! We are sorry to see that you are having trouble, how can we help?*	Apology
Can you send us a screenshot of what you're seeing?	*Hey Craig, help's here! Can you send us a screenshot of what you're seeing?*	Greet
Let's discuss it in DM.	*We want to help. Let's discuss it in DM.*	Assurance
What is happening with your internet?	*Oh no that's not good. I can help! What is happening with your internet?*	Empathy
Enjoy your show while flying!	*Thanks for your kind words and enjoy your show while flying!*	Appreciation

Table 1: Examples of Courteous Responses

more comfortable in using these systems leading to an increase in user association with the brand or product. This would eventually lead to customer satisfaction with an increase in customer retention. Moreover, such language conditioning shall ensure that responses are more human-like. Thus, the major motivation behind this task is to create systems that are able to converse with humans efficiently and generate replies in accordance with the emotions of the customer. Courteousness is a virtue of humans and to be able to make a machine behave courteously is a challenging task.

Unlike a generic NLG system that focuses in generating responses, our system adds courteous nature and emotional sense to the replies, thereby, making the responses interesting and engaging to the users. Such systems have high applications in many areas/companies that employ chatbots to deal with the customers. We thus propose a novel research direction of inducing courteous behavior in the natural language responses for the customer care domain whilst being contextually consistent. The key contributions of our work are summarized as follows:

(i) Creation of a high quality and a large conversational dataset, Courteously Yours Customer Care Dataset (CYCCD) prepared from the actual conversations on Twitter. We provide both forms of agent responses: generic and courteous.

(ii) Proposal of a strong benchmark model based on a context and emotionally aware reinforced pointer-generator approach which demonstrates very strong performance (both on quantitative and qualitative analyses) on established and task-specific metrics, both automatic and human evaluation based.

The rest of the paper is structured as follows: In section 2, we discuss the related works. In Section 3 we explain the proposed methodology followed by the dataset description in section 4. Experimental details, evaluation metrics and results are presented in section 5 and 6 respectively. In section 7, we present the concluding remarks followed by future directions.

2 Related Work

Natural language generation (NLG) module has been gaining importance in wide applications such as dialogue systems (Vinyals and Le, 2015; Shen et al., 2018; Wu et al., 2018; Serban et al., 2017a; Raghu et al., 2018; Zhang et al., 2018; Li et al., 2016), question answering systems (Reddy et al., 2017; Duan et al., 2017), and many other natural language interfaces. To help the users achieve their desired goals, response generation provides the medium through which a conversational agent is able to communicate with its user. In (Serban et al., 2017b), the authors have proposed an hierarchical encoder-decoder model for capturing the dependencies in the utterances of a dialogue. Conditional auto-encoders have been employed in (Zhao et al., 2017), that generates diverse replies by capturing discourse-level information in the encoder. Our work differentiates from these previous works in dialogue generation in a way that we embellish the appropriate response content with courteous phrases and sentences, according to the conversation. Hence, our system is an accompaniment to any standalone natural language generation system to enhance its acceptability, usefulness and user-friendliness.

Emotion classification and analysis (Herzig et al., 2016) in customer support dialogue is important for better understanding of the customer and to provide better customer support. Lately, a number of works have been done on controlled text generation (Hu et al., 2017; Li et al., 2017; Subramanian et al., 2017; Fedus et al., 2018; Peng et al., 2018) in order to generate responses with desired attributes. Emotion aware text generation (Zhou and Wang, 2018; Zhou et al., 2018; Huang et al., 2018) have gained popularity as it generates responses depending on a specific emotion. Previous works in conditioned text generation have worked on inducing specific biases and behaviors (Herzig et al., 2017) while generation (like emotion, style, and personality trait). Our work is different in the sense that it can encompass different emotional states (like joy, excitement, sad-

ness, disappointment) and traits (like friendliness, apologetic, thankfulness, empathy), as is the demand of the situation.

