Development of an Empathetic Conversational Agent

Michael Behr* and Ye Fan†
Department of Electrical and Computer Engineering,
University of Waterloo
Waterloo, Ontario, Canada
Email: *mbehr@uwaterloo.ca, †y245fan@uwaterloo.ca,
Student Number: *20868195, †20868356

***Abstract*—There has been a large push in recent years to develop improved emotional conversational agents (chatbots) from both an industry and academia standpoint. An important application that has been rarely explored to this date, is the usage of an empathetic, counselling conversational agent. To this end, in this paper we propose the Empathetic Conversational Agent (ECA) which is designed to generate emotional and empathetic responses to the user. ECA is built upon a sequence to sequence (Seq2Seq) model utilizing LSTM layers, and is trained on specifically designed empathetic dialogue (ED) corpus. Most importantly, ECA utilizes emotion enriched word embeddings that are built on the global vectors for word representation (GloVe) embeddings. The combination of these factors allows ECA to adjust its tone and responses in an empathetic manner. We compared ECA to three other chatbot models including a model trained on the Cornell movie dialogue corpus (Baseline Chatbot), a model trained on the ED corpus without pre-trained embedding (baseline ECA) and a model trained on the ED corpus using GloVe embedding (GloVe ECA). We found that all 3 models trained on the ED corpus have superior performance based on human evaluation in terms of relevance, understandability and emotional awareness. Overall, participants reported that ECA and GloVe ECA have significantly ($p<.05$) higher relevance, and understandability scores. Additionally, ECA was rated as having the highest emotional awareness although it was not statistically significantly higher than GloVe ECA. The perplexity and cosine similarity were calculated as quantitative evaluation metric for the chatbot models, though these metrics alone are not sufficient to fairly represent the model performance. In conclusion, we present a small, well tuned, generative conversational agent that has the capability of producing empathetic responses in emotionally grounded conversations.**

Index Terms – Empathetic Conversational Agent (ECA), sequence-to-sequence model, emotional awareness, model evaluation

I. Introduction

In recent years, there has been a large amount of focus on the development of conversational agents from both an academic and industry standpoint [1], [2],[3]. One important aspect that comes with the increased relevance of conversational agents, is that for a chatbot to have engaging, and natural conversation it must have the capability to detect the user's emotional state and respond accordingly [4],[5],[6],[7]. There have been multiple studies that have demonstrated that by introducing emotion representation/data into a conversational agent, they positively improved the user experience [8], [9]. In general, it seems that by including emotion information into a model, conversational agents are able to engage in natural and improved interactions with the user along with reducing the amount of misrepresentation and miscommunication[9],[10].

A very intriguing application for conversational agents is their usage as a therapeutic conversationalist. Findings from a recent study suggest that youth diagnosed with neurodevelopmental conditions (NDC) or mental health conditions (MHC) may face substantial barriers to attending and completing postsecondary education[11]. Exasperating this trend, is the fact that research shows that there is insufficient funding for mental health infrastructure on Canadian University campus's [12].

Unfortunately, there are still many unknowns in the area of using conversational agents as a solution for mental health issues. While an average user may find an emotionally sensitive chabot easier to talk to, it is unknown as to whether users with MHC will respond in a similar manner. It is possible that these users will be more receptive to chatbots that maintain more neutral tones, as they could find them this conversational agent to be more trusted and intelligent. Another possibility is that with a softer tone, they may resemble an empathetic counsellor. In this case, it's possible it will provide a more trusted conversational companion where they will feel like opening up more in the dialogue. However the case, the overall lack of available MH services make it incredibly difficult to provide prompt, and proper care to affected individuals. This is where we hope that chatbots can provide easy accessibility, scalability, and a low entrance barrier to helping those with MHC.

With this in mind, we introduce the design of the Empathetic Conversational Agent (ECA), which is able to respond in an empathetic manner to the users prompts. ECA provides an empathetic conversation and is able to detect the emotion in the user's questions and prompts in order to provide a wellness stimulating conversation. We approached this in a two-fold manner: first by introducing a well curated and improved empathetic dialogue corpus[8], and second by utilizing emotion enriched word representations [13]. We evaluated ECA by utilizing validation conversation data and compared it's performance against a baseline chatbot trained on the cornell movie dialogue corpus [14]. We also results from a survey comprised of randomly selected validation questions/prompts and the generated responses form our models.

Overall, we hypothesize that with simple but powerful modern chatbot architecture [15], emotion-rich word embedding and a well curated dataset we can develop an effective, and performant conversional agent that can help relieve part of the mental health infrastructure burden.

II. Related Work

To the best of our knowledge, the earliest public attempt to develop a chatbot was Eliza which was an early natural language processing (NLP) computer program created in the 1960's [16]. Eliza simulated conversation by utilizing an in depth pattern matching and substitution technique, that gave the illusion of comprehension and understand on the part of the program. Of course, Eliza had no contextual understanding of conversations and had no capability of conversing with true understanding. Building off of this achievement, another chat bot was created in the 1970's named Parry [17] which while still utilizing rule based approaches implemented an advanced mental model that demonstrated the ability to stimulate emotion in users. This can be considered as the first chatbot to include emotion in its design and implementation. More modern day rule based chatbots include Artifical Linguistic Internet Computer Entity (ALICE) [18], and Microsoft's XiaoIce [19].

