# Incorporating Politeness across Languages in Customer Care Responses: Towards building a Multi-lingual Empathetic Dialogue Agent

**Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya**
Indian Institute of Technology Patna
Bihta, Bihar-801106, India
{mauajama.pcs16, asif, pb}@iitp.ac.in

**Abstract**

Customer satisfaction is an essential aspect of customer care systems. It is imperative for such systems to be polite while handling the customer requests or demands. In this paper, we present a large multi-lingual conversational dataset for English and Hindi. We choose data from Twitter having both generic and courteous responses between customer care agents and aggrieved users. We also propose strong baselines that can induce courteous behaviour in generic customer care response in a multi-lingual scenario. We build a deep learning framework that can simultaneously handle different languages and incorporate polite behaviour in the customer care agent's responses. Our system is competent in generating responses in different languages (here, English and Hindi) depending on the customer's preference and also is able to converse with humans in an empathetic manner to ensure customer satisfaction and retention. Experimental results show that our proposed models can converse in both the languages and the information shared between the languages helps in improving the performance of the overall system. Qualitative and quantitative analysis show that the proposed method can converse in an empathetic manner by incorporating courteousness in the responses and hence increasing customer satisfaction.

**Keywords:** Courteous, Response generation, Deep Learning, Multilingualism

## 1. Introduction

With the growth and progress in artificial intelligence (AI) along with natural language processing (NLP), dialogue systems have made a huge impact on humans by helping them in their daily works. Dialogue systems are perfect examples of human-computer interactions. Such systems are highly prevalent nowadays in the form of chatbots and personal assistants like Apple's Siri, Microsoft's Cortana, Amazon's Alexa, and many more.

The primary objective of these automated applications is to assist humans and help them in the smallest possible ways as humanly as possible. Natural language generation (NLG) module of every dialogue system is an essential component as it presents the information to the user. To enhance the interactions between human and computers, recently researchers have focused on adapting different styles, emotions and personalities in text generation. Recent research has been inclined to make the system understand different languages giving rise to multi-lingual applications. With the significant focus on making the computer understand different languages as in (Masumura et al., 2018b; Masumura et al., 2018a; Upadhyay et al., 2018), researchers are aiming to make the dialogue systems language invariant as in real-world scenario. Also, for more extensive applications, it is crucial for these systems to be able to converse with humans in their preferred language, thereby increasing the usage and advancement of technology.

Providing assistance to the customer through social media channels is attaining high popularity. The centre of our current work focuses on incorporating politeness in customer care responses belonging to different languages. Due to the unavailability of large scale data for Hindi language, we create a conversational dataset from Twitter belonging to different companies having generic and polite/courteous responses for customer care agents. In this work, we aim at inducing courteous and empathetic phrases and modify the existing generic customer care responses for English and Hindi languages using adversarial training that helps in learning language invariant features. For better efficiency and proliferation of any organization, it is essential for the customer care agents to understand the different languages the users choose to converse in and hence be friendly and amiable to the customers.

While handling user queries, it is imperative for the customer care agents to provide customer satisfaction by acknowledging different situations the customers face with any company or application. For example, empathizing with the customers when they are in problem, apologizing when at fault, greeting and appreciating feedback, thereby ensure strong customer relations. In Table 1, we provide different use-cases for both the languages where the customer care agent can behave politely safeguarding better customer experience.

The primary objective of this work is to present an efficient deep learning framework that can generate polite customer care responses in both Hindi and English languages, enhancing the performance and usability of the existing Natural Language Generation (NLG) systems by being conversationally coherent and aware of customer's emotional state. For either goal-oriented or open domain (chit-chat) conversational agents, the polite response is a significant advantage for the NLG

| Generic Response | Polite Response | Behaviour |
|---|---|---|
| Provide your booking info via dm. | We're here to help you, please provide your booking info via dm. | *Assurance* |
| हम मामले पर गौर करेंगे। (We will look into the matter.) | यह सुनने के लिए निराशाजनक है, कृपया धैर्य रखें जब तक कि हम मामले को न देखें। (That's disappointing to hear, please have patience until we look into the matter.) | *Empathy* |
| Our team is working on getting your bags. | We're sorry for the extended travel time, our team is working hard on getting your bags, please have patience. | *Apology* |
| What are you looking for? | Hey good evening, good to have you with us, please tell what are you looking for? | *Greet* |
| हमने सभी जानकारी प्रदान की है। (We have provided all the information.) | हमारी सेवाओं का उपयोग करने के लिए धन्यवाद, हमने सभी जानकारी प्रदान की है। (Thanks for using our services, we have provided all the information.) | *Appreciation* |

Table 1: Examples of polite responses for both Hindi and English languages

module for making the system human-like and increasing user interactions. The ability to respond politely in any given language can be incorporated in any NLG application for providing humanly essence to the system and making it more comfortable for the users. Thus, the primary motivation of this work is to develop systems that can converse with humans in their preferred language by making the responses polite and courteous, eventually leading to user satisfaction and high customer retention for any given brand or company.