Style transfer has been an emerging field in natural language processing (NLP). A couple of works have been done in changing the style of an input text and designing the output text according to some particular styles. In (Rao and Tetreault, 2018), a dataset has been introduced for formality style transfer. Unsupervised text style transfer has encouraged in transforming a given text without parallel data (Shen et al., 2017; Carlson et al., 2017; Fu et al., 2018; Li et al., 2018; Niu and Bansal, 2018). Overall our system is novel as it is motivated by the need for inducing specific behavior and style in an existing NLG systems (neural, or template-based) as a means of post editing, by simultaneously being emotionally and contextually consistent. We have successfully demonstrated this behavior through empirical analysis for a specific application of customer care.

3 Methodology

Given the Conversation History (previous few exchanges in the dialog), and the Generic Response, the task is to generate the Courteous Response. The architectural diagram of our proposed model is in Figure 1.

3.1 Conversational History Representation

The conversation history C is a sequence of utterances $(u_1, u_2, \ldots, u_D)$ and each utterance u_d is a sequence of words $w_1, w_2, \ldots, w_N$ which are represented by their embeddings. For encoding the emotional states associated with these utterances, we use the output distribution from DeepMoji (Felbo et al., 2017) which is pre-trained on the emoji prediction task.

Let the utterance u_d be a sequence of sentences $s_1, s_2, \ldots, s_N$, where the n^{th} sentence has an emotional embedding $e_{n,d}$. Then the emotional representation of the utterance is:

$$e_d[i] = \max_n e_{n,d}[i] \tag{1}$$

The first bi-directional layer over any utterance u_d yields the hidden states $h^1_{1d}, h^1_{2d}, \ldots, h^1_{Nd}$, where N is the word length of the utterance. The final representation of any utterance r_d is given by the concatenation of the emotional representation as well as the last hidden state of the Bi-directional Long Short Term Memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) encoder.

$$r_d = [e_d \cdot h^1_{Nd}] \tag{2}$$

The second hierarchical layer Bi-LSTM encodes the utterance representations $r_1, r_2, \ldots, r_D$ as hidden states $h^2_1, h^2_2, \ldots, h^2_D$. The last hidden state h^2_D is the representative of the conversational history, and is renamed as the conversational context vector c.

3.2 Encoder states

Another single layer unidirectional LSTM network encodes the generic response word embedding sequence to obtain the encoder hidden states h_i.

3.3 Decoder states and Attention calculation

At the decoder time step t, the decoder LSTM state s_t is used to calculate the attention distribution over the encoder states α^t:

$$e^t_i = v^T \tanh(W_h h_i + W_s s_t + b_{\text{attn}}) \tag{3}$$
$$\alpha^t = \text{softmax}(e^t) \tag{4}$$

where v, W_h, W_s and b_{attn} are trainable parameters.

This attention distribution helps to identify the relevant encoder states necessary for the current decoding step. The representation of the encoder for this time step is an attention weighted sum of its states, called the context vector:

$$h^*_t = \sum\nolimits_i \alpha^t_i h_i \tag{5}$$

The LSTM state s_t is updated using s_{t-1}, the previous time step's context vector h^*_{t-1}, word embedding of the previously generated word $w_{emb}(y_{t-1})$, and the conversation context vector c.

$$s_t = LSTM(s_{t-1}, W_p[w_{emb}(y_{t-1}), h^*_{t-1}, c] + \tilde{b}) \tag{6}$$

where, W_p and $\tilde{b}$ are the trainable parameters.

3.4 Output distribution calculation

To aid the copying of words from the generic response while generating the courteous response, we use the mechanism similar to (See et al., 2017). For the pointer generator network, the model computes two distributions, one over the