However, most modern chatbot technologies utilize NLP techniques along with neural network based approaches instead of rule-based and pattern matching techniques. One of the most popular Deep Learning models is the Sequence-to-Sequence (Seq2Seq) model and has had widespread usage in Neural Machine translation applications [20]. The Seq2Seq model has also been utilized in a multitude of applications like speech recognition, text generation, and text summarization. By default, this model includes an encoder (encodes source sentences) and decoder (emits translation of target sentence) structure utilizing vanilla recurrent neural network (RNN) designs [21]. In general, the encoder will take the input sentence or paragraph and encode it's meaning in a thought vector. The thought vector is just a vector space representation of the input containing a sequence of numbers to represent the meaning of the input. The decoder works by processing this vector and then decoding it into a translation, or output response/sentence. While this model gained initial success in it's application, the vanilla RNN layers struggle with longer input (long sentences or paragraphs). The next big research development came with the creation of long short-term memory (LSTM) neural network layers, as they have a increased capability to handle long sequences [21]. This LSTM-Seq2Seq model implementation has been demonstrated to have great success in the literature with Deep-Probe[22], SuperAgent [23], MILABOT [24], and RubyStar [25] all utilizing the general framework.

Several studies, have built on the LSTM-Seq2Seq model and attempted to improve it by including emotion awareness into the chatbots like the Emotional Chatting machine (ECM) [26]. This was one of the first successful attempts to make a large-scale emotionally-aware conversational agent by using a deep learning neural network approach. Further developments in this direction have attempted to introduce emotion embedding into the word vector embedding process [13],[27],[28], [29] and changing the model architecture to include reinforcement learning [30],[31]. Other studies have also attempted to introduce better training datasets [8], in order to improve model performance when dealing with emotional contexts in conversation.

Overall from the literature, it seems to have a successful, empathetic, and emotion sensitive chatbot utilizing a powerful neural network based approach like the LSTM-Seq2Seq model along with a well tuned dataset is a must. Additionally, to the best of our knowledge, there has not been prior work that has combined these two aspects with emotionally rich word embedding as well. Consequently, even though our neural network model lacks complexity as compared to some of the state of the art industrial conversational agents, our hypothesis is that we can still construct a performant generative agent by combining the above 3 factors.

III. Methods

A. Data Retrieval

1) Dialogue Corpus: As described earlier, the dataset we trained ECA on was collected from Facebook's AI research team, called EMPATHETICDIALOGUES [8]. This is a novel dialogue corpus that includes 25k conversations split into training, validation and test sets. The conversations from this dialogue dataset contain a large range of emotions (32) including the top 5: surprised, excited, angry, proud and sad. This dataset was collected by, recording a one-on-one conversation between two people where the topic of the conversation is grounded in one of these emotions. The users had a back and forth exchange for up to 6 turns in total, with the users giving understanding, empathetic and human responses to each other. Interestingly, the described initial experiments utilizing this data have indicated that models utilizing this dialogue corpus have improved empathy as determined by human user evaluation. This dataset was downloaded from a footnote link in the paper, where the data is stored as comma-separated values (CSV) files. From there, the training, validation and testing sets were easily imported into python for preprocessing using the pandas module.

Additionally, as a comparison for ECA, we decided to train our model architecture on the cornell movie dialogue corpus [14] to create a baseline conversational agent. This model will be referred to as Baseline for the reset of this paper, and we used it to provide an understanding of the performance improvements gained from our methodology. This dataset was downloaded directly from the Cornell website, and is stored as a text file. This was read into python line by line and made ready for preprocessing.

2) Emotion Embedding: The general idea behind word embedding is to allow for words to be represented by an encoding of numbers, enclosed in a vector format. This is an improvement over simple one-hot encoding methods for text mapping, as similar words are not placed close to each other in the computed embedding space. They can be described more specifically as a ”distributed word representation” [13], where each word in the vocabulary list is mapped to a encoded, continues vector. The goal of this procedure is to map words with similar/same contexts to similar/same regions in this new encoded vector space. In this way, the encoded words can predict words that it often appears in the context of in sentences.

Embeddings can be learned by use of a neural network applied to a supervised task. The idea is to obtain the embeddings from the weight parameters of the network which is constantly adjusting them to minimize the loss of the supervised task. The resulting embedding matrix (weight matrix) is the representation of words that are similar to words of a similar context. Two of the most popular pre-trained embedding techniques are GloVe [32] and CBOW [33], where they have shown success in many NLP applications. While both techniques do a satisfying job of mapping words that occur in semantically similar contexts, they do a poor job of mapping emotionally dissimilar words with similar contexts [13]. More specifically, they will map opposite emotions (love, hate) as highly similar (Cosine Similarity) due to the fact that they will often appear in the the same context. For emotion to be included into the embedding, we would want this to be represented by a proper Cosine similarity between emotion representing words. For example, happy and sad are mapped with a Cosine similarity of 0.643 and 0.535 for GloVe and CBOW respectively [13].

It's for this reason that we will utilize the Emotion Word Embeddings (EWE) for ECA, which is a 300 dimension word representation designed to project emotionally similar words into neighboring spaces [13]. Additionally, we compare ECA's performance with and without EWE in order to demonstrate the impact of EWE on the chatbot's performance.

B. Preprocessing

1) Data Extraction: The data preprocessing is done in a series of steps, in which the first is to extract the data from the Pandas data structure. To extract the data, for each conversation we mapped each prompt and corresponding response to two separate data structures: questions and answers. Each prompt was placed in the question structure, whereas the response to that prompt was placed into the answer structure. In this sense, for a 6 line conversation between 2 people we generated 5 pairs of questions and answers. From this, we created 60746 pairs of questions and answers.

