Eliciting Rich Positive Emotions in Dialogue Generation

Ziwei Gong
Department of Computer Science
Columbia University
zg2272@columbia.edu

Qingkai Min
School of Engineering
Westlake University
minqingkai@westlake.edu.cn

Yue Zhang
School of Engineering
Westlake University
zhangyue@westlake.edu.cn

Abstract

Positive emotion elicitation aims at evoking positive emotion state in human users in open-domain dialogue generation. However, most work focuses on inducing a single-dimension of positive sentiment using human annotated datasets, which limits the scale of the training dataset. In this paper, we propose to model various emotions in large unannotated conversations, such as joy, trust and anticipation, by leveraging a latent variable to control the emotional intention of the response. Our proposed emotion-eliciting-Conditional-Variational-AutoEncoder (EE-CVAE) model generates more diverse and emotionally-intelligent responses compared to single-dimension baseline models in human evaluation.

1 Introduction

In human communication theory, intentionality (intention of speakers) and effectiveness (effects of conversations) are key factors to a conversation (Littlejohn and Foss, 2010; Lindquist et al., 2015; Morick, 1971), both of which can be exhibited by emotions (Dezecache et al., 2013). There has been research on dialogue systems for generating human-like, emotionally intelligent responses (Huang et al., 2018; Zhou et al., 2017; Liu et al., 2016). However, existing work focuses on generating utterances with targeted emotion to express, yet few studies explore how one's emotion is affected by utterances, nor the intentionality of generated sentences (Kao et al., 2019; Zhou et al., 2017).

One exception is emotion elicitation, which considers generating responses that elicit a pre-specified emotion in the other party (Hasegawa et al., 2013). Though natural for humans to recognize and intentionally influence other's emotions, eliciting pre-specified emotions is challenging for dialogue models (Rashkin et al., 2019). Prior work has evolved from statistical response generator (Hasegawa et al., 2013) to neural networks (Lubis et al., 2018; Li et al., 2020). All existing models focus on eliciting a single coarse-grained sentiment: positive emotion (Ma et al., 2020; Rashkin et al., 2019). However, as shown in Figure 1, positive sentiment can include more fine-grained emotions such as "*Hopeful*", "*Joy*" and "*Surprise*", which can further serve to deepen the model's understanding of *effect*, if not *intention*. By incorporating more emotions in training, it ameliorates the performance in the elicitation of positive emotions. Besides, existing work is mostly based on small-scale human-annotated datasets, which limits its capacity of eliciting various emotions.

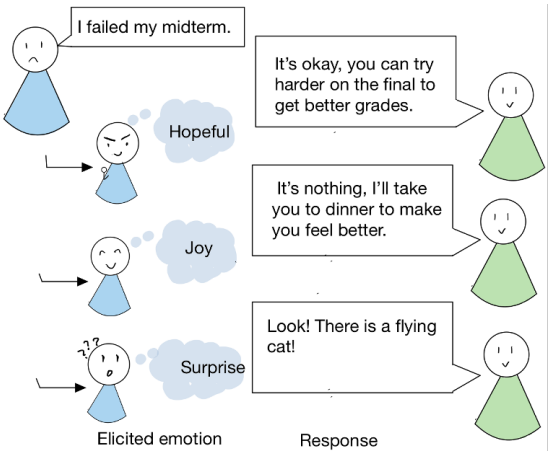


Figure 1: Examples of different responses that elicit different positive emotions.

We fill this gap by proposing the first model for emotion elicitation that controls the generation of responses that elicit various pre-specified emotions. Due to difficulties in annotation, we represent the elicited emotions using latent variables in order to take full advantage of the large-scale unannotated dataset, choosing Conditional Variational Auto-encoder (CVAE) as a backbone (Zhao et al., 2017; Tikhonov et al., 2019; Chen et al., 2019). Two discriminators are further used to control the generation of responses.