The ability of such systems to understand the emotions of the users in different languages and responding in accordance with the emotion is a challenging task. Also, politeness is a virtue of humans, and to make a machine understand and behave amicably and courteously is an additional task for such systems. Hence, in this work, we propose a large-scale Hindi dataset for this task and evaluate using the baseline approach of (Golchha et al., 2019) to incorporate politeness in customer care responses belonging to different languages and providing new research directions for showcasing the differences in politeness and courteous behavior across the languages.

We summarize the key contributions as follows:

**(i)** We create a large-scale Hindi conversational data, prepared from the actual conversations on Twitter.

**(ii)** We propose a robust response generation model for both Hindi and English languages by modeling the conversational history and the emotional state of the user by learning language invariant representation using adversarial training.

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

## 2. Related Work

Natural language generation (NLG) module provides a platform to conversational agents through which they can communicate with the users, thereby assisting them in achieving their desired objectives. Natural language generation is one of the core components of every dialogue system (Shen et al., 2018; Vinyals and Le, 2015; Wu et al., 2018; Serban et al., 2017a; Serban et al., 2017b; Zhao et al., 2017; Zhang et al., 2018). The authors in (Li et al., 2016) proposed a reinforcement learning-based approach for generating interesting, diverse and coherent dialogues. While the authors in (Raghu et al., 2018) employed a hierarchical pointer generator memory network for generating responses by handling out-of-vocabulary (OOV) words.

In this work, we make the responses more engaging by incorporating politeness in them, thereby differentiating it from the existing NLG systems. Hence, our system can add value to these existing NLG systems by making it polite, diverse and interesting. Therefore, it improves its usability and enhances its growth in terms of customer retention.

Recently, emotion classification in conversations (Majumder et al., 2018; Herzig et al., 2016) has been an interesting research area, which aims at making the system aware of different human emotions. Specifically, in customer support systems, it is crucial to understand the feelings of the user for providing proper assistance to them as investigated in (Herzig et al., 2016). Generating emotional responses (Zhou and Wang, 2018; Zhou et al., ; Huang et al., 2018) has been addressed in the past to give the systems humanly essence. Unlike the existing emotional response generation systems where emotions are explicitly provided, in our work we model the customers' emotions through conversational history and provide polite responses by being emotionally aware of the users' emotional state.

Lately, style transfer has been a growing research area with several works done in incorporating specific styles in the output texts which is different from the input texts (Carlson et al., 2017; Li et al., 2018a; Shen et al., 2017; Niu and Bansal, 2018; Fu et al., 2018) in an unsupervised fashion. The authors in (Golchha et al., 2019) proposed a reinforced pointer generator network for inducing courteous behavior in customer care responses. Also, there is a recent shift in building systems that are capable of understanding different languages (Li et al., 2018b; Do and Gaspers, 2019; Masumura et al., 2018a), hence making conversational agents robust in their applications.

In this work, we propose a novel system that is ca-

pable of generating polite responses in different languages (in our case, Hindi and English) by learning language invariant features through adversarial training and modeling the differences in politeness across the languages.

## 3. Dataset

Customer satisfaction is the ultimate goal of every human-machine interactions hence, making the machine behave humanly is an important aspect of such systems. Courteousness is a virtue of humans that they imbibe in conversations. Polite behavior is imperative while dealing with customers and providing responses to them. With the recent shift in research to make the dialogue systems language invariant, it is important to have datasets of different languages. Our goal is to incorporate courteous behavior in a customer care response belonging to different languages. Due to the unavailability of multilingual conversational datasets involving less resource language like Hindi, we prepare a large dataset for the Hindi language from Twitter conversations belonging to different companies. We then focus on polite response generation for Hindi and English languages. The complete details of the prepared dataset are given below.

### 3.1. Hindi conversational dataset

As there does not exist any Hindi conversational dataset[1], we create our dataset for this particular task for the experiments. We follow the guidelines of the English dataset (Golchha et al., 2019) and similarly prepare the Hindi dataset. For preparing the Hindi dataset, we mine the real user interactions from Twitter in Hindi. We collected Hindi conversations between users and customer care agents for different companies. The statistics for both Hindi and English datasets are provided in Table 2. We prepare the generic responses by filtering out polite expressions, phrases and sentences from the actual responses.

#### 3.1.1. Data description and source:

The conversations between various customers and the trained customer care agents of different companies on their Twitter handles were used for building the dataset. The Twitter data for Hindi was mined from twitter for different companies. To make a large-scale dataset we also translated some conversations from the Twitter data made available on Kaggle using Google translator. The translated conversations have been manually checked by three human translators proficient in the Hindi language. The inter-annotator agreement for the translated conversations was observed to be more than 90% indicating correct translation of the conversations in the Hindi language. The tweets in the dataset consist of the company names, unidentified user-ids along with time stamps and ids of the response tweets which are important for conversation reconstruction and proper analysis. We extract the conversations that begin with the company tweets and also filter out conversations having numerous responses to a single tweet. It is done to ensure the correct conversation flow and to retain exchanges based on suggestions/complaint, respectively.

#### 3.1.2. Data information:

In the absence of a generic and courteous version of utterances in a conversation, we create our own dataset having both the generic information type utterances and the courteous utterances. The generic response utterances are prepared by removing the polite and courteous words, phrases, expressions and sentences from the actual response utterance. We assume that the actual responses to be the courteous form of the responses.