2) Data Cleaning: Next, we needed to actually clean our question and answer data. We utilized Python's regular expressions to do the following: remove common contractions (I'm, he's, she's, .. etc.) and replace them with the respective words and remove unnecessary characters and symbols such as (,),#,@ etc. For example, the sentence: "Why she's my friend!!" will become: "why she is my friend". Additionally, each utterance in the data is given a start and end tag for usage in the decoder input. These tags will mark the beginning and end of each utterance, so the model is trained on the correct data. Lastly, any missing entries in the data were deleted.

3) Tokenization: The vocabulary contained in the empathetic dialogue training data is made of 19529 unique words. As explained earlier, one of the ways to create a word embedding representation is by using the training data itself. The first step to do this, is by calling keras's utility function 'Tokenizer' and fitting it on the data. This then provides a standard bag-of-words text encoding scheme based upon the actual training data itself, where each word is encoded with a specific integer in the vocabulary dictionary.

4) Trimming and Padding: Unfortunately, most of the input sentences are of variable length, making it impossible to use as input. For this reason, we trim and pad (with zeros) all of our sentences to a length of 100 words. In this way, we can handle all sentences of varying length.

5) Embedding: As mentioned earlier, the embedding matrix is the matrix of weights that connects the embedding layer to the rest of the model architecture. The input to this embedding layer, is the input data that is tokenized from the previous step. For this step, we import GloVe and EWE embedding matrices into python and use these to create our embedding layer. The EWE embedding is a 300 dimension representation for each word, much like GloVe's. For ECA we do not train the embedding layer weights, as they have already been fine-tuned. These two models we will refer to as GloVe ECA (for GloVe embedding) and ECA (for EWE embedding).

Additionally, we explored learning our own word embedding for our problem by incorporating it as part of the overall model training process to serve as a comparison. We do this by creating the embedding layer out of an embedding matrix created from the empathetic dialogues training data vocabulary. We will call this model Baseline ECA.

C. Model Architecture

Recurrent neural networks (RNN) are designed to extract patterns of time series data, specifically the sentences for a language model. RNN units not only generate output for prediction but also a state at each time step. The state keeps track of the influence from previous time steps and feeds to the RNN unit at the next time step. Such a mechanism is the key for an RNN model for learning time series data. A widely used RNN model, Long Short Term Memory (LSTM) was employed for encoding the input questions and decoding the output answers. LSTM takes advantage of four gate functions and generates two states (Shown in figure 1) to control the flow of information to the following units. Its capability in remembering the inputs from long and short time steps makes it a popular model in natural language tasks.

For the project, a sequence to sequence (seq2seq) architecture was adopted with two LSTM models as encoder and decoder respectively shown in figure 2. As questions and sentences differ in length, such architecture is able to establish the mapping between two sets of data in different lengths. The aim of the encoder is to extract the information in the question sentence and its output state is fed into the decoder. The decoder then takes the encoded state and the start label as initial input. The LSTM unit in the decoder uses the predicted word and state from the previous time step and outputs the next word until it predicts the end label or reaches the maximum length.

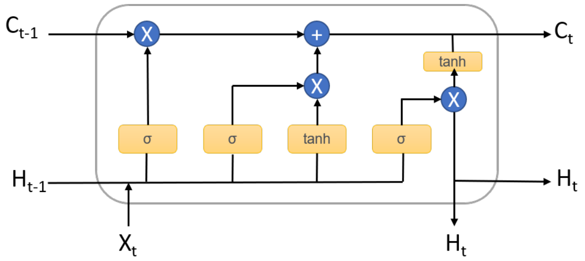

Fig. 1. Long Short Term Memory Unitl

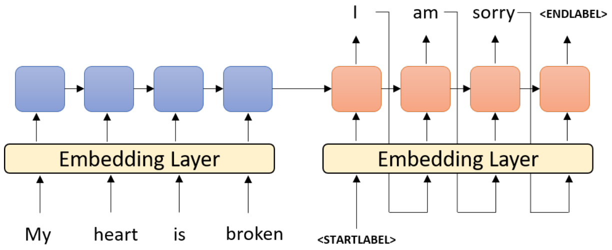

Fig. 2. Sequence-to-Sequence Model Architecture

The ECA model involves several important parameters, including the length of questions and answers, the number of units in LSTM, the vocabulary size and the embedding weights. The maximum length of 114 and 119 words in questions and answers respectively are determined as the length for input and output. Though the vocabulary contains 19529 words, the most frequently 12000 words were adopted to reduce the dimension of the one hot vector of each word. Various number of units in LSTM and word embeddings were experimented to improve the model performance, which will be discussed in the next section.

D. Training

We trained our models utilizing a computer using the freely available Google Collaboratory computer clusters. The specifications included a dual core Intel(R) Xeon(R) CPU running at 2.3 GHz, using 12.7 GB of RAM. Additionally, GPU acceleration was utilized witha Tesla K80 GPU with 12 GB of GDDRB VRAM. The two models using pre-trained word embedding were trained on average for 8 hours (4 hour training periods) over 70 epochs in total. The other two models which developed the word embedding from scratch, were ran for 10 hours, for over 5 epochs in order to properly train the model. Table 1 tabulates and explains these four chatbot models, and the data and embedding each is trained on.