We reconstruct a recent multi-modal MEmoR dataset (Shen et al., 2020), extracting useful text data for our task, and conduct experiments on nine primary emotions[1]. A large-scale TV show dataset is used to pretrain the model in an unsupervised fashion. Results show that our model outperforms the state-of-the-art single-emotion elicitation model (Li et al., 2020), achieving higher accuracy for eliciting positive emotions. Using all emotions in pretraining and finetuning produces the best performance in eliciting positive emotions. In addition, our results show that rich emotion elicitation is a challenging task for current neural models and there is a need for more effective few-shot learning. Our code and data will be available at https://github.com/taolusi/EECVAE.

2 Related Work

Emotion Elicitation Hasegawa et al. (2013) investigates a statistical response generator guided by predicted future emotions. Recent approaches extend the Hierarchical Recurrent Encode-Decoder model (Serban et al., 2016) by adding a separate layer of emotion modules to induce a positive emotion (Lubis et al., 2018), and propose an encoder-decoder adversarial model with two discriminators to increase emotion-awareness or empathetic dialogue generation (Li et al., 2020). Emotion-grounded generation is also used to guide empathetic dialogue generation (Majumder et al., 2020; Lin et al., 2019). Different from the above, we are the first to model the elicitation of rich positive emotions using one neural network.

Conditional Variational Autoencoder (CVAE) CVAE is an extension of VAE (Sohn et al., 2015; Bowman et al., 2016; Kingma and Welling, 2013; Salimans et al., 2015), which has been used for dialogue generation (Chen et al., 2019) by introducing a latent variable to capture discourse-level variations (Zhao et al., 2017). We take CVAE as a basis for extension, adding two discriminator components, which has been shown useful for single-emotion elicitation (Hu et al., 2017).

3 Baseline: EmpDG

As shown in Figure 2(a), EmpDG (Li et al., 2020) is a sequence-to-sequence dialogue response generation model that enhances the elicitation of positive emotion through empathy. During encoding, the dialogue context is represented as a vector c; during decoding, the generator uses two CNN discriminators to generate an n-token response x. Specifically, a semantic discriminator D_{sem} measures the distance from the generated response to the gold response, while an emotional discriminator D_{emo} specifies the degree of empathy in responses. Both discriminators are used to extend a Transformer model (Vaswani et al., 2017), serving as semantic and emotional enhancements, respectively. For training, the loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{gen} + \mathcal{L}_{sem} + \mathcal{L}_{emo}, \quad (1)$$

where $\mathcal{L}_{gen}$ denotes the objective for the auto-regressive generator, which uses a standard maximum likelihood function, and $\mathcal{L}_{sem}$ and $\mathcal{L}_{emo}$ denote the loss functions of the two discriminators, both of which are calculated by minimizing the Wassertein-1 distance between distributions of golden responses and the generated responses.

D_{sem} uses the next utterance directly as user semantic feedback in $\mathcal{L}_{sem}$ and D_{emo} extracts user emotional feedback from the emotional words in the utterance in $\mathcal{L}_{emo}$. Instead of using explicit feedback from annotated labels, EmpDG extracts implicit information from the next utterance as feedback for semantic and emotional guidance of targeted response. Although such method alleviates the burden of annotating emotion labels, the extracted feedback can be sparse and noisy, which introduces uncertainty in empathetic generation. To address this issue, we introduce a latent variable to represent the emotion labels, which can be learned in an unsupervised way.

4 Model

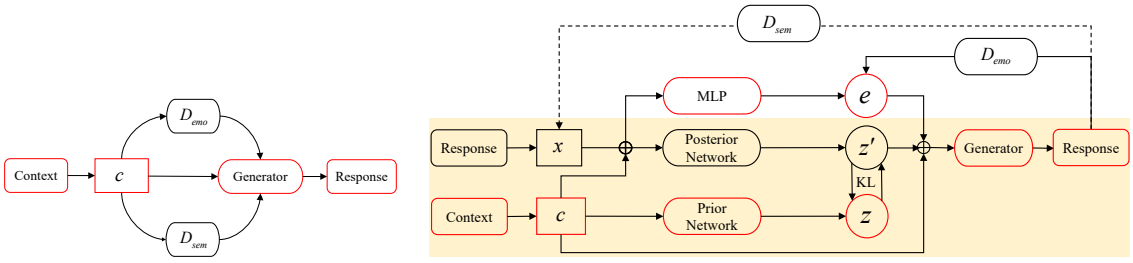
The overall structure of our model is shown in Figure 2(b). It can be seen as an extension of CVAE (in yellow) with a latent variable and two discriminators to elicit multiple emotions.

