GENERATING EMPATHETIC RESPONSES BY LOOKING AHEAD THE USER'S SENTIMENT

Jamin Shin*   *Peng Xu**   *Andrea Madotto*   *Pascale Fung*

Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
jayshin.nlp@gmail.com   pasacle@ece.ust.hk

ABSTRACT

An important aspect of human conversation difficult for machines is conversing with empathy, which is to understand the user's emotion and respond appropriately. Recent neural conversation models that attempted to generate empathetic responses either focused on conditioning the output to a given emotion, or incorporating the *current* user emotional state. However, these approaches do not factor in how the *user* would feel *towards* the generated response. Hence, in this paper, we propose *Sentiment Look-ahead*, which is a novel perspective for empathy that models the *future* user emotional state. In short, Sentiment Look-ahead is a reward function under a reinforcement learning framework that provides a higher reward to the generative model when the generated utterance improves the user's sentiment. We implement and evaluate three different possible implementations of sentiment look-ahead and empirically show that our proposed approach can generate significantly more empathetic, relevant, and fluent responses than other competitive baselines such as multitask learning.

Index Terms— Natural Language Processing, Dialogue Systems, Empathetic Chatbots, Sentiment Look-ahead

1. INTRODUCTION

Showing empathy is a core attribute of human conversations as our daily interactions often involve discussions of emotional situations ranging from recent job promotions to even funerals, or about the difficulties of raising children. Hence, it is natural to think that modeling empathy and eliciting it in chatbots are crucial towards bringing them even closer to humans.

Moreover, there are also practical benefits to incorporating such emotional understanding capabilities. For instance, [1] has reported that addressing the users' emotions reduces the probability of dialogue breakdowns, and [2] has shown that it can enhance overall user satisfaction. In addition, talking about emotional experiences is known to help relieve stress [3] which can also be seen from the large number of users talking about their emotional situations with Microsoft XiaoIce [4].
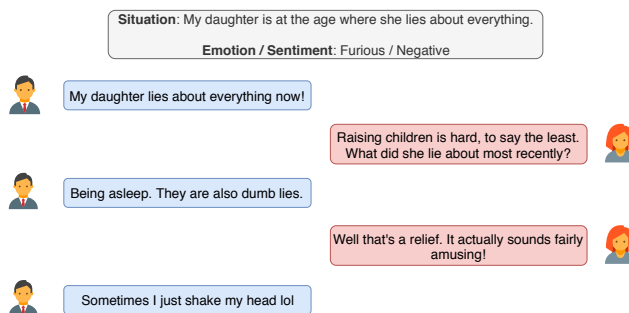

Fig. 1. Example conversation from Rashkin et al., 2019 [7]. A speaker (blue) talks about his emotional situation, and the listener (red) is supposed to react empathetically. The task of empathetic dialogue generation aims to model the latter.

As a result, there have been several attempts to model empathy in dialogues. Some initial works include that of [5] which manages the dialogue based on user emotions, and creating affective listeners that can respond in terms of both content and affect level [6]. However, these studies were mainly rule-based systems that are limited in scalability to dataset size and generalizability to different domains and situations.

Recently, neural conversation models [8] have been successful in generating fluent and relevant responses. However, their responses are widely known to promote *dull* and *generic* responses due to the maximum likelihood objective [9, 10] that does not factor in any kind of emotional exchange, or empathy. Meanwhile, several recent works have tackled the problem of empathetic dialogue response generation, which is understanding the user's emotion and responding appropriately, mainly on two directions. The first line of work has been successful in controlling and conditioning the generated responses to certain sentiments, emotions, and emojis [11, 12, 13, 14]. Meanwhile, others have worked on more data-driven approaches by training a model to jointly predict the current emotional state and generate a response [15, 7].