We utilized categorical Cross-entropy as our loss function, as we are predicting in a multi-class classification problem (each word is it's own exclusive class). For cross-entropy loss, each predicted probability is compared to the actual class output and a score is computed that will penalize the probability based on how the distance to the expected value. It is this distance we are attempting to minimize, with lower cross-entropy loss values representing a better prediction than larger ones. Lastly, since we have a sentence which is a collection of predictions, the cross-entropy is averaged across the entire sentence after the per-word basis is calculated. The equation for cross-entropy can be demonstrated as:

$$Categorical\ Cross\ Entropy = -\sum_{i=1}^{classes} y_i \cdot log\ \hat{y}_i \tag{1}$$

The whole dataset contains over 60000 pairs of 114-word questions and 119-word answers. After one hot encoding, each word was transformed into a 12001-dimension (12000 words and 1 for the padded 0) vector. Loading the whole dataset into the RAM led to out-of-memory error. To resolve this issue, a custom data generator was built to feed the batches of data to the training process. The data generator encapsulates the prepossessing steps of one-hot encoding and generates the specified batch size of paired input data for training the ECA model in real time.

Due to limited computing resource, we trained and optimized the ECA model with 30000 sample pairs of questions and answers. As far as the epoch is concerned, the training was performed initially using 120 epochs with default parameters. The train and validation loss dropped to around 0.2 and 0.3 respectively at epoch 70, and the decrease in loss from epoch 70 to 120 was much less significant. Therefore, epoch 70 was adopted for as the default value of training length. It's worth mentioning that as a common practice to prevent oscillations in language learning tasks, RMSprop optimizer was employed for training the ECA model. Lastly, parameters such as LSTM units, batch size, and learning rate were tuned through the use of a grid-search procedure that can be shown in Table 2.

Lastly, to show the effect of the number of epochs we explored ECA's response quality compared to the number of epochs using several examples. This was done by computing the perplexity score for each generated response by using the expected response and is shown in Table 3.

E. Model Evaluation

Model evaluation was conducted by both quantitative methods and human evaluation for the four models, Chatbot Baseline, Baseline ECA, GloVe ECA and ECA.

1) Quantitative Methods: It's worth mentioning that there's lack of standardized procedure for chatbot evaluation [34]. Several metrics have been proposed, including lexical diversity, average cosine-similarity, BLEU-2 score and perplexity, but none of them alone is recognized as the standard metric. In this project, we implemented perplexity and word-averaged cosine-similarity based on GloVe embeddings.

Perplexity was calculated using equation (2) [35]. The loss was obtained by taking the average of each word's categorical cross entropy loss between the prediction and ground truth.

$$perplexity = 2^{loss} \tag{2}$$

Cosine similarity was calculated by first encoding each word into GloVe embedding vectors and then taking the average

TABLE I
COMPARING DIFFERENT CHATBOT MODELS

Model Name	Dataset	Pre-trained Embedding	Sample Size	Epochs
Baseline Chatbot	Cornell Movie Dialogue Corpus	None	30k	70
Baseline ECA	EMPATHETIC DIALOGUES	None	30k	70
GloVe ECA	EMPATHETIC DIALOGUES	GloVe 300 dimension	30k	70
ECA	EMPATHETIC DIALOGUES	EWE 300 dimension	30k	70

TABLE II
PARAMETER GRID-SEARCH

Epochs	Sample Size	Batch Size	Learning Rate	LSTM Units
50,70,100,120	10k, 20k, 30k, 60k	32,64,128,256	0.01, 0.001, 0.0001	64, 128, 200, 300

of each word vectors for the prediction and ground truth response, respectively. The cosine similarity between these two sentences can be calculated by the following equation:

$$cosine\ similarity = cos(\theta) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\cdot\|\vec{B}\|} \qquad (3)$$

2) Human Evaluation: To further evaluate the chatbot, we designed a survey that contains 10 randomly selected validation questions and the responses generated from each of the four models. Each of the four chatbot model responses were blindly evaluated with randomly selected users by scoring the relevance, understandability, and emotional awareness of each response out of 10. These responses were collected and compared using a one-way ANOVA across all user's in order to gain a human interpretability of the four chatbot models performance.

IV. RESULTS

A. The Four Chatbot Models

As seen in figure 3, the training and validation losses for all four chatbot models are displayed. As seen by the plots, the pre-trained word embedding models reach a lower training and validation loss than both the baseline bot and baseline ECA. Additionally, the baseline bot validation loss stopped improving around 15 epochs, performing far worse than all three bots trained on the ED corpus.

Additionally, the perplexity and cosine similarity score of each of the 4 models generated responses on 500 validation data samples is shown in figure 4. The figure shows that the perplexity for the Baseline Chatbot is lower than 0.2 while the other three range from 2.3 to 2.6. From this figure, we can see the difference in cosine similarity between these four models was insignificant, while the GloVe ECA and ECA had slightly higher values. It is noticeable that the variance for perplexity was relatively large while the cosine similarity generally varied from 0.7 to 0.9. The difference in model performance in terms of quantitative metric will be discussed in the following section.

B. Fine-tuning ECA

As explained in above section, the final ECA model and hyperparameters were finely tuned by use of a grid search procedure, with the combinations being shown in Table 2. The optimal model was found to have the following parameters: Learning Rate = 0.001, LSTM units = 300, batch size = 128. The resulting training loss, and validation loss from this procedure was 0.200 and 0.268.