4.1 CVAE for Dialogue Generation

A dialogue-based CVAE (Hu et al., 2017) generates responses conditioned on the dialogue context. Briefly, the generative process of a dialogue-based CVAE is composed of two steps:

1. Sample a latent vector z from prior network $p_\theta(z|c)$, where c is the dialogue context.
2. Generate a response x through a generator $p_\theta(x|z,c)$, given dialogue context c and latent

[1] Plutchik (1980)'s 9 primary emotions: joy, anger, disgust, sadness, surprise, fear, anticipation, trust and neutral.



(a) Baseline mode: EmpDG

(b) Our EE-CVAE model.

Figure 2: Training illustration of our model and a baseline model. (Red components are used for testing. CVAE in yellow background. Dashed arrow denotes a discriminator.)

vector z, where θ denotes the parameters of a generative network.

For training, the objective formula is:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\theta,\phi) =& \mathbf{E}_{q_\phi(z|c,x)}[\log p_\theta(x|z,c)] \\ & - \text{KL}(q_\phi(z|c,x)\|p_\theta(z|c)) \\ & \leq \log p(x|c). \end{aligned} \quad (2)$$

4.2 Adding Emotion Elicitation Function

To model elicited emotion, we augment CVAE with a latent variable e, which is used to control the generation of a response together with the unstructured variable z. The training objective is:

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\theta,\phi) =& \mathbf{E}_{q_\phi(z|c,x)q_\phi(e|c,x)}[\log p_\theta(x|z,c,e)] \\ & - \text{KL}(q_\phi(z|c,x)\|p_\theta(z|c)) \\ & \leq \log p(x|c), \end{aligned} \quad (3)$$

where the first term is used to minimize the reconstruction error given the posterior network $q_\phi(z|c,x)$ and $q_\phi(e|c,x)$, and the second term is the KL-divergence of the posterior network $q_\phi(z|c,x)$ and the prior network $p_\theta(z|c)$, which can be viewed as a regularisation term.

Inspired by the idea of style transfer (Hu et al., 2017), a discriminator D_{emo} is used to enforce the generator to produce coherent emotions:

$$\mathcal{L}_{\text{Attr},e}(\theta) = \mathbb{E}_{p(z)p(e)}\left[\log q_{D_{emo}}\left(e \mid \widetilde{G}_\tau(z,e)\right)\right], \quad (4)$$

where $\widetilde{G}_\tau(z,e)$ denotes the generated response.

Similarly, the variational encoder is reused to separate unrelated attributes from e by enforcing them to be fully captured by z. It can be considered as another discriminator D_{sem}:

$$\mathcal{L}_{\text{Attr},z}(\theta) = \mathbb{E}_{p(z)p(e)}\left[\log q_{D_{sem}}\left(z \mid \widetilde{G}_\tau(z,e)\right)\right]. \quad (5)$$

Combining Eqs.(2)-(4), the formal objective is:

$$\min \mathcal{L}_G = \mathcal{L}_{\text{VAE}} + \lambda_e \mathcal{L}_{\text{Attr},e} + \lambda_z \mathcal{L}_{\text{Attr},z}, \quad (6)$$

where λ_e and λ_z are balancing parameters.

To accurately infer elicited emotions expressed in a sentence, the discriminator D_{emo} is formulated as a sentence classifier. In contrast to the latent variable z, which is learned in fully unsupervised autoencoder training, e is further trained to entail designated emotions using a small set of labeled examples. Specifically, we follow a wake-sleep training schedule (Hu et al., 2017), training the generator before the discriminator.