While both cases have been successful in generating empathetic and emotional responses to a certain extent, but have neglected some crucial points in empathetic dialogue response
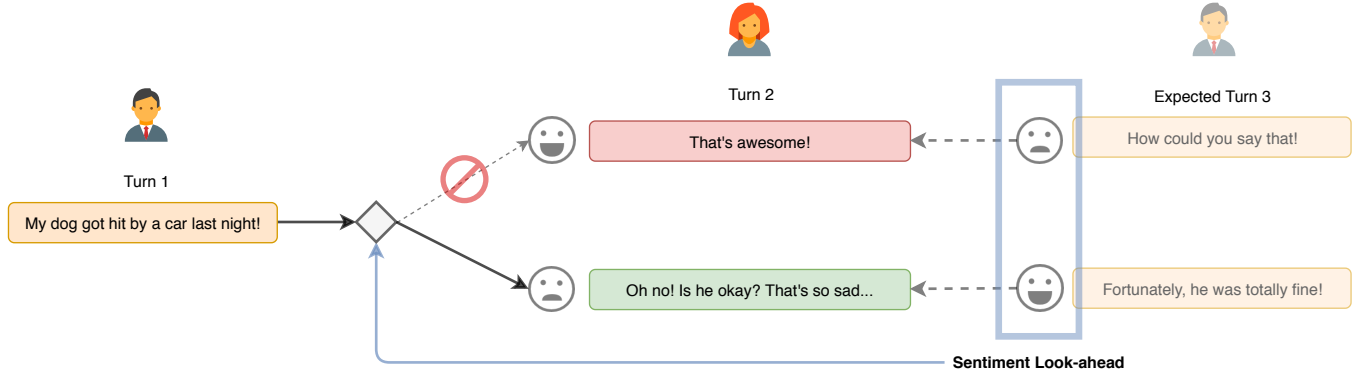
Fig. 2. Illustration of Sentiment Look-ahead. During conversations, empathetic people consider the impact of their utterances before speaking.

generation. 1) The first approach - namely controlled text generation - assumes that the emotion to be conditioned on is given as an input, but we often do not know which emotion is appropriate in order to generate an empathetic response. 2) The latter takes the assumption that by understanding the user's *current* emotion, the model will implicitly learn how to respond empathetically. However, without providing any additional inductive bias, recognizing the current emotional state does not necessarily imply that the model will respond appropriately towards that emotion.

On the other hand, to cope with the above issues, we propose *Sentiment Look-ahead* to directly address the problem of *responding appropriately*. Intuitively, an empathetic person would naturally consider the conversation partner's feelings before speaking and, in turn, trigger a more positive sentiment response from the user. To elaborate, we cast this problem as reinforcement learning in which the reward signal to maximize is the (predicted) sentiment of the *next user turn*.

Finally, we propose three different implementations of the *sentiment look-ahead* reward function. Based on thorough automatic / human evaluations, we show that our approach significantly outperforms other models in terms of Empathy, Relevance, and Fluency, verifying the effectiveness of our novel viewpoint about modeling empathy.

2. METHODOLOGY

Our model has two modules: the policy model and the reward. The policy model takes input x and generates a response $\hat{y}$ and the reward model evaluates the generated response with a predefined criteria and updates the policy model accordingly.

2.1. Notations

Considering a conversation between a user and the system, we can represent an input dialogue history as an alternating sequence of turns between the two parties as such: $x =$

$[u^1; s^1; u^2; s^2; \cdots ; u^T]$, where T is the number of user/system turns and u^* and s^* denote user and system utterances, respectively. The turns in x are flattened as a single sequence of words: $x = [w^1, w^2, \cdots, w^N]$, where N is the number of words in the dialog history. Subsequently, we denote the true output system response s^T towards the input x as y and the prediction as $\hat{y}$, where y and $\hat{y}$ are both sequences of words. We also denote d as the dimensionality of the encoder and decoder, and V as the size of the vocabulary.

2.2. Reward Functions: Sentiment Look-ahead

In a nutshell, given the generated system response s^t, an arbitrary function f will output a score $R_* \in [0, 1]$ that determines how good the input is, which is then used as the reward for learning the optimal policy. In short, the reward score comes from the next user turn, rather than the currently generated system turn. In the following, we propose and explain three different implementations for sentiment look-ahead reward function and one method for control. We denote function $g(\cdot) \in [0, 1]$ as a pre-trained sentiment classifier that outputs the probability of the input sentence's sentiment being positive.

Forward R_F is the most naive implementation of sentiment look-ahead by directly aiming to predict the next user sentiment towards the generated system response. This reward function outputs the probability of the next user turn sentiment being positive. Hence, f is a GRU trained to predict $g(u^{t+1})$ given s^t.