An example of the conversation from the dataset is shown below:

**Customer Utterance (Conversational Context):**
मेरा आई फोन एक मिनट पहले पूरी तरह से काम कर रहा था और फिर अचानक काम करना बंद कर दिया।

(So my iPhone was perfectly working a minute back and then it stopped working just out of nowhere.)

**Customer care agent response utterance:**
मुझे क्षमा करें, यह अच्छा नहीं है। आपको क्या परेशानी हो रही है? हम आपके साथ इस पर गौर करेंगे और देखेंगे कि क्या हो रहा है।

(I am sorry, that's not good. What trouble are you having? We'll look into this with you and see what's going.)

The above conversation is used to prepare the generic and courteous responses.

- Generic Response:

  आपको क्या परेशानी हो रही है?

  (What trouble are you having?)

- Courteous/Polite Response:

  मुझे क्षमा करें, यह अच्छा नहीं है। हम आपके साथ इस पर गौर करेंगे और देखेंगे कि क्या हो रहा है।

  (I am sorry, that's not good. We'll look into this with you and see what's going.)

For our task, we need to filter out the courteous words, phrases, expressions, and sentences from a given customer care tweet. Hence, we divide the tweets into sentences. The entirely courteous (polite and non-informative) sentences are removed. Completely informative sentences are retained, while the sentences having both the courteous and informative information are transformed (to extract only the courteous information from these sentences). We characterize these three types of sentences as the following:

- Informative sentences: These are the sentences that comprise of the actual content of the tweet without having any courteous phrases or expressions. These sentences mainly consist of instructions, suggestions, assertions, and imperatives.

  Example:

  हमें फ़ोन का संपूर्ण कॉन्फ़िगरेशन भेजें।

  (Send us the entire configuration of the phone.)

[1]The dataset is available in `https://www.iitp.ac.in/~ai-nlp-ml/resources.html#courteousH`

- Courteous sentences: These are the sentences that do not contain any information and are purely polite. The sentences may comprise of greetings and expressions illustrating apology, enthusiasm, appreciation, empathy or assurance to the customers for increasing customer satisfaction.

  Example:

  असुविधा के लिए हमें बेहद खेद है।

  (We are extremely sorry for the inconvenience.)

- Hybrid sentences: These are the sentences that comprise of both the information as well as the polite/courteous expressions in them.

  Example:

  हम यह सुनकर खुश हैं, रसीद के लिए हमें अपनी मेल जानकारी भेजें।
  (We are happy to hear this, for the receipt send us your mail information.)

#### 3.1.3. Data creation process:

To prepare a large scale dataset, we follow the procedures discussed below for every company separately. To speed-up the annotation, we annotate the utterances individually by grouping similar sentences together. The detailed processes are described below.

- Segmentation of Sentences: The tweets from the customer care agents are first obtained. Each tweet is then divided into the three types of sentences, (i). informative if they have purely information, and no courteous expression in it; (ii). purely courteous utterances; and (iii). hybrid sentence denoting both informative and courteous.

- Clustering: The customer care agents of a particular company mainly use expressions and sentences belonging to similar patterns. Hence grouping these similar expressions and sentences before annotation helps in making the annotation process faster. The vector representation of the utterances using the FastText embeddings (Bojanowski et al., 2017) for the Hindi language is used to represent the utterances. We then use the K-means clustering algorithm (Aggarwal and Zhai, 2012) with k = 300 to cluster these sentences. Basically, by clustering, we intend to divide the sentences into groups, where the sentences in a particular group are highly similar to each other in comparison to sentences in other groups.

- Annotation: The segmented and clustered sentences are annotated by three annotators proficient in the Hindi language. The annotators were asked to label each sentence into three categories that are courteous, informative and hybrid. The sentences labeled as hybrid was then used to prepare the generic sentences. Annotators were asked to remove the courteous phrases from the hybrid sentences to obtain only the informative part which is considered as the generic response. The annotators were also requested to remove the non-Hindi sentences, if any, from the dataset and filter them out. The phrases of words of any other language (if any) were also asked to be replaced by its subsequent Hindi words and phrases. For annotation, we observe a multi-rater Kappa agreement of 90%, which can be considered as reliable and substantial.

- Generic response preparation: For a given utterance $U$ of the customer care agent having multiple sentences $u_1, u_2, \ldots, u_m$, we assume it is comprised of both informative and courteous sentences. In order to obtain only the generic part, we exclude the courteous sentences from the given utterance $U$, thereby retaining only the informative sentences. Also, for the hybrid sentences, we remove the courteous phrases/expressions while keeping the informative part of the utterance.