5 Dataset

We reconstruct the multi-modal MEmoR dataset (Shen et al., 2020) to fit our task and conducted human evaluations to validate the usability in a single modality. MEmoR contains video, audio, and text information of clips from the TV show The Big Bang Theory, with emotion labels given on each character in every clip. We use only the textual data and consider non-speakers' emotions to be the elicited emotions by an utterance. Manual decision is made on whether a target emotion can be elicited using text context only, in order to filter dialogues. Our reconstructed dataset has a annotator agreement of 80% accuracy ($Cohen's\ \kappa = 0.491$). The reconstructed corpus has 22,732 utterances and we split the data into training (18,943), dev (1,894), and test (1,894). Nine emotions are labeled in total in the dataset according to the emotion classification of (Plutchik, 1980), out of which 3 positive emotions are chosen as the model output [2].

6 Experiments

Experimental Setup For both EmpDG and EE-CVAE, we use more than 200k utterances from Friends (Zahiri and Choi, 2017) and Open Subtitles[3] datasets for pre-training the generator module, and the reconstructed MEmoR dataset to train the discriminators. Since one EmpDG model can

[2]We show in Section 6 that including negative emotions in model training helps better generates positive emotions.

[3]http://www.opensubtitles.org/

Model	TBBT - 9			
	PPL	**Avg. len**	**KL**	**Acc.**
EmpDG	667.4	8.7	-	
EmpDG$_{pre}$	462.2	9.2	-	0.290
Ours	196.4	14.3	25.9	
Ours$_{pre}$	91.5	13.2	14.0	0.448

Table 1: Results of models generation in comparison. "-" indicates not applicable, the average length for EmpDG is not reported because the generation results are unacceptable for most emotion categories. Human evaluations are conducted for selected models due to limited resources.

only be trained with a single emotion category, we split the dataset according to different emotion categories to train different models.

We follow former dialogue generation and emotion elicitation studies (Li et al., 2020; Shen et al., 2019; Hasegawa et al., 2013) and report perplexity (PPL) results, KL loss term, average length of sentences, and emotion accuracy (human evaluation).

Results The overall results are shown in Table 1. The perplexity of our method is lower than that of the EmpDG baseline, which shows that our model can generate more fluent responses. This can be because of the two-channel model structure by EmpDG, where implicit emotions can interfere with generation by introducing noise. In contrast, the use of CVAE can be also more effective in isolating the influence of emotion signals.

The output length of our model is 13.17 on average, 43% longer than that of the baseline. We find that the baseline model tends to give shorter and more generic outputs such as "*really*' and "*what*", as can be seen in Table 2. The maximum output lengths are similar, 31 (baseline) and 39 (ours) respectively. This again shows the efficaciousness of the discriminator D_{sem} in our model, as compared to that of the baseline.

For both models, the perplexity is lower when pre-training is added for the generation module. Compared with EmpDG, the rate of PPL reduction is markedly larger for our model. This suggests that our CVAE structure can make better use of wake-sleep learning. In addition, pre-training reduces the KL of our model, demonstrating that more data allows the model to better separate the emotion signal e and the semantic context z.

With regard to the emotion accuracy, we conducted human evaluation, where experts are asked to assign a binary value indicating whether the elicited emotion is correct as expected. Our model gives 44.8% accuracy, significantly better than

Context:Well, you be sure to let us know when you win the nobel prize for boysenberry.
Golden (anticipation): Hey. **EmpDG** (anticipation): yeah . **Ours** (joy): Oh , what a gentleman? **Ours** (trust): Wow , I really appreciate it.
Context: Aw, Amy, that was lovely. You know, this is fun. Let's do more. Someone else say something wonderful about me.
Golden (joy) Sheldon, I don't think everyone ... **EmpDG** (joy): What is great. **Ours** (joy) Oh, sure. Mmm. I told you, he's got too many. **Ours** (anticipation) And you.

Table 2: Samples with emotion elicitation.