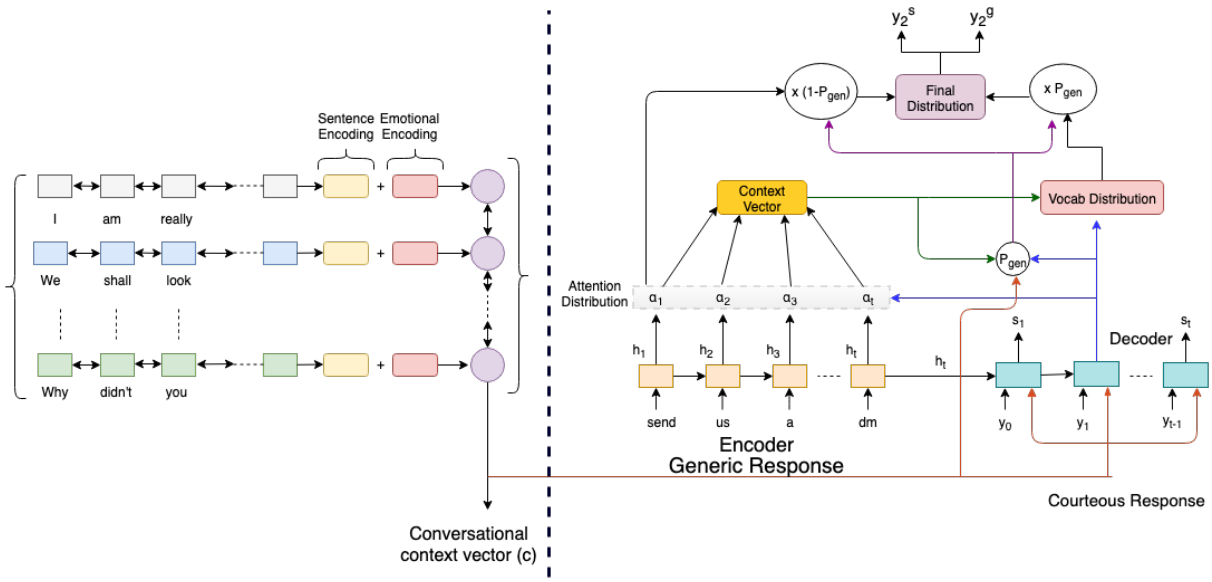

Figure 1: Architectural Diagram of the Proposed Model. Inputs to the model: Conversation History (left), Generic Response (centre) Output: Courteous Response (right). The Conversation History is encoded by hierarchical BiLSTM to a Conversational Context vector c. The encoder encodes the Generic Response into hidden states h_i. Response tokens are decoded one at a time. Attention α_i, and vocabulary distributions (p_{vocab}) are computed, and combined using p_{gen} to produce output distribution. Sampling it yields y_i^s and taking its argmax yields y_i^g.

encoder words (α^t) and one over the vocabulary (p_{vocab}).

$$p_{vocab} = \text{softmax}(W'(W[s_t, h_t^*] + b) + b') \quad (7)$$

where W, W', b and b' are the trainable parameters.

The trade-off between the two distributions is computed dynamically in the form of the *generation probability* $p_{gen} \in [0, 1]$ from the context vector h_t^*, the decoder state s_t, the decoder input x_t, and conversational context vector c:

$$p_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + w_c^T c + b_{\text{gen}}) \quad (8)$$

where vectors w_{h^*}, w_s, w_x, w_c and scalar b_{gen} are trainable parameters and σ is the sigmoid function.

The final distribution over the union of the vocabulary words and the words of the generic response is calculated by:

$$P(w) = p_{gen} p_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} \alpha_i^t \quad (9)$$

3.5 Model training

We use the joint reinforcement learning (RL) and machine learning (ML) training as used in (Paulus et al., 2017). If $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{n'}\}$ is the gold output tokens for given generic response tokens x_1 and conversation history x_2, the maximum-likelihood objective using teacher forcing is given by:

$$L_{ML} = -\sum_{t=1}^{n'} \log p(\tilde{y}_t | \tilde{y}_1, \ldots, \tilde{y}_{t-1}, x_1, x_2) \quad (10)$$

Along with training with the maximum likelihood error, we also use reinforcement learning to learn from maximizing discrete metrics that are task specific (which we design as the rewards). We use the self-critical policy gradient algorithm proposed in (Rennie et al., 2017).

Here the REINFORCE (Williams, 1992) algorithm is baselined with the reward obtained by the inference time algorithm (which performs greedy decoding), without the need for training a critic network for estimating value functions. During training, two output sequences are produced: y^s, obtained by sampling $p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$ probability distribution, and y^g, the baseline output, obtained by greedily maximizing the output probability distribution at each time step.

$$L_{RL} = (r(y^g) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x_1, x_2) \quad (11)$$

Our reward function $r(y)$, used for evaluating y against the gold standard output is

$$r(y, \tilde{y}) = \lambda_1 \cdot m1(y, \tilde{y}) + \lambda_2 \cdot m2(y, \tilde{y}) \quad (12)$$

It is the weighted mean of the two terms:

(i) BLEU metric $m1$: Ensures the content matching between the reference and the decoded outputs.