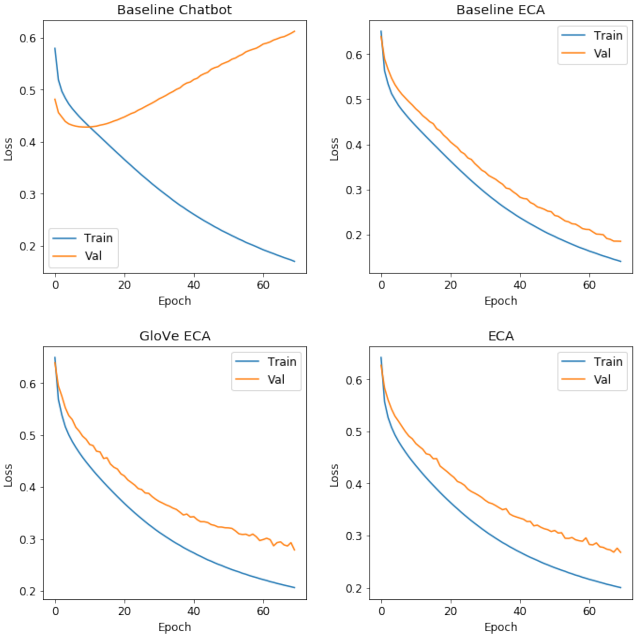

Fig. 3. Training and Validation Loss for each model

C. Generated Responses vs. Epochs

In table 3 we demonstrate examples of the fine-tuned ECA's generated responses vs the number of epochs used for training. The perplexity score is computed for each response, and for both examples it's shown to decrease as the number of epochs increases up to a point. In general, we found that 70 epochs was a sweet spot in terms of subjective quality of responses along with quantitative measurements like perplexity and cross-entropy loss.

D. Model Evaluation through Questionnaire

In table 4, we demonstrate the questions and generated responses from each model from our questionnaire. As explained earlier, we took 10 randomly selected questions from

TABLE III
EXAMPLES OF FINE-TUNED ECA GENERATED RESPONSES VS. TRAINING EPOCHS

Model	**Responses** for ”my girlfriend seems to be distant and talking to other guys”	**Perplexity**	**Cos Similarity**
50 Epochs	1. really that is a lot of people who can do that to be rude	1.696	0.908
70 Epochs	2. really do you have a new one or what do you want to do	1.832	0.908
100 Epochs	3. that is really sweet do you love	3.207	0.853
120 Epochs	4. everything all have good lives and do not all easy enough to know that you should not be able to know that to happen there has changed a lot	1.37	0.95
Expected Response	have you talked to her about it	1	1
Model	**Responses** for ”no but we did end up going out on a date”	**Perplexity**	**Cos Similarity**
50 Epochs	1. thats good im glad you had a good grade to get	1.692	0.859
70 Epochs	2. wow it is always good to hear that is the good job times	1.829	0.939
100 Epochs	3. how much did you do	5.127	0.92
120 Epochs	4. that is really annoying maybe you should just to treat them not to be alone at work a bowl it up a couple days to come back to the good memories	1.352	0.931
Expected Response	that is great did you have fun	1	1

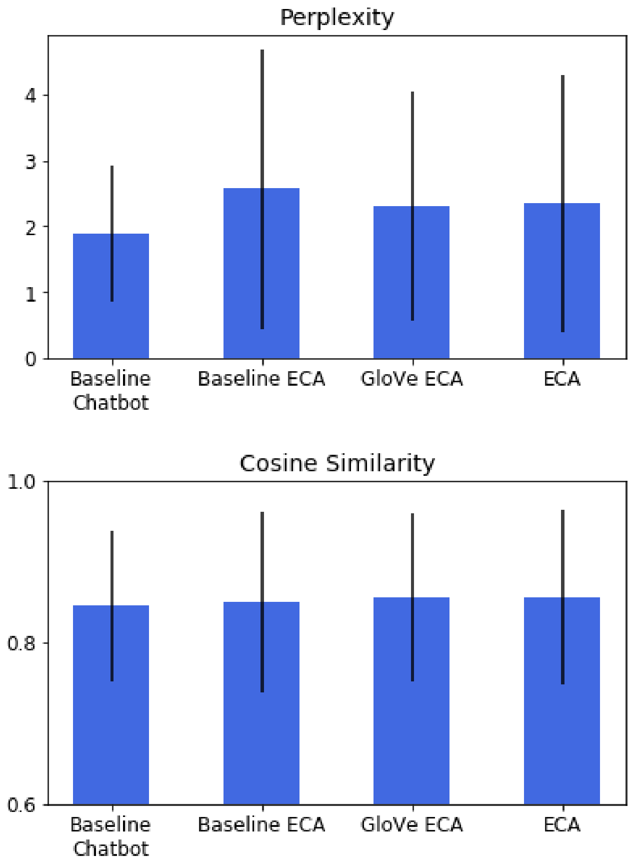

Fig. 4. Perplexity and Cosine Similarity Comparison Between Models

the validation corpus and generated a response for each model. The models are all randomized so the user was unable to know which response belonged to which bot. The results from this questionnaire were then tabulated and statistically compared in figure 5, where each category is compared across all of the models. As seen by this figure, GloVe ECA and ECA have statistically significant higher response ratings than the other two models for all three response categories ($p<0.05$).

