Emotional Neural Language Generation Grounded in Situational Contexts

Sashank Santhanam and Samira Shaikh
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC, USA
{ssantha1,samirashaikh}@uncc.edu

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.

1 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 (Vinyals and Le, 2015). 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 (Vinyals and Le, 2015; Li et al., 2016c). These issues also affect engagement with the conversational agent, that leads to short conversations (Venkatesh et al., 2018). Apart from producing engaging responses, understanding the situation and producing the right emotional response to a that situation is another desirable trait (Rashkin et al., 2019).

Emotions are intrinsic to humans and help in creation of a more engaging conversation (Poria et al., 2019). Recent work has focused on approaches towards incorporating emotion in conversational agents (Asghar et al., 2018; Zhou et al., 2018; Huang et al., 2018; Ghosh et al., 2017), 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 (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019). However, their success in language modeling tasks have been inconsistent (Ziegler et al., 2019). 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.* (2019) 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.

2 Approach

Consider the example show in Table 1 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 pro-

duce responses to conversation that are emotionally appropriate to the situation and emotion portrayed. We approach this problem through a lan-

Emotion: Confident **Situation:** I just knew I was going to do well at work this morning.
Speaker: I just knew I was going to do well at work this morning. I was prepared **Listener:** That is the way to go! Keep it up!

Table 1: Example of conversations between a speaker and a listener

guage 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.* (2019) 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* (2019). We ensure that that the sequences are separated by special tokens.

3 Experiments

3.1 Data

In our experiments we use the Empathetic Dialogues dataset made available by Rashkin *et al.* (2019). 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[1].Table 2 provides the basic statistics of the corpus.

[1]More information about the dataset made available on the (Rashkin et al., 2019)

	Train	Valid.	Test
Num. Conversations	19433	2770	2547
Utterances	84324	12078	10973
Avg Length Conversations	4.31	4.36	4.31

Table 2: Statistics of Empathetic Dialogue dataset used in our experiments

3.2 Implementation

In all our experiments, we use the GPT-2 pre-trained 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.[2] The input sentences are tokenized using byte-pair encoding(BPE) (Sennrich et al., 2016) (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 (Holtzman et al., 2019; Ippolito et al., 2019). 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.

3.3 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 (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) from machine translation and text summarization (Liu et al., 2016) 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 (Liu et al., 2016). We report on the BLEU (Papineni et al., 2002) 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* (2017), we also report the diversity metric that helps us measure the ability of the model to promote diversity in responses (Li et al.,

[2]https://github.com/huggingface/pytorch-transformers

2016a). Diversity is calculated as the as the number of distinct unigrams in the generation scaled by the total number of generated tokens (Mei et al., 2017; Li et al., 2016c). 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 (Novikova et al., 2017). We measure readability through the Flesch Reading Ease (FRE) (Kincaid et al., 1975) 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.* (2018).

4 Results

4.1 Automated Metrics

We first compare the performance of our approach with the baseline results obtained from Rashkin *et al.* (2019) that uses a full transformer architecture (Vaswani et al., 2017), consisting of an encoder and decoder. Table 3 provides a comparison of our approach with to the baseline approach. In Table 3, 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* (2019) 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.

4.2 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:

1. Readability - Is the response easy to understand, fluent and grammatical and does not have any consecutive repeating words.
2. Coherence - Is the response relevant to the context of the conversation.
3. Emotional Appropriateness- Does the response convey emotion suitable to the context of the conversation?

Table 5 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.

5 Related Work

The area of dialogue systems has been studied extensively in both open-domain (Niu and Bansal, 2018) and goal-oriented (Lipton et al., 2018) situations. Extant approaches towards building dialogue systems has been done predominantly through the *seq2seq* framework (Vinyals and Le, 2015). 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 (Vinyals and Le, 2015; Venkatesh et al., 2018). Researchers have tackled this problem of dull and generic responses through different optimization function such as MMI (Li et al., 2016b) and through reinforcement learning approaches(Li et al., 2016d). 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 (Li et al., 2016c; Zhang et al., 2018; Wolf et al., 2019).

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.* (Ghosh et al., 2017). This work leverages the LIWC (Pennebaker et al., 2001) 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 al*(2018) that uses internal and external memory, Asghar *et al.* (2018) that models emotion through af-

Experiment	Valid PPL	BLEU	Readability	Coherence	Length	Diversity
Baseline Fine-Tuned (Rashkin et al., 2019)	21.24	6.27	x	x	x	x
Baseline Emo-prepend (Rashkin et al., 2019)	24.30	4.36	x	x	x	x
Our Model Fine-Tuned	**18.32**	7.71	0.78	0.93	9.77	0.0031
Our Model Emo-prepend	19.49	**7.78**	**0.79**	0.93	9.71	**0.0033**

Table 3: Comparison of the performance of our model to the baseline model proposed by Rashkin *et al* (2019) across a variety of automated metrics to provide a thorough comparison. **x** indicates that these metrics were not provided in the Rashkin *et al* (2019) work.

Input Conversation	Model	Outputs
Person A: I think I aced my supervisor exam!	Ground Truth	Woo hoo! That is awesome! Congrats! When will you find out
	Fine Tuned	What was it?
	Emo Prepend	Wow, that's impressive!
Person A: I had pizza at a restaurant a couple days ago and found a hair on my pizza!	Ground Truth	It is disgusting. What did you do?
	Fine Tuned	I was so angry
	Emo Prepend	That's terrible, did you clean it up?

Table 4: Example generations from our two model along with the ground truth responses.

	Readability	Coherence	Emotional Appropriateness
Our Model Fine-Tuned	4.14	3.50	3.70
Our Model Emo-prepend	3.54	3.4	3.19
Ground Truth	3.92	3.86	4

Table 5: Human ratings demonstrating a comparison between our models to the ground truth responses on the metrics of readability, coherence and emotional appropriateness

fective embeddings and Huang *et al* (2018) 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 (Vaswani et al., 2017). These transformers models have also helped created large pre-trained language models such as BERT (Devlin et al., 2019), XL-NET (Yang et al., 2019), GPT-2 (Radford et al., 2019). However, these pre-trained models show inconsistent behavior towards language generation (Ziegler et al., 2019).

6 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 (Santhanam and Shaikh, 2019) and analyzing the models focus when emotion is prepended to the sequence (Clark et al., 2019). 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.

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

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

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2018. Augmenting neural response generation with context-aware topical attention. *arXiv preprint arXiv:1811.01063*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016c. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016d. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2018. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2017. Coherent dialogue with attention-based language models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, San Francisco, CA.

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

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *arXiv preprint arXiv:1905.02947*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. *arXiv preprint arXiv:1909.10122*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*.

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

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.