	Setting1	Setting2	Setting 3	Tie
Anticipation	.47	.32	.19	.02
Joy	.55	.215	.215	.02
Trust	.54	.17	.27	.02
All	.51	.25	.22	.02

Table 3: Results comparing three settings with the percentage of times one model is considered the best when eliciting different positive emotions. Setting 1: modeling all emotions in pretraining and fine-tuning. Setting 2: modeling all emotions in pretraining, fine-tuning with only positive emotions. Setting 3: modeling only positive emotion.

29.0% of the baseline. This shows the advantage of using a latent variable for modeling rich emotions, compared to hard-coding one emotion in a multi-encoder model. It also shows the effectiveness of our model in pretraining.

The Effect of Modeling Negative Emotions Intuitively, adding negative emotions to model training can improve positive emotion elicitation due to two reasons. First, the amount of training data is enlarged in both pretraining and fine-tuning. Second, awareness of negative emotions enhances that of positive emotions, which is similar to adversarial learning in principle (Goodfellow et al., 2015; Miyato et al., 2016; Mądry et al., 2017).

We conduct ablation by removing negative utterances in pretraining and fine-tuning, respectively, leading to three settings (Table 3)[4]. We randomly select 164 samples and perform human evaluation to select a response from the three models that can best elicit different positive emotions. As can be seen from the results, our model produces the best results in all positive emotions in setting 1, verifying our intuition above.

[4]We use a pretrained sentiment-analysis classifier to remove utterances that elicit negative emotions from raw data: https://huggingface.co/transformers/task_summary.html?highlight=sentimen%20analysis

7 Conclusion

We provided the first discussion on rich emotion elicitation in open-domain dialogue generation, incorporating various positive emotions with a framework that extends CVAE with a latent emotion variable equipped with two discriminators. Results show that rich emotion elicitation is a challenging task and our model gives more reliable utterances compared with a state-of-art model for single emotion elicitation, and introducing negative emotions in pretraining benefits the model's ability to elicit positive emotions.

8 Ethical Statement

8.1 Annotation

To facilitate research, we reconstruct a dataset from a large unannotated dataset Open Subtitle and a small annotated dataset MEmoR, which is annotated with speakers and non-speaker emotions. Both datasets are publicly available and are collected from TV shows. We use the emotion elicited in actors (transcripts) as elicited emotions in our research. To verify that the approach is valid, a blind check was conducted on a randomly selected set, where two annotators were asked to make manual decisions on whether a target emotion can be elicited. Annotators are recruited college students from universities whose primary teaching language is English, and compensated with course credit. Our reconstructed dataset has a annotator agreement of 80% accuracy ($Cohen's\ \kappa$ = 0.491). In our researches, for the purpose of validating the dataset and evaluate model results, annotators are only asked to evaluate if the emotional labels were valid, not to offer personal emotion feedback. To ensure reprehensibility, we would release the reconstructed dataset along with the paper at http://XXX.

8.2 Elicit Rich Emotions

Our model elicits only positive emotions, but our dataset contains labeling of negative emotions, which exist in the TV show dialogues naturally. We demonstrate that using all emotions would not only benefit the differentiation between all emotions, but also help the model to better elicit positive emotions. Naturally, there are emotions that are considered to be more positive and the ones that are more negative. We intend to model various emotions so that a system is more aware of the correlation between intention and response. Consequently, a model can, for example, be aware that a certain type of answer may result in sadness and thus avoid it. In addition, a model can better understand user attitudes also by capturing such intentions in them. However, the modeling of multi-various emotions is not necessarily for the purpose of eliciting them. In application, we only elicit emotions that are considered to be positive, as our goal is to better elicit rich positive emotions in dialogue.

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Junfan Chen, Richong Zhang, Yongyi Mao, Binfeng Wang, and Jianhang Qiao. 2019. A conditional vae-based conversation model. In *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding*, pages 165–174, Singapore. Springer Singapore.