### 3.2. English conversational dataset:

English dataset is based on our prior work (Golchha et al., 2019). This dataset consists of real interactions between the users and customer care agents of different companies on their Twitter handles taken from Kaggle. The dataset comprises of conversations between agents and customers with every utterance in the conversation labeled as courteous, informative or hybrid. The utterances in the conversation having only polite expressions are tagged as courteous, while the utterances having pure information are marked as informative. Finally, the utterances having both courteous expressions and information are labeled as a hybrid. The authors labeled the dataset with these three labels for preparing the generic responses for the proposed system.
The conversations are divided into train, validation and test sets as shown in Table 2. Each training instance is of the form: conversational history (last three utterances), generic response and courteous response.

| ***Language*** | ***Type*** | ***Train*** | ***Valid*** | ***Test*** |
|---|---|---|---|---|
| *English* | *Conversation* | 140203 | 20032 | 40065 |
| | *Utterances* | 179034 | 25642 | 51238 |
| *Hindi* | *Conversation* | 43207 | 8415 | 13472 |
| | *Utterances* | 68745 | 11428 | 17315 |

Table 2: Dataset Statistics

## 4. Methodology

The focus of our current task is to generate polite responses for different languages, given the conversational history (that is previous utterances of the conversation) and the generic response. The architecture of our proposed model showcasing the joint training of both English and Hindi utterances for generating courteous responses for both the languages is depicted in Figure 1. This has been developed from the English model of (Golchha et al., 2019). For the individual model we use the same approach as in (Golchha et al., 2019).

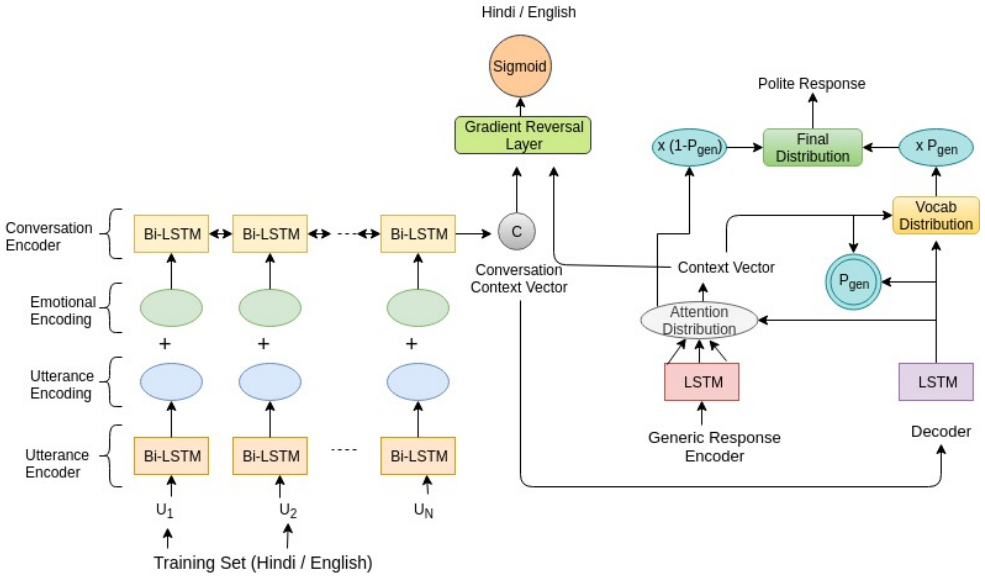


Figure 1: Architectural diagram of the proposed model for joint training of Hindi and English. Inputs to the model are the conversational history and generic response; Gradient reversal layer is used to learn language invariant features; Polite responses for both Hindi and English are the generated outputs of the proposed model.

**Embedding Layer:** Word embeddings are usually trained through an unsupervised manner on a huge dataset, and then the embeddings are fine-tuned by the supervised training process. For word embedding, we use the pre-trained embedding model, FastText[2] for both English and Hindi. The monolingual embeddings for Hindi and English are mapped in the same vector space using linear transformation as illustrated in (Artetxe et al., 2018). With this technique, embeddings for every language exist in the same vector space and maintain the property that words with similar meanings (regardless of language) are close together in the vector space. Hence, the words in English appears close to the words in Hindi in the embedding space. Thus, we train on one or more languages and learn a model that operates on words of a particular language that was not present during training.

**Contextual Encoder:** The context encoder captures the conversational history $C$, which is a sequence of user utterances $u_1, u_2, ....., u_n$, where $n$ is the total number of utterances in a given conversation. Each user utterance $u_n$ comprises of a sequence of words $w_1, w_2, ..., w_{n'}$ where $n'$ is the total number of words in a given utterance, and every word is represented by their pre-trained embeddings. We use the DeepMoji (Felbo et al., 2017) output distribution that is pre-trained on the emoji prediction task to encode the utterances with their corresponding emotional states. A Bi-directional Long Short Term Memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) layer is used for encoding the utterances, and their representations are denoted by $h_i^1, h_2^1, ....h_n^1$, where $n$ denotes the $n_{th}$ word in the utterance. The last hidden state $h_n^1$ of the Bi-LSTM denoting the utterance representation is concatenated with the emotional representation of the given utterance to give the final utterance representation $f_d$. The final utterance representation $f_1, f_2, ..., f_{\tilde{n}}$ is encoded using a second hierarchical Bi-LSTM layer as hidden states $h_1^2, h_2^2, ...., h_{\tilde{n}}^2$. The conversational history is represented by the last hidden state $h_{\tilde{n}}^2$, and is thereby referred to as the conversational context vector $c$.