(ii) Emotional accuracy $m2$: Measured by the cosine similarity of the emoji distributions of the gold and generated responses (using pretrained DeepMoji). It ensures that the emotional states of the generated courteous behavior is consistent with the gold.

We first pre-train using the maximum likelihood (ML) objective (Eq. 10) and then using a mixed objective function with a reduced learning rate:

$$L_{mixed} = \eta L_{RL} + (1 - \eta) L_{ML}, \quad (13)$$

3.6 Baselines

We develop the following models:

1. **Model-1**: This is a Seq2Seq model with attention (Luong et al., 2015) and decoder conditioned on the conversational context vector c (without concatenating emotional embedding i.e. instead of Eq. 2, $r_d = h^1_{Nd}$)

2. **Model-2**: This model is developed using Model-1 along with the copying mechanism of Pointer Generator Network.

3. **Model-3**: This model is developed using Model-2 along with emotional embeddings in the conversational context vector as in E.g., 2.

	Train	***Valid***	***Test***
# Conversation	140203	20032	40065
# Utterances	179034	25642	51238

Table 2: Dataset Statistics

4 Dataset

In this section we describe the details of the dataset that we create for our experiments.

4.1 Dataset source and description

We use the data of the interactions between customers and professional customer care agents of companies on their Twitter handles. We source the requisite Twitter data from the dataset made available on Kaggle by Thought vector [1]. Tweets have labels of company names, anonymized user ids, time stamps, and response tweet ids - essential for reconstructing conversations, and nuanced analyses. We filter out conversations having multiple responses to a single tweet, and those starting by a tweet by a company. This was done to ensure correct conversation flow and to acquire suggestion / complaint based exchanges, respectively.

[1] https://www.kaggle.com/thoughtvector/customer-support-on-twitter

4.2 Process for data creation

As there exists no dataset with generic and courteous versions of utterances we create our own dataset. We prepare responses of generic styles by filtering out courteous sentences, phrases and expressions from the actual responses. We presume actual responses as the courteous form of response.

An example conversation:

Customer utterance (conversation history): *y'all just came to my house like last week and I'm having problems with my internet again smh*

Tweet by the Customer Care professional: *Oh no that's not good. I can help! What is happening with your internet?*

We use this conversation to prepare the courteous and the generic response

1. **Courteous response**: *Oh no that's not good. I can help! What is happening with your internet?*

2. **Generic response**: *What is happening with your internet?*

As we want to filter out courteous phrases / sentences from a given customer care tweet, we segment the tweet into sentences. Purely courteous (and non-informative) sentences must be removed, purely informative sentences must be retained, and informative sentences with courteous expressions must be transformed (to remove only the courteous part from the sentence). We define these three forms of sentences as:

(i) *Courteous sentences*: Sentences which do not contain any information/ suggestions, and are purely non-informative. These may include personalized greetings and expression of appreciation, apology, empathy, assurance, or enthusiasm. Example: *Sorry to hear about the trouble!*

(ii) *Informative sentences without courteous expressions*: These sentences contain the actual content of the tweet and are generally assertions, instructions, imperatives or suggestions. Example: *Simply visit url_name to see availability in that area!*

(iii) *Hybrid-Informative sentences with courteous expressions*: These are the sentences of the second type also containing some expressions of the first type. Example: ***We appreciate the feedback,*** *we'll pass this along to the appropriate team.*

4.3 Scaling up for large data creation

We annotate sentences in isolation by grouping similar sentences together to speed up annotations and then reconstruct the generic sentences by post-processing rules. We follow the following procedure to prepare the dataset for each company separately:

1. *Sentence segmentation*: We first extract the tweets from customer care agents. Each tweet is segmented into sentences to eventually identify three forms of the sentences.