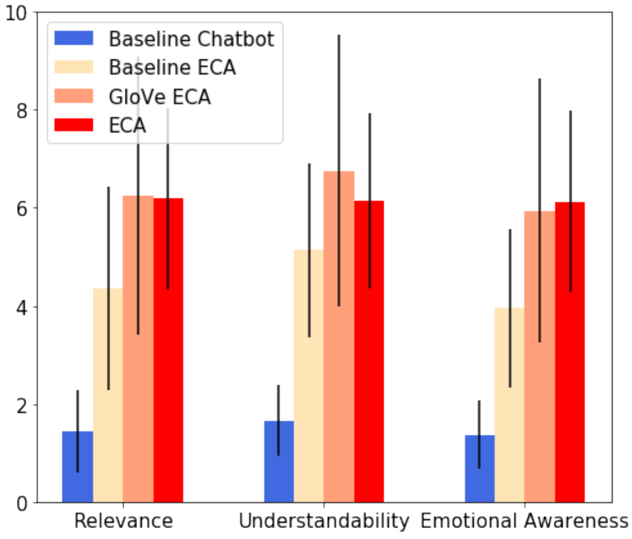

Fig. 5. Human Evaluation

V. DISCUSSION

A. Model Comparison

It is clear from the sample responses, the Baseline Chatbot trained on the general movie dialogue barely gave human readable responses. It's probably due to the lack of emotional context in the Cornell Movie dataset, making it difficult for the baseline model to generate meaningful responses to prompts/questions with emotion. The scores in relevance, understandability and emotional awareness from the evaluation survey were significantly lower than other models, providing concrete evidence in the model's lower performance.

Training on the EMPATHETIC DIALOGUES significantly improved the model performance in generating meaningful responses, to questions/prompts with emotional context. The Baseline ECA model achieved over 4 and 5 on relevance and understandability respectively. Such improvement could be related to the similarity between the training and validation data, in that both training and validation data are grounded in emotional situations. Once the model extracted the latent

TABLE IV

SAMPLE PROMPTS AND GENERATED RESPONSES FOR HUMAN EVALUATION

Model	**Responses** for ”i was extremely relieved when i found out my dad beat cancer”
Baseline	1. recognize we were getting out of him in the light
Baseline ECA	2. oh no i would be so worried
GloVe ECA	3. i am so sorry you had to be so glad you are proud of
ECA	4. oh really good luck for you
Model	**Responses** for ”my girlfriend has not talked to me in a few days and i am struggling to keep it together but i am at the moment”
Baseline	1. recognize we were a good then we could tell the same as they come here
Baseline ECA	2. i can not believe it are you guys taking a break off
GloVe ECA	3. i am sorry to hear do you think you can talk to her about it
ECA	4. oh that is really annoying when you can not be very fortunate of her
Model	**Responses** for ”i have not been able to talk to my girlfriend in a few days and it is making me quite sad”
Baseline	1. recognize about what then it is the most important thing it would like to go to the eyes on that that is the only thing that is what i have been doing this this is the most fun i am looking for a job if you are right i am not even your father scared is this
Baseline ECA	2. that is tough it is always nice to be able to get a good person though
GloVe ECA	3. why
ECA	4. im on that sucks you can get through through it
Model	**Responses** for ”my daughter is very smart she learns things really quickly”
Baseline	1. recognize we were like a and then were then they were not for the same time
Baseline ECA	2. he must be very proud
GloVe ECA	3. that is great you must be so proud of her
ECA	4. that is nice of you to get her family
Model	**Responses** for ”i started a new job about a month ago and it is going really well”
Baseline	1. oh wow what kind of car is it
Baseline ECA	2. oh wow what is your first exam
GloVe ECA	3. wow that is quite exciting you are a great way to the first one is the first time you love and leave on her
ECA	4. wow that is quite exciting congratulations i hope you feel that way because it is tough
Model	**Responses** for ”but that is an awesome grade you should be so proud”
Baseline	1. recognize then and then be a mangs thing can not please matter i am not sure
Baseline ECA	2. i am i am i am just going to expect this work there though
GloVe ECA	3. i am i he really really proud that i loved it too much fun though i am so happy with him
ECA	4. really proud
Model	**Responses** for ”i never expected to fail my final exams that is a major blow for me”
Baseline	1. recognize you were a good could let me see a thing
Baseline ECA	2. yeah are you nervous too much for the one thanks
GloVe ECA	3. what did you do
ECA	4. oh no wow that is a tough situation i hope you have a great time with your new smart to do not get any easier
Model	**Responses** for ”my younger brother always get the best dates and beautiful girls to date i hate it”
Baseline	1. recognize we were about him there was the only a a a a a a a a a a a a a a a
Baseline ECA	2. that is a good thing i am sure that will make you so happy with that
GloVe ECA	3. aw i have not seen that in my life
ECA	4. do you ever be friends
Model	**Responses** for ”i do not like it when people talk with their mouth full of food”
Baseline	1. recognize we were well we would be there for a while we lose the end we could not have a party
Baseline ECA	2. there i am a bit scared to people because i have to find a way to go to my place i am sorry but thats good of you
GloVe ECA	3.i agree with that
ECA	4. oh that would be so annoying
Model	**Responses** for ”i am relaxed right now ate a nice dinner now i am relaxing on the couch watching tv”
Baseline	1. recognize we might be dead for our new friend there s only time before we get back to work on the took it if you could find it off an old man
Baseline ECA	2. i would definitely tell because that is why i am not sure that happened
GloVe ECA	3. i that sounds so relaxing what you did do it
ECA	4. yes it sounds like fun
Model	**Responses** for ”yes surgery is so scary but i am glad it is over”
Baseline	1. recognize we were going to help you were there
Baseline ECA	2. same here
GloVe ECA	3. you should know something to him about the job to tell you
ECA	4. then will be fine

representations within the training data, it successfully applies this to the dialogues with similar context.