**Generic Response Encoder:** The word embedding sequence of the generic response for either Hindi or English language is encoded using the unidirectional LSTM network, and the obtained utterance representation is denoted by $h_i$.

**Decoder:** To calculate attention distribution over the encoder state, the decoder LSTM state $s_t$ is used at every decoder time step $t$.

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{attn}) \quad (1)$$

$$\alpha^t = softmax(e^t) \quad (2)$$

where $v$, $W_h$, $W_s$ and $b_{attn}$ are trainable parameters. This attention distribution guides decoder to focus on relevant encoder states at every time step. Context vector is calculated using weighted sum of the encoder states.

$$h_t^* = \sum\nolimits_i \alpha_i^t h_i \quad (3)$$

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

$$s_t = LSTM(s_{t-1}, W_p[w_{emb}(y_{t-1}), h_{t-1}^*, c] + \tilde{b}) \quad (4)$$

[2]https://fasttext.cc/

**Pointer Generator Network:** We use the mechanism analogous to (See et al., 2017) to help copy expressions from the generic response while producing a courteous reply for both the languages. In pointer generator network, the model calculates two distributions, one over the vocabulary ($p_{vocab}$) and the other over the encoded words ($\alpha^t$).

$$p_{vocab} = softmax(\hat{W}(\tilde{W}[s_t, h_t^*] + \tilde{b}) + \hat{b}) \quad (5)$$

where $\tilde{W}$, $\hat{W}$, $\tilde{b}$ and $\hat{b}$ are the trainable parameters.
The *generation probability* $p_{gen}$ dynamically measures the trade-off between the two distributions using the information from the decoder state $s_t$, context vector $h_t^*$, the decoder input $x_t$, and conversational context vector $c$:

$$p_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + w_c^T c + b_{gen}) \quad (6)$$

where vectors $w_{h^*}$, $w_s$, $w_x$, $w_c$ and scalar $b_{gen}$ are trainable parameters and $\sigma$ is the Sigmoid function.
The final distribution over the union of the words of the generic response and the vocabulary words is calculated by:

$$P(w) = p_{gen} p_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} \alpha_i^t \quad (7)$$

**Training and Inference:** We jointly use teacher forcing, reinforcement learning and adversarial learning paradigm to train our model. If $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{n'}\}$ is the gold output tokens for given generic response tokens $c_1$ and conversation history $c_2$, the maximum-likelihood objective using teacher forcing is given by:

$$L_{MLE} = -\sum_{t=1}^{n'} \log p(\tilde{y}_t | \tilde{y}_1, \ldots, \tilde{y}_{t-1}, h_t^*, c) \quad (8)$$

The teacher forcing algorithm discussed in the previous section suffers from exposure bias problem (Ranzato et al., 2016) due to the mismatch in training and inference procedures. One way to bridge the gap between training and inference is by augmenting the standard supervised learning with a reinforcement learning framework (Norouzi et al., 2016).
In other words, designing a task-specific reward function helps distribute the probability mass of the model among the valid sequences that do not occur in the training set. We are particularly interested in self-critical reinforcement learning (RL) (Rennie et al., 2017) algorithm for our polite generation task. This approach tackles the high variance problem in REINFORCE (Williams, 1992) estimator by choosing the greedy decoding score as a baseline. This results in an inference time algorithm without the need for training an additional "critic" network for quick baseline (value function) estimation.
During training, two output sequences are produced: $y^s$, obtained by sampling from the probability distribution $P(y_t^s | y_1^s, \ldots, y_{t-1}^s, \mathcal{D})$, and $y^g$, the greedy-decoding output sequence. We define $r(y, y^*)$ as the reward obtained for an output sequence $y$, when the ground truth sequence is $y^*$. The self-critical RL loss can be written as:

$$\mathcal{L}_{RL} = (r(y^s, y^*) - r(y^g, y^*)) \sum_{t=1}^{n'} \log P(y_t^s | y_1^s, \ldots, y_{t-1}^s, h_t^*, c) \quad (9)$$

Our reward function formulation for the task is the same as (Golchha et al., 2019) i.e., BLEU score and emotion accuracy(the cosine similarity of the emoji distributions of the gold and generated responses).
Along with training using teacher forcing and reinforcement learning, we also use adversarial learning to learn language invariant representation. We use the gradient reversal layers to efficiently use stochastic gradient descent based training for optimizing adversarial language network. It allows the input vectors during forward propagation, and sign inversion of the gradients during backpropagation, to be utilized (Ganin et al., 2016; Masumura et al., 2018a). The probability of the language label can be computed as:

$$\hat{y}_l = \sigma(GRL(c, h_t^*)) \quad (10)$$

$$L_{AL} = -\log p(\hat{y}_l | y_l, c, h_t^*) \quad (11)$$

where $y_l$ is the ground-truth language label (in our case it will be Hindi or English).
The maximum likelihood (ML) objective function is used to pre-train the model (Eq. 8) and then using a mixed objective function with a reduced learning rate:

$$L_{mixed} = \alpha L_{RL} + \beta L_{MLE} + (1 - \alpha - \beta) L_{AL} \quad (12)$$

where, $\alpha$, $\beta$ are hyperparameters. Baselines: We develop the following models:
1. **Seq2Seq**: This is a Seq2Seq model with attention (Luong et al., 2015) and decoder conditioned on the conversational context vector $c$ (without concatenating emotional embedding).
2. **Seq2SEq+P**: This model is developed using the previous Seq2Seq model along with the copying mechanism of Pointer Generator Network.
3. **Seq2Seq+P+EE**: This model is developed using Seq2Seq + P along with emotional embeddings in the conversational context vector.
4. **Seq2Seq+P+EE+RL**: This model is developed using the previous model by adding mixed training of both machine learning and reinforcement learning.