2. *Clustering*: As expressions and sentences used by professionals of a company often follow similar patterns. Grouping similar sentences together before annotation would therefore significantly make the process faster. The vector-semantic representations of sentences are obtained using the sentence encoder(Conneau et al., 2017) trained on the SNLI corpus(Bowman et al., 2015). We use the K-Means clustering(Aggarwal and Zhai, 2012)(k = 300) to cluster these sentences.

3. *Annotations*: Three annotators proficient in the English language were assigned to annotate the sentences into the three categories: *purely courteous, purely informative, hybrid*. For sentences having both informative and courteous clauses/expressions (hybrid), they were asked to manually prepare the generic sentence by removing the courteous part. Also they were asked to identify non English conversations (and filter them). We observe the multi-rater Kappa agreement ratio of approximately 80%, which may be considered as reliable.

4. *Preparing generic responses*: Now let us assume we have a courteous response S with n sentences $s_1, s_2, \ldots, s_n$. We obtain the generic response by removing the courteous sentences, retaining the informative sentences, and replacing the hybrid sentences with the prepared generic equivalents.

We divide the conversation into train, validation and test sets as given in Table 2. Each training example is of the form: conversational history (last three utterances), generic response and courteous response.

5 Experiments

Implementation Details: We use a vocabulary of size 30k for the task (as the range of courteous expressions is limited, and informative contents can be copied even if they are out-of-vocabulary-OOV). We use 256 dimensional hidden states and 128 dimensional word embeddings (not pre-trained). We use AdaGrad as the optimizer with gradient clipping (magnitude 2). We train with batches of size 16, and use the same size for beam search decoding. We monitor smoothened running loss on the validation set for early stopping and finding the best models for decoding. We use η = 0.99 (similar to (Paulus et al., 2017)) for the joint loss. For the reward function the values of λ_1 and λ_2 are 0.75 and 0.25, respectively.

Automatic Evaluation: For automatic evaluation, in addition to the standard metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and perplexity, we also use two task-specific metrics:

1. Content preservation (CP): We want to measure how much of the informative content from the original generic response(X) is reflected in the generated courteous response(Y). We use a measure similar to ROUGE-L recall.

$$CP = LCS(X, Y)/len(X) \tag{14}$$

where LCS is the longest common subsequence.

2. Emotional accuracy (EA): To measure the consonance between the generated courteous expressions (source of emotion) and the gold, we find the cosine similarity between the MojiTalk emoji distributions of the two responses (X_e and Y_e).

$$EA = X_e \cdot Y_e/(|X_e||Y_e|) \tag{15}$$

Human Evaluation: In order to understand the quality of the responses we adopt human evaluation to compare the performance of different models. We randomly sample 500 responses from the test set for human evaluation. Given a generic response along with conversation history, three human annotators with post-graduate exposure were assigned to evaluate the courteous responses generated by the different models for the three metrics:

1. Fluency (F): The courteous response is grammatically correct and is free of any errors.

2. Content Adequacy (CA): The generated response contains the information present in the generic form of the response and there is no loss of information while adding the courteous part to the responses.

3. Courtesy Appropriateness (CoA): The courtesy part added to the generic responses is in accordance to the conversation history.

The scoring scheme for fluency and content adequacy is 0: incorrect or incomplete, 1: moderately correct, 2: correct, whereas for courtesy appropriateness the scoring scheme is -1: inappropriate, 0: non-courteous, 1: appropriate, respectively. We computed the Fleiss' kappa (Fleiss, 1971) for the above metrics to measure inter-rater consistency. The kappa score for fluency is 0.75 and courtesy appropriateness is 0.77 indicating "substantial agreement" and the score is 0.67 for content adequacy denoting "considerable agreement".

6 Results and Analysis

Automatic evaluation results: Results of the different models are presented in Table 3. The proposed model performs significantly better than the other baselines for all the evaluation metrics and the improvement in each model is statistically significant compared to the other models.[2]. The attention based sequence to sequence model (Model 1) is a decent baseline with good scores (56.80 BLEU). The Pointer generator model (Model 2) is aided by the copying mechanism. Thus, it is better modeled to include portions of the content from the generic response into the courteous response. This is corroborated by the increased score in CP (+9.33%). Its emotional accuracy is slightly reduced from Model 1 (-0.45%), probably because of eagerness to copy rather than generate. The advantage of the emotional embedding in Model 3 over Model 2 is reflected with the increased scores(+3.77%), because of its ability to better understand the emotional states and generate more appropriate courteous responses. The perplexity values are slightly reduced in Model 3 and Model 4, apparently because of the emotion embedding confusing the actual content from the conversation history. The final model performs decently better than other models. The reinforcement learning objective helps it to improve upon the desired metrics rather than just learn to be accurate at the token level.