Improvement R_I takes a slightly different approach. Because the labels for learning R_F and g are the same 70% of the time, we hypothesize that R_F will not learn how to predict the next sentiment, but rather the current one. Instead, we reward our system when the sentiment of the next user turn is predicted to *improve*. Thus, this reward function predicts *whether the sentiment of* u^{t+1} *improves* compared to the current user turn u^t. Hence, f is a GRU trained to predict $\mathbb{1}(g(u^{t+1}) > g(u^t))$ given s^t.

Simulation R_S is directly given by the sentiment of the *simulated* next user turn $\hat{u}_{t+1}$. As both of the above methods use the *predicted* user reaction instead of the *actual* one, we aim to cope with this issue through user simulation with a pre-trained Seq2Seq model. The sentiment difference $R_S = g(\hat{u}_{t+1}) - g(x)$ is used as the reward.

Current R_C is given by the current sentiment of the input $R_C = g(s^t)$. This model is used as control in order to isolate the contributions of sentiment look-ahead models.

2.3. Policy Model: Seq2Seq with Attention

The Policy Model is a Seq2Seq model [16] based on GRU [17] and dot product attention [18]. To elaborate, the encoder BiGRU takes the flattened input x and generates the encoder hidden states $\mathbf{h}_{enc} = [h_{enc}^1, h_{enc}^2, \cdots h_{enc}^N] \in \mathbb{R}^{d \times N}$. For each decoding step t, the decoder generates the decoder hidden state $h_{dec}^t \in \mathbb{R}^{1 \times d}$ given the previous decoder state h_{dec}^{t-1}, the previous target token y^{t-1}, and the context vector using attention c^t:

$$\mathbf{h}_{enc} = \text{BiGRU}(x) \tag{1}$$

$$h_{dec}^t = \text{GRU}(h_{dec}^{t-1}, y^{t-1}, c^{t-1}) \tag{2}$$

$$a_t = \text{softmax}(h_{dec}^t \cdot \mathbf{h}_{enc}) \tag{3}$$

$$c^t = \sum_{i=1}^{N} a_t^i h_{enc}^i \tag{4}$$

, where y^0 is <SOS> token, $h_{dec}^0 = \vec{0}$, and $c^0 = \vec{0}$.

The decoder hidden state and context vector is then used for generating a conditional probability distribution the next token $\hat{y}_t$ is sampled from:

$$h_t = [c^t; h_{dec}^t] \tag{5}$$

$$h_t^* = \tanh(h_t \cdot \mathbf{W}_r) \tag{6}$$

$$\hat{y}_t \sim p_\theta(y \mid y^{t-1}, h_{dec}^{t-1}) = \text{softmax}(h_t^* \cdot \mathbf{W}_{out}/\tau) \tag{7}$$

, where $\mathbf{W}_r \in \mathbb{R}^{2d \times d}$ is a trainable parameter that reduces the hidden state dimensionality, $\mathbf{W}_{out} \in \mathbb{R}^{d \times V}$ is a trainable matrix that maps the reduced state to the output vocabulary space, τ refers to softmax temperature, and θ is parametrized by $\{\text{BiGRU}, \text{GRU}, \mathbf{W}_r, \mathbf{W}_{out}\}$.

2.4. Policy Learning

We follow the MIXER algorithm [19] which basically pre-trains the Policy Model and incrementally updates it with REINFORCE [20]. First, we optimize the following Maximum Likelihood Estimation (MLE) loss function $\mathcal{L}_{\text{MLE}}$ to pre-train the Policy Model:

$$\mathcal{L}_{\text{MLE}} = -\log p(y) = -\log p(y^1, \cdots, y^M) \tag{8}$$

$$= -\frac{1}{M} \sum_{t=1}^{M} \log p(y^t \mid y^1, \cdots, y^{t-1}) \tag{9}$$

, where M is the number of words to generate and $p(y^t \mid \cdot)$ is parametrized by Equation 7.

We then use REINFORCE which defines the loss function $\mathcal{L}_{RL}$ as the negative expected reward:

$$\hat{R}_t = h_{dec}^t \cdot \mathbf{W_b} \tag{10}$$

$$\mathcal{L}_{\text{RL}} = -\frac{1}{M} \sum_{t=1}^{M} (R_* - \hat{R}_t) \log p(\hat{y}_t \mid \hat{y}_1, \cdots, \hat{y}_{t-1}) \tag{11}$$

, where $\mathbf{W_