## 5. Experiments

**Implementation Details:** All the implementations are done using the PyTorch[3] framework. We use the same vocabulary for both the encoder and decoder. The vocabulary is collected from the training data, and we keep the top 50,000 frequent words. We use 128 dimensional fasttext(Smith et al., 2017) word embeddings. We use the dropout(Srivastava et al., 2014) with probability 0.45. During decoding, we use a beam

[3]https://pytorch.org/

| Model | Type | English | | | | | Hindi | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ***BLEU*** | ***ROUGE-L*** | ***PPL*** | ***EA*** | ***CP*** | ***BLEU*** | ***ROUGE-L*** | ***PPL*** | ***EA*** | ***CP*** |
| *Seq2Seq* | *Without Joint Training* | 56.80 | 64.52 | 58.21 | 82.43 | 68.34 | 48.67 | 53.61 | 62.47 | 70.33 | 59.41 |
| *Seq2Seq + P* | | 66.11 | 66.40 | 42.91 | 81.98 | 77.67 | 54.33 | 56.22 | 55.11 | 69.86 | 64.75 |
| *Seq2Seq + P + EE* | | 68.16 | 71.17 | 43.52 | 85.75 | 76.05 | 55.75 | 57.58 | 54.36 | 74.54 | 66.72 |
| *Seq2Seq + P + EE + RL* | | 69.22 | 72.37 | 43.77 | 86.87 | 77.56 | 56.82 | 58.88 | 53.81 | 75.23 | 67.16 |
| *Seq2Seq* | *With Joint Training* | 57.18 | 65.75 | 57.48 | 82.56 | 69.43 | 51.16 | 54.26 | 60.54 | 72.45 | 61.52 |
| *Seq2Seq + P* | | 68.38 | 69.25 | 41.66 | 82.21 | 78.52 | 56.03 | 57.82 | 48.67 | 71.32 | 66.71 |
| *Seq2Seq + P + EE* | | 70.84 | 72.77 | 42.98 | 86.34 | 76.85 | 57.14 | 59.48 | 46.23 | 76.81 | 67.71 |
| *Seq2Seq + P + EE + RL* | | 71.22 | 73.37 | 43.11 | 87.41 | 78.33 | 57.82 | 59.93 | 45.79 | 77.15 | 68.09 |
| *Our Model* | | **72.45** | **75.21** | **41.89** | **87.96** | **79.20** | **59.66** | **61.48** | **44.18** | **77.93** | **68.52** |

Table 3: Results of various models. Here, P: Pointer generator model; EE: Emotional embedding; RL:Reinforcement learning; Our model: Seq2Seq + P + EE + RL + Adversarial Training; PPL: Perplexity; CP: Content Preservation; EA: Emotion Accuracy

search with beam size 10. We initialize the model parameters randomly using a Gaussian distribution with Xavier scheme (Glorot and Bengio, 2010). The hidden size for all the layers is 512. We employ AMSGrad (Reddi et al., 2018) as the optimizer for model training to mitigate the slow convergence issues. We use uniform label smoothing with $\epsilon = 0.1$ and perform gradient clipping when gradient norm is over 5. We monitor smoothened running loss on the validation set for early stopping and finding the best models for decoding.

| Language | Model | F | | | CA | | | PC | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *0* | *1* | *2* | *0* | *1* | *2* | *-1* | *0* | *1* |
| *English* | *Seq2Seq* | 16.88 | 41.32 | 41.80 | 16.74 | 40.33 | 42.93 | 24.56 | 48.71 | 26.73 |
| | *Our Model* | 9.87 | 42.05 | 48.08 | 13.52 | 39.27 | 47.21 | 13.24 | 37.19 | 49.57 |
| *Hindi* | *Seq2Seq* | 15.42 | 40.54 | 44.04 | 17.23 | 41.63 | 41.14 | 25.84 | 50.66 | 23.50 |
| | *Our Model* | 10.56 | 41.28 | 48.16 | 14.11 | 38.77 | 47.12 | 14.62 | 38.39 | 46.99 |

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

**Automatic evaluation:** In addition to conventional metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and perplexity, we also use two task-specific metrics such as Emotional Accuracy (EA) and Content preservation (CP) for automatic evaluation as described in (Golchha et al., 2019).

