Emotional Neural Language Generation Grounded in Situational Contexts

Abstract

Emotional language generation is one of the keys to human-like artificial intelligence. Humans use different type of emotions depending on the situation of the conversation. Emotions also play an important role in mediating the engagement level with conversational partners. However, current conversational agents do not effectively account for emotional content in the language generation process. To address this problem, we develop a language modeling approach that generates affective content when the dialogue is situated in a given context. We use the recently released Empathetic-Dialogues corpus to build our models. Through detailed experiments, we find that our approach outperforms the state-of-the-art method on the perplexity metric by about 5 points and achieves a higher BLEU metric score.

Introduction

Rapid advancement in the field of generative modeling through the use of neural networks has helped advance the creation of more intelligent conversational agents. Traditionally these conversational agents are built using seq2seq framework that is widely used in the field of machine translation BIBREF0. However, prior research has shown that engaging with these agents produces dull and generic responses whilst also being inconsistent with the emotional tone of conversation BIBREF0, BIBREF1. These issues also affect engagement with the conversational agent, that leads to short conversations BIBREF2. Apart from producing engaging responses, understanding the situation and producing the right emotional response to a that situation is another desirable trait BIBREF3.

Emotions are intrinsic to humans and help in creation of a more engaging conversation BIBREF4. Recent

work has focused on approaches towards incorporating emotion in conversational agents BIBREF5, BIBREF6, BIBREF7, BIBREF8, however these approaches are focused towards seq2seq task. We approach this problem of emotional generation as a form of transfer learning, using large pretrained language models. These language models, including BERT, GPT-2 and XL-Net, have helped achieve state of the art across several natural language understanding tasks BIBREF9, BIBREF10, BIBREF11. However, their success in language modeling tasks have been inconsistent BIBREF12. In our approach, we use these pretrained language models as the base model and perform transfer learning to fine-tune and condition these models on a given emotion. This helps towards producing more emotionally relevant responses for a given situation. In contrast, the work done by Rashkin et al. BIBREF3 also uses large pretrained models but their approach is from the perspective of seq2seq task.

Our work advances the field of conversational agents by applying the transfer learning approach towards generating emotionally relevant responses that is grounded on emotion and situational context. We find that our fine-tuning based approach outperforms the current state of the art approach on the automated metrics of the BLEU and perplexity. We also show that transfer learning approach helps produce well crafted responses on smaller dialogue corpus.

Approach

Consider the example show in Table TABREF1 that shows a snippet of the conversation between a speaker and a listener that is grounded in a situation representing a type of emotion. Our goal is to produce responses to conversation that are emotionally appropriate to the situation and emotion portrayed.

We approach this problem through a language modeling approach. We use large pre-trained language model as the base model for our response generation. This model is based on the transformer

architecture and makes uses of the multi-headed self-attention mechanism to condition itself of the previously seen tokens to its left and produces a distribution over the target tokens. Our goal is to make the language model $p(y)=p(y_1,y_2,....,y_t;\theta)$ learn on new data and estimate the conditional probability $p(y|x)$. Radford et al. BIBREF10 demonstrated the effectiveness of language models to learn from a zero-shot approach in a multi-task setting. We take inspiration from this approach to condition our model on the task-specific variable $p(y_t|x,y_{< t})$, where x is the task-specific variable, in this case the emotion label. We prepend the conditional variable (emotion, situational context) to the dialogue similar to the approach from Wolf et al BIBREF13. We ensure that that the sequences are separated by special tokens.

Experiments ::: Data

In our experiments we use the Empathetic Dialogues dataset made available by Rashkin et al. BIBREF3. Empathetic dialogues is crowdsourced dataset that contains dialogue grounded in a emotional situation. The dataset comprises of 32 emotion labels including surprised, excited, angry, proud, grateful. The speaker initiates the conversation using the grounded emotional situation and the listener responds in an appropriate manner.Table TABREF4 provides the basic statistics of the corpus.

Experiments ::: Implementation

In all our experiments, we use the GPT-2 pretrained language model. We use the publicly available model containing 117M parameters with 12 layers; each layer has 12 heads. We implemented our models using PyTorch Transformers. The input sentences are tokenized using byte-pair encoding(BPE) BIBREF14 (vocabulary size of 50263). While decoding, we use the nucleus sampling ($p=0.9$) approach instead of beam-search to overcome the drawbacks of beam search BIBREF15, BIBREF16. All our models are trained on a single TitanV GPU and takes around 2 hours to fine-tune the model. The

fine-tuned models along with the configuration files and the code will be made available at: https://github.com/sashank06/CCNLG-emotion.