Guillaume Dezecache, Hugo Mercier, and Thomas C. Scott-Phillips. 2013. An evolutionary approach to emotional communication. *Journal of Pragmatics*, 59:221 – 233. Biases and constraints in communication: Argumentation, persuasion and manipulation.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *stat*, 1050:20.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 964–972, Sofia, Bulgaria. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Chenyang Huang, Osmar Zaïane, 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, New Orleans, Louisiana. Association for Computational Linguistics.

C. Kao, C. Chen, and Y. Tsai. 2019. Model of multi-turn dialogue in emotional chatbot. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 1–5.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Kristen A. Lindquist, Jennifer K. MacCormack, and Holly Shablack. 2015. The role of language in emotion: predictions from psychological constructionism. *Frontiers in Psychology*, 6:444.

Stephen W Littlejohn and Karen A Foss. 2010. *Theories of human communication*. Waveland press.

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, Austin, Texas. Association for Computational Linguistics.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and S. Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *AAAI*.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50 – 70.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *stat*, 1050:9.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *CoRR*, abs/1904.02295.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *stat*, 1050:7.

Harold Morick. 1971. Intentionality, intensionality and the psychological. *Analysis*, 32(2):39–44.

R Plutchik. 1980. Plutchik's wheel of emotions. *Accessed (Dec 2, 2019) at: https://www. researchgate. net/publication/234005320_Discovering_Basic_*

Emotion_Sets_via_Semantic_Clustering_on_a_TwitterCorpus/figures.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Tim Salimans, Diederik Kingma, and Max Welling. 2015. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226, Lille, France. PMLR.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. 2019. Towards generating long and coherent text with multi-level latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2079–2089, Florence, Italy. Association for Computational Linguistics.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 493–502. ACM.

Kihyuk Sohn, Xinchen Yan, and Honglak Lee. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3483–3491, Cambridge, MA, USA. MIT Press.

Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. 2019. Style transfer for texts: Retrain, report errors, compare with rewrites. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3936–3945, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 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 (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *CoRR*, abs/1704.01074.

Parameter	Value
Embedding Size	200
Latent Variable Size	200
Batch Size	30
Learning Rate	0.001
Optimizer	Adam
Grad Clip	5

Table 4: Model parameter settings.

A Model Parameter Settings

The parameters of our model are shown in Table 4.

B Human evaluation

B.1 MEmoR

Since the MEmoR dataset was originally annotated in a multi-module setting, we did human envaluation on the MEmoR dataset to ensure the emotion annotated through multimodal scenario can also be perceived through plain text. we used methods propsed by (Mir et al., 2019). Firstly, 400 preprocessed dialogues were randomly selected from the MEmoR dataset. Two group of annotators were asked to choose "yes" or "no" on the original emotion labels based purely on the text, or scripts, of each dialogue. The final resultsa: both annotator marked "yes" on 80 percent of the samples.

Used to measure the inner annotator agreement, the Cohen's Kappa value, calculated without weights, is 0.491 (z = 10.3). According to Landis and Koch's interpretation, 0.491 means two annotators reached moderate agreement.

B.2 Generations

In evaluation of the generation, 300 randomly dialogues spread across nine pre-specified emotions and their corresponding generations from both models are sampled and evaluated by two annotators. Each utterance were evaluated by two annotators. Annotators were asked to judge if the utterance in given content could successfully elicit target emotion, and ignore minor grammar mistakes in generation. Generation that are so grammatical incorrect that one cannot tell the meaning were marked as unsuccessful.

Used to measure the inner annotator agreement, the Cohen's Kappa value calculated without weights is 0.323 and 0.319 when evaluating the generation results in EmpDG and CVAE. For each model 300 randomly dialogues spread across 9 prespecified emotions are sampled and evaluated.

C Data Preprocessing

The train, dev, test set split is 10:1:1 for the MEmoR dataset using random spliting. The dataset and splits will be published together with our code.

To use the dataset on chosen baseline EmpDG, we split the MEmoR data by emotion categories and run a EmpDG model on each category. None of the models used any meta information.