**Human evaluation:** We adopt a human assessment to compare the efficiency of various models to comprehend the quality of the generated polite responses same as (Golchha et al., 2019). For human evaluation, we randomly chose 700 samples from the test set for both the languages. Six human annotators with postgraduate exposure on both Hindi and English languages (three each for a language) were allocated to assess the polite responses produced by the different models for the specified three metrics, given the generic response along with conversation history for a particular language. The metrics for human evaluation are (i). Fluency (F): checks whether the generated responses are grammatically correct; (ii). Content Adequacy (CA): no loss of information in the generated response in comparison to the generic responses; (iii). Politeness Consistency (PC): politeness added to the generic responses is consistent with the history of the conversation.

The fluency and content adequacy scoring system is 0: inaccurate or incomplete, 1: mildly right, 2: correct, whereas the scoring system for politeness consistency is -1: improper, 0: inadequate/not-polite, 1: suitable, respectively. For the above metrics, we calculated the Fleiss kappa (Fleiss, 1971) to evaluate consistency between raters. The fluency, content adequacy and politeness consistency kappa score is 0.77 indicating "considerable agreement".

## 6. Results and Analysis

**Automatic evaluation results:** In Table 3, we present the results of the different models. The experimental results without joint training showcase the results of the various models when Hindi and English languages were trained independently (i.e., training on a particular language and testing on the same language)[4]. While the experiment results with joint training showcased the results of various models when both Hindi and English were trained simultaneously. From the results, it is evident that the joint training of both the languages has helped in improving the performance of both Hindi and English in comparison to the individual models (i.e., only Hindi or English). For English, there is a significant improvement of 1.9% in BLEU score while the model is jointly trained on both the languages. While there is an increased improvement of 1.33% (overall 3.23%) using our proposed model which incorporates adversarial training. The proposed model performs significantly better than the other models for all the evaluation metrics, and the performance of each model is statistically significant compared to the baselines[5].

Similarly, the emotional accuracy of our proposed model by jointly training both Hindi and English is 2.7% and 1.09% better than the individual models, respectively. In Table 5, we present a few examples of the polite responses generated by our proposed model for both English and Hindi.

**Human evaluation results:** In Table 4, we present the human evaluation results of joint training for both the languages. In the case of fluency, our proposed model for both the languages performs better than the baseline sequence-to-sequence model. There is an

[4]The results for English dataset without joint training is same as (Golchha et al., 2019)

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

| Conversational History | Generic Response | Polite Response |
|---|---|---|
| Somebody please help me, I'm having the worst luck with your customer service | How can we help? | Help has arrived! We are sorry to see that you are having trouble, how can we help? |
| मेरा फोन एक मिनट पहले पूरी तरह से काम कर रहा था और फिर अचानक काम करना बंद कर दिया। (My phone was perfectly working a minute back and then it stopped working suddenly.) | आपको क्या परेशानी हो रही है? (What trouble are you having?) | हम आपके साथ इस पर गौर करेंगे, कृपया हमें बताएं कि आपको क्या परेशानी हो रही है? (We'll look into this with you, please tell us what trouble are you having?) |
| Is the Radio X radio within the app branded station, or just a station called Radio X? | Can you send us a screenshot of what you're seeing? | No worries at all! Help's here! Please send us a screenshot of what you're seeing? We'd like to take a close look. |
| पिछले हफ्ते ही घर आया था और मुझे फिर से अपने इंटरनेट से समस्या हो रही है। (Just came home last week and I'm having problems with my internet again.) | आपके इंटरनेट के साथ क्या हो रहा है? (What is happening with your internet?) | अरे नहीं यह अच्छा नहीं है। मैं मदद कर सकता हूँ! कृपया बताएं कि आपके इंटरनेट में क्या समस्या है? (Oh no that's not good. I can help! Please tell whats the problem with your internet?) |

Table 5: Examples of polite responses generated by the proposed model for both the languages given the conversational history and generic response.

improvement of more than 5% in comparison to the baseline model. Also, in the case of content adequacy, the proposed model shows an increase of 5% and 6% for English and Hindi, respectively, in contrast to the baseline models. Similarly, for the politeness consistency, the polite responses generated by the proposed model for Hindi has a score of 46.99% indicating that the model has been successful in generating polite responses in accordance to the conversational history. Also, for English, the score is 49.57% for politeness consistency. Hence, it validates the fact that our proposed model is capable of generating polite and courteous responses than the baselines. From this assessment, we can infer that the responses generated by our model are not only sufficient in terms of preserving information but can also induce polite behaviour by making the responses diverse and interesting.

**Error Analysis:** To conduct a thorough qualitative analysis of the generated responses, we further analyze the outputs produced by our proposed model. Some common types of errors include:

**1.** Incomplete responses: There are instances when the generated responses are left incomplete as unknown token gets predicted due to out-of-vocabulary (OOV) words. This phenomenon is common for Hindi responses as the number of unknown tokens are more in comparison to English. For example, Gold: *.. which store in minneapolis did you visit ?*; Predicted: *.. which store in unk unk ....*

**2.** Extra information: There are instances when the predicted responses contain more information than the generic responses. This happens mostly for English responses as the number of training examples is more in comparison to Hindi. For example, Gold: *Send us a dm.*; Predicted: *Please send us a dm. Did you want to know about other information ?*