[2]we perform statistical significance tests (Welch, 1947) and it is conducted at 5% (0.05) significance level

	Model	BLEU	ROUGE			PPL	CP	EA
			1	2	L			
1	***Seq2Seq***	56.80	63.8	59.06	64.52	58.21	68.34	82.43
2	***Seq2Seq + P***	66.11	69.92	64.85	66.40	**42.91**	**77.67**	81.98
3	***Seq2Seq + P + EE***	68.16	72.18	67.92	71.17	43.52	76.05	85.75
4	***Proposed Model***	**69.22**	**73.56**	**69.92**	**72.37**	43.77	77.56	**86.87**

Table 3: Results of various Models; P: Pointer Generator Model; EE: Emotional embedding

Model	F			CA			CoA		
	0	1	2	0	1	2	-1	0	1
Model 1	15.70	42.50	41.80	16.21	41.69	42.10	23.71	51.08	25.21
Model 2	14.23	42.77	43.00	15.62	39.65	44.73	22.05	39.43	38.52
Model 3	11.15	44.10	44.75	13.66	41.12	45.22	15.23	41.22	43.55
Our Model	10.05	44.90	44.60	13.85	38.48	47.67	14.11	41.11	44.78

Table 4: Human evaluation results for Fluency, Content Adequacy and Courtesy Appropriateness (All values are in percentages.)

Human evaluation results: In Table 4, we present the results of human evaluation. In case of fluency, our proposed model and the third model show similar performance, whereas Models 1 and 2 are relatively less fluent. Model 2 shows great improvement with respect to Model 1 as it is able to copy the content from the input. Also, for content adequacy our proposed model has been able to generate 38.48% moderate replies that have adequate amount of information in it while it generates around 47.67% correct responses that contain all the information present in the input. For courtesy appropriateness, Model 1 and Model 2 show lower performance while our proposed model has been able to capture the courteous behavior. As score 1 is given to the responses that are courteous as well as the nature of courteousness is in accordance to the conversation, it can be seen that our model achieves 44.78% performance level which is higher than the other models. From this evaluation, we can infer that the responses generated by our model are not only adequate in terms of information preservation, but also able to induce the courteous behavior by making the responses interesting and informative.

Error Analysis: We further analyse the outputs generated from our proposed model to perform a detailed qualitative analysis of the responses. In Table 5, we present few examples of the responses generated by the different models given the generic input. Some common forms of mistakes include:

1. Unknown Tokens: As Model 1 does not have the copying mechanism, the number of unknown

Generic Input	Model 1	Model 2	Model 3	Our Model
dm us more info and well take a look into it for you	we'll look into it	im sorry to hear this please dm us more info and we'll take a look into it for you	were here to help please dm us more info and well take a look into it for you	were here to help please dm us more info and well take a look into it for you at the earliest
adjust the brightness via your display settings on your device	whos the brightness via your display settings on your device	were here to help adjust the brightness via your display settings on your device	we have several ways to change the display brightness on your device and were happy to help	thanks for reaching out we have several ways to change the display brightness on your device and were happy to help
we'll follow up with the store	we'd like to help well follow up	were here to help well follow up with the store	sorry to hear that well follow up with the store	thats disappointing to hear, we'll follow up with the store
can you confirm which platform you are using for video access ? what is the error ?	what is the error ?	I am sorry for the frustration ! can you confirm which platform you are using for video access ? what is the error ?	I am sorry to hear this can you confirm which platform you are using for video access? what is the error?	I am sorry for any frustration, can you please confirm which platform you are using for video access? Please tell us what is the error.
fill this form <url>	please fill this form <url>	were here to help fill this form <url>and I'll contact you at the earliest a	apologies for the hassle, please fill this form <url>and we'll contact you thank you for reaching out to us we will follow up with the store	i am sorry for the hassle, please fill this form <url> and ill contact you at the earliest

Table 5: Examples of Courteous Responses Generated by the Different Models

tokens is predicted the most in this. Also often the model predicts 'end of sequence' token just after the 'out of vocabulary' token, thus leaving sequences incomplete.