Experiments ::: Metrics

Evaluating the quality of responses in open domain situations where the goal is not defined is an important area of research. Researchers have used methods such as BLEU , METEOR BIBREF17, ROUGE BIBREF18 from machine translation and text summarization BIBREF19 tasks. BLEU and METEOR are based on word overlap between the proposed and ground truth responses; they do not adequately account for the diversity of responses that are possible for a given input utterance and show little to no correlation with human judgments BIBREF19. We report on the BLEU BIBREF20 and Perplexity (PPL) metric to provide a comparison with the current state-of-the-art methods. We also report our performance using other metrics such as length of responses produced by the model. Following, Mei et al BIBREF21, we also report the diversity metric that helps us measure the ability of the model to promote diversity in responses BIBREF22. Diversity is calculated as the as the number of distinct unigrams in the generation scaled by the total number of generated tokens BIBREF21, BIBREF1. We report on two additional automated metrics of readability and coherence. Readability quantifies the linguistic quality of text and the difficulty of the reader in understanding the text BIBREF23. We measure readability through the Flesch Reading Ease (FRE) BIBREF24 which computes the number of words, syllables and sentences in the text. Higher readability scores indicate that utterance is easier to read and comprehend. Similarly, coherence measures the ability of the dialogue system to produce responses consistent with the topic of conversation. To calculate coherence, we use the method proposed by Dziri et al. BIBREF25.

Results ::: Automated Metrics

We first compare the performance of our approach with the baseline results obtained from Rashkin et al. BIBREF3 that uses a full transformer architecture BIBREF26, consisting of an encoder and decoder. Table TABREF9 provides a comparison of our approach with to the baseline approach. In Table TABREF9, we refer our “Our Model Fine-Tuned” as the baseline fine-tuned GPT-2 model trained on the dialogue and “Our-model Emo-prepend” as the GPT-2 model that is fine-tuned on the dialogues but also conditioned on the emotion displayed in the conversation. We find that fine-tuning the GPT-2 language model using a transfer learning approach helps us achieve a lower perplexity and a higher BLEU scores. The results from our approach are consistent with the empirical study conducted by Edunov et al BIBREF27 that demonstrate the effectiveness of the using pre-trained model diminishes when added to the decoder network in an seq2seq approach. We also perform a comparison between our two models on the metrics of length, diversity, readability and coherence. We find that our baseline model produces less diverse responses compared to when the model is conditioned on emotion. We find that the our emo-prepend model also higher a slightly higher readability score that our baseline model.

Results ::: Qualitative Evaluation

To assess the quality of generations, we conducted a MTurk human evaluation. We recruited a total of 15 participants and each participant was asked to evaluate 25 randomly sampled outputs from the test set on three metrics:

Readability - Is the response easy to understand, fluent and grammatical and does not have any consecutive repeating words.

Coherence - Is the response relevant to the context of the conversation.

Emotional Appropriateness- Does the response convey emotion suitable to the context of the

conversation?

Table TABREF15 shows the results obtained from the human evaluation comparing the performance of our fine-tuned, emotion pre-pend model to the ground-truth response. We find that our fine-tuned model outperforms the emo-prepend on all three metrics from the ratings provided by the human ratings.

Related Work

The area of dialogue systems has been studied extensively in both open-domain BIBREF28 and goal-oriented BIBREF29 situations. Extant approaches towards building dialogue systems has been done predominantly through the seq2seq framework BIBREF0. However, prior research has shown that these systems are prone to producing dull and generic responses that causes engagement with the human to be affected BIBREF0, BIBREF2. Researchers have tackled this problem of dull and generic responses through different optimization function such as MMI BIBREF30 and through reinforcement learning approachesBIBREF31. Alternative approaches towards generating more engaging responses is by grounding them in personality of the speakers that enables in creating more personalized and consistent responses BIBREF1, BIBREF32, BIBREF13.

Several other works have focused on creating more engaging responses by producing affective responses. One of the earlier works to incorporate affect through language modeling is the work done by Ghosh et al. BIBREF8. This work leverages the LIWC BIBREF33 text analysis platform for affective features. Alternative approaches of inducing emotion in generated responses from a seq2seq framework include the work done by Zhou et alBIBREF6 that uses internal and external memory, Asghar et al. BIBREF5 that models emotion through affective embeddings and Huang et al BIBREF7 that induce emotion through concatenation with input sequence. More recently, introduction of transformer based approaches have helped advance the state of art across several natural language understanding tasks

BIBREF26. These transformers models have also helped created large pre-trained language models such as BERT BIBREF9, XL-NET BIBREF11, GPT-2 BIBREF10. However, these pre-trained models show inconsistent behavior towards language generation BIBREF12.

Conclusion and Discussion

In this work, we study how pre-trained language models can be adopted for conditional language generation on smaller datasets. Specifically, we look at conditioning the pre-trained model on the emotion of the situation produce more affective responses that are appropriate for a particular situation. We notice that our fine-tuned and emo-prepend models outperform the current state of the art approach relative to the automated metrics such as BLEU and perplexity on the validation set. We also notice that the emo-prepend approach does not out perform a simple fine tuning approach on the dataset. We plan to investigate the cause of this in future work from the perspective of better experiment design for evaluation BIBREF34 and analyzing the models focus when emotion is prepended to the sequence BIBREF35. Along with this, we also notice other drawbacks in our work such as not having an emotional classifier to predict the outcome of the generated sentence, which we plan to address in future work.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No FA8650-18-C-7881. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of AFRL, DARPA, or the U.S. Government. We thank the anonymous reviewers for the helpful feedback.