**3.** Mixture of languages: The generated responses, sometimes in the case of Hindi, contains English words as well. This is mainly because the number of training examples is more for English. Gold: अपने डिवाइस पर सेटिंग्स समायोजित करें।; (Adjust the settings on your device.) Predicted: कृपया अपने device पर सेटिंग्स adjustकरें।

**4.** Wrong polite expression: The proposed model, for both the languages, sometimes generates wrong polite expressions with respect to the conversational context. Gold: *we are here, reply by dm about the bag details and ..*; Predicted: *thanks for the feedback, enjoy your day..*

## 7. Conclusion and Future Work

In this paper, we have proposed a novel research direction of incorporating politeness across languages. For Hindi and English languages, we transform the generic responses to courteous responses, thereby providing user satisfaction and helping the companies/organisations in building strong customer relations leading to customer retention. With different languages, the politeness varies as courteous expressions changes, thereby making the task even more challenging. For this work, we have prepared a large Hindi customer care conversational dataset by mining real user interactions from Twitter. Our proposed models can handle both the languages and can simultaneously generate polite responses for a given language by modelling the conversational history and being emotionally aware of the state of the user through the emotional embeddings. Experimental results show that the proposed models have been capable of inducing politeness in the generic responses for both Hindi and English by being contextually correct and in accordance to the emotional state of the user without any loss of information that is present in the generic responses.

In future, we would like to extend this work for more languages and investigate the change in politeness across multiple languages. Also, we would look into a code-mixed scenario, when two languages are mixed and see the usage of politeness in this situation.

## Acknowledgement

Authors duly acknowledge the support from the Project titled "Sevak-An Intelligent Indian Language Chatbot", Sponsored by SERB, Govt. of India (IMP/2018/002072). Asif Ekbal gratefully acknowledges Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

## 8. Bibliographical References


Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Carlson, K., Riddell, A., and Rockmore, D. (2017). Zero-shot style transfer in text using recurrent neural networks. *arXiv preprint arXiv:1711.04731*.

Do, Q. N. T. and Gaspers, J. (2019). Cross-lingual transfer learning for spoken language understanding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5956–5960. IEEE.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (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), Copenhagen, Denmark, September 9-11, 2017*.

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

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 249–256.

Golchha, H., Firdaus, M., Ekbal, A., and Bhattacharyya, P. (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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 851–860.

Herzig, J., Feigenblat, G., Shmueli-Scheuer, M., Konopnicki, D., Rafaeli, A., Altman, D., and Spivak, D. (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, 13-15 September 2016, Los Angeles, CA, USA*, pages 64–73.

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

Huang, C., Zaiane, O., Trabelsi, A., and Dziri, N. (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), New Orleans, Louisiana, USA, June 1-6, 2018,*, volume 2, pages 49–54.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., and Gao, J. (2016). Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1192–1202.

Li, J., Jia, R., He, H., and Liang, P. (2018a). Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1865–1874.

Li, Y., Zhao, X., Xu, W., and Yan, Y. (2018b). Cross-lingual multi-task neural architecture for spoken language understanding. In *Interspeech*.

Lin, C.-Y. (2004). Rouge: a package for automatic evaluation of summaries. July.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2018). Dialoguernn: An attentive rnn for emotion detection in conversations. *arXiv preprint arXiv:1811.00405*.

Masumura, R., Shinohara, Y., Higashinaka, R., and Aono, Y. (2018a). Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 633–639.

Masumura, R., Tanaka, T., Higashinaka, R., Masataki, H., and Aono, Y. (2018b). Multi-task and multi-lingual joint learning of neural lexical utterance classification based on partially-shared modeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018,*

*Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3586–3596.

Niu, T. and Bansal, M. (2018). Polite dialogue generation without parallel data. *TACL*, 6:373–389.

Norouzi, M., Bengio, S., Chen, z., Jaitly, N., Schuster, M., Wu, Y., and Schuurmans, D. (2016). Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1723–1731.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Raghu, D., Gupta, N., et al. (2018). Hierarchical pointer memory network for task oriented dialogue. *arXiv preprint arXiv:1805.01216*.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Reddi, S. J., Kale, S., and Kumar, S. (2018). 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*.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Serban, I. V., Klinger, T., Tesauro, G., Talamadupula, K., Zhou, B., Bengio, Y., and Courville, A. C. (2017a). Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3288–3294.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A. C., and Bengio, Y. (2017b). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6830–6841.

Shen, X., Su, H., Niu, S., and Demberg, V. (2018). Improving variational encoder-decoders in dialogue generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5456–5463.

Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Upadhyay, S., Faruqui, M., Tür, G., Dilek, H.-T., and Heck, L. (2018). (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 6034–6038. IEEE.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

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

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

Wu, X., Martinez, A., and Klyen, M. (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) New Orleans, Louisiana, USA, June 1-6, 2018,*, volume 1, pages 2049–2059.

Zhang, H., Lan, Y., Guo, J., Xu, J., and Cheng, X. (2018). Reinforcing coherence for sequence to sequence model in dialogue generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4567–4573.

Zhao, T., Zhao, R., and Eskénazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1:*

*Long Papers*, pages 654–664.

Zhou, X. and Wang, W. Y. (2018). Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1128–1137.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. ). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739.