The adoption of pre-trained word embedding promoted the model performance even further. In the three criteria of human evaluation, both GloVe ECA and ECA obtained scores around 6. As shown in Table IV, some responses perfectly mimicked the human reactions, which were not only relevant and understandable, but also correctly sensed the emotions in the prompts. It illustrates that the pre-trained word embedding captures the hidden features of words, helping the model converge faster and generate human readable responses.

Additionally, what must be noted is that there was limited differences between the general GloVe embedding and EWE embedding human evaluation scores. Interrestingly, the EWE embedding model (ECA) did score highest on average in the emotional understanding evaluation score. This is what we initially hypothesized, in that we can improve a chatbot model by by adding emotion context into the word embedding. While the improvement is minor in our study, we suspect with more data, longer training a better model architecture this emotional embedding will provide even greater performance gains.

B. Evaluation Methods

The statistics from the perplexity and cosine similarity metrics are not in perfect alignment with the human evaluation. Though the GloVe ECA and ECA achieved slightly higher cosine similarity scores (than the baseline bots), the difference was minor relative to the variance. For perplexity, the Baseline Chatbot obtained the lowest value, indicating it having the most optimal performance, while its responses barely made sense in terms of readability, context (emotion) awareness, and understandability. Compared to the human evaluation results from our survey, we would expect GloVe ECA and ECA to have significantly lower average perplexity scores across responses and significantly higher cosine similarity.

We suspect the main reason for this discrepancy is related to the way both metrics are computed. More specifically, the perplexity follows an exact one-to-one word pairing strategy to compare the prediction with the ground truth. Any changes in word ordering or sentence length would greatly decrease the perplexity score, even though such changes may make little difference in human interpretation. On the other hand, cosine similarity simply compares the distance between the averaged word embedding vectors of two responses. Though it captures the hidden features within each word of the sentences, it is invariant with regard to word ordering and many other aspects in a sentence. This backs up our results, in that we did find higher cosine similarity scores for our best performing models even though they were not significantly higher (as expected).

This finding reflected the general complication in evaluating the chatbot performance with one single, quantitative metric. In this sense, it seems as if the cosine similarity scores are a better metric for evaluating the chat responses compared to the perplexity scores.

C. Further Improvements

While we have demonstrated the effectiveness of combining emotion word embedding with empathetic dialogue corpus, there are still many improvements that can be made.

1) Ensemble Methods: Taking a close look at the current modern, chatbot literature it seems as if most of the current models employ some sort of neural based model, built on an encoder-decoder framework [2]. More specifically, utilizing an encoder-decoder framework to provide seq2seq learning, much like we have implemented in our current model architecture. However there are multiple differences between our smaller, simpler model and some of the state of the art industrial architectures which utilize ensemble methods [25],[30],[31]. In general, an ensemble methodology includes a combination of the three main conversational agent techniques including neural network based, rule based, and retrieval based approaches. This combination approach, seems to provide a more ”human-like” conversational agent, with a good example being the chatbot RubyStar [25]. RubyStar utilizes this ensemble approach to handle a large number of general topics with no specific task-oriented conversation. For example, if the user asks a question about their favourite movie RubyStar utilizes a combination of a template selection model and Named Entity Disambiguation model to parse the input question and determine specific question topics and tags. Using this information, RubyStar is able to search the Wikipedia for information involving the topics and tags and will pull important information about the film into RubyStar's response using a entity-based template. Additionally, it can use these topic keywords to create a fresher response by using Twitter's API to search the top 100 tweets in the last 7 days. From these 100 tweets, a reply is selected creating a human-like response not based on a specific template. Additionally, RubyStar employs neural dialog generation utilizing a seq2seq model much like our model to also generate completely novel candidate responses. Overall, while a neural dialog generation model can produce quality responses it often can struggle to respond with human-like and concrete answers. We feel as if an ensemble approach is the current best approach available to make a performant emotionally aware chatbot.

2) Larger Dialogue Corpus: While ECA is trained on over 30K question and answer pairs, this is still considered to be a small dialogue corpus. Due to this fact, there are many contextual situations ECA is unable to respond correctly to due to their absence in the training data. More specifically, when the user prompts conversation towards a specific topic it is unlikely that ECA can generate a quality response. In general, while ECA generates empathetic responses to emotionally charged prompts or questions it may fail to generate an appropriate response to a more neutral prompt/question.

A solution to increase ECA's general conversational capabilities, would be to expand the model's training to include a larger dialogue corpus. For example, in the paper that introduced the empathetic dialogues dataset [13], they describe using a large corpus of Reddit conversations (1.7 billion) to pre-train their conversational model. They then fine-tuned their models using the empathetic dialogues dataset in order to improve the model's empathetic responses. By first pre-training the model on a large general conversation dataset, they were able to provide an excellent conversational base for their chatbot model. In this way, the generative dialogue produced by the bot improved in its capabilities to handle general conversation. For further improvements of ECA, we would consider utilizing a larger corpus to pre-train before fine-tuning with the empathetic dialogues.