2. Wrong copying: Sometimes pointer network makes mistakes while copying (being influenced by language model): Gold: .. *which store in gillingham did you visit ?*; Predicted: .. *which store in belgium did you visit ?*

3. Mistakes in emotion identification: These mistakes are more prominent in Models 1 and 2 (they don't have emotional embeddings), where the generated courteous phrases denote mistakes in identifying the emotional state of the customer. For example, Gold: *you're very welcome, hope the kids have an amazing halloween !*; Predicted: *we apologize for the inconvenience. hope the kids have an amazing halloween !*

4. Extra information: Models 1, 2, 3 sometimes generate extra informative sentences than in the generic response: Gold: *please send us a dm*; Predicted: *please send us a dm please let us know if you did not receive it*

5. Contextually wrong courteous phrases: These mistakes are common across models while generating courteous phrases with content in them: Gold: *we want to help, reply by dm and ..*; Predicted: *im sorry you havent received it. please reply by dm and ..*

6. Difference in phrases: Generated responses differ from reference responses in their use of (equivalent) courteous phrases, and are hence wrongly penalized by some metrics.

7 Conclusion and Future Work

In this paper, we propose a new research problem of accentuating customer care responses with courteous behavior. Incorporation of courteousness is important for attaining user satisfaction and to improve the performance of the application leading to user retention. We successfully prepare a large benchmark corpus, created from the actual showcasing of courteous behavior by human professionals on Twitter. Our developed models appropriately model the dialogue history and are informed of the past emotional states through emotional embeddings. We have used both automatic and human based metrics for evaluating the performance of our model. In qualitative and quantitative analyses of the generated responses, we observe contextually correct courteous behavior and content preservation, along with minor inaccuracies as discussed in the error analysis section. Overall the performance of our model shows the variations in responses with the other models keeping the information and courtesy nature of the generated responses intact.

In future, along with the opportunity of extending the architectural designs and training methodologies to enhance the performance of our systems, we look forward to designing a specific component to enhance the natural language generation component of an end to end chatbot, by including appropriate mechanisms to interact with all its components (memory, database, and the dialog manager). Moreover, studies will be conducted on courtesy transfer for the other domains, and also transfer learning from one domain to the another (like customer care to hospitality).

Acknowledgement

Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Min-

istry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Charu C Aggarwal and ChengXiang Zhai. 2012. *Mining text data*. Springer Science & Business Media.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2017. Zero-shot style transfer in text using recurrent neural networks. *arXiv preprint arXiv:1711.04731*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

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.

William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: Better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*.

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)*.

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

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

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 *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 64–73.

Jonathan Herzig, Michal Shmueli-Scheuer, Tommy Sandbank, and David Konopnicki. 2017. Neural response generation for customer service based on personality traits. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 252–256.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *ICML*.

Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 49–54.

Jiwei Li, Will Monroe, Alan Ritter, Daniel Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *EMNLP*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Daniel Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*.

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

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Tong Niu and Mohit Bansal. 2018. 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 *ACL*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

Yehong Peng, Yizhen Fang, Zhiwen Xie, and Guangyou Zhou. 2018. Topic-enhanced emotional conversation generation with attention mechanism. *Knowledge-Based Systems*.

Dinesh Raghu, Nikhil Gupta, et al. 2018. Hierarchical pointer memory network for task oriented dialogue. *arXiv preprint arXiv:1805.01216*.

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)*, volume 1, pages 129–140.

Sathish Reddy, 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*, volume 1, pages 376–385.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C Courville. 2017a. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*, pages 3288–3294.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

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.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *AAAI*.

Sandeep Subramanian, Sai Rajeswar, Francis Dutil, Chris Pal, and Aaron Courville. 2017. Adversarial generation of natural language. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 241–251.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Xianchao Wu, Ander Martinez, and Momo Klyen. 2018. Dialog generation using multi-turn reasoning neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2049–2059.

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.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *ACL*.