VI. Conclusion

In conclusion, we present a small, well tuned, generative conversational agent (ECA) that has the capability of producing empathetic responses in emotionally grounded conversations. Our results demonstrate that with a smaller dialogue corpus, pre-trained word embedding provides significant improvement in the generative responses. Additionally, the type of pre-trained word embedding had an impact on the response quality in emotional awareness as EWE was demonstrated to be marginally better than using the more general baseline GloVe embeddings. Due to ECA's success, with improved model architecture, larger training datasets and more powerful computing resources we expect to achieve a highly performant, conversational agent which can help perform duties as an empathetic counsellor in the future.

References

[1] Hao Fang et al. "Sounding Board: A User-Centric and Content-Driven Social Chatbot". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 96–100. DOI: 10.18653/v1/N18-5020. URL: https://www.aclweb.org/anthology/N18-5020.

[2] Endang Wahyu Pamungkas. "Emotionally-aware chatbots: A survey". In: *arXiv preprint arXiv:1906.09774* (2019).

[3] Braden Hancock et al. "Learning from dialogue after deployment: Feed yourself, chatbot!" In: *arXiv preprint arXiv:1901.05415* (2019).

[4] Timothy Bickmore and Justine Cassell. "Relational agents: a model and implementation of building user trust". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2001, pp. 396–403.

[5] Jonathan Gratch et al. "Creating rapport with virtual agents". In: *International workshop on intelligent virtual agents*. Springer. 2007, pp. 125–138.

[6] Gale M Lucas et al. "It's only a computer: Virtual humans increase willingness to disclose". In: *Computers in Human Behavior* 37 (2014), pp. 94–100.

[7] Asma Ghandeharioun et al. "EMMA: An Emotion-Aware Wellbeing Chatbot". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 1–7.

[8] Hannah Rashkin et al. "Towards empathetic open-domain conversation models: A new benchmark and dataset". In: *arXiv preprint arXiv:1811.00207* (2018).

[9] Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. "TickTock: A non-goal-oriented multimodal dialog system with engagement awareness". In: *2015 AAAI Spring symposium series*. 2015.

[10] Helmut Prendinger and Mitsuru Ishizuka. "THE EMPATHIC COMPANION: A CHARACTER-BASED INTERFACE THAT ADDRESSES USERS'AFFECTIVE STATES". In: *Applied artificial intelligence* 19.3-4 (2005), pp. 267–285.

[11] Rubab Arim, Marc Frenette, et al. *Are Mental Health and Neurodevelopmental Conditions Barriers to Postsecondary Access?* Tech. rep. Statistics Canada, Analytical Studies Branch, 2019.

[12] Dimitris Giamos et al. "Understanding campus culture and student coping strategies for mental health issues in five Canadian colleges and universities". In: *Canadian Journal of Higher Education/Revue canadienne d'enseignement supérieur* 47.3 (2017), pp. 136–151.

[13] Ameeta Agrawal, Aijun An, and Manos Papagelis. "Learning emotion-enriched word representations". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 950–961.

[14] Cristian Danescu-Niculescu-Mizil and Lillian Lee. "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*. 2011.

[15] Daniel Adiwardana et al. "Towards a human-like open-domain chatbot". In: *arXiv preprint arXiv:2001.09977* (2020).

[16] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[17] Kenneth M Colby. "Human-computer conversation in a cognitive therapy program". In: *Machine conversations*. Springer, 1999, pp. 9–19.

[18] Robert Epstein, Gary Roberts, and Grace Beber. *Parsing the Turing test*. Springer, 2009.

[19] Li Zhou et al. "The design and implementation of xiaoice, an empathetic social chatbot". In: *Computational Linguistics* 46.1 (2020), pp. 53–93.

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

[21] Alex Sherstinsky. "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm)

network". In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306.

[22] Zi Yin, Keng-hao Chang, and Ruofei Zhang. "Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 2131–2139.

[23] Lei Cui et al. "Superagent: A customer service chatbot for e-commerce websites". In: *Proceedings of ACL 2017, System Demonstrations*. 2017, pp. 97–102.

[24] Iulian V Serban et al. "A deep reinforcement learning chatbot". In: *arXiv preprint arXiv:1709.02349* (2017).

[25] Huiting Liu et al. "Rubystar: A non-task-oriented mixture model dialog system". In: *arXiv preprint arXiv:1711.02781* (2017).

[26] Hao Zhou et al. "Emotional chatting machine: Emotional conversation generation with internal and external memory". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[27] Nabiha Asghar et al. "Affective neural response generation". In: *European Conference on Information Retrieval*. Springer. 2018, pp. 154–166.

[28] Pierre Colombo et al. "Affect-driven dialog generation". In: *arXiv preprint arXiv:1904.02793* (2019).

[29] Roman Shantala, Gennadiv Kyselov, and Anna Kyselova. "Neural dialogue system with emotion embeddings". In: *2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC)*. IEEE. 2018, pp. 1–4.

[30] Jia Li et al. "Reinforcement Learning Based Emotional Editing Constraint Conversation Generation". In: *arXiv preprint arXiv:1904.08061* (2019).

[31] Xiao Sun, Xiaoqi Peng, and Shuai Ding. "Emotional human-machine conversation generation based on long short-term memory". In: *Cognitive Computation* 10.3 (2018), pp. 389–397.

[32] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[33] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[34] Joao Sedoc et al. "Chateval: A tool for chatbot evaluation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 2019, pp. 60–65.

[35] Saizheng Zhang et al. "Personalizing dialogue agents: I have a dog, do you have pets too?" In: *arXiv preprint arXiv:1801.07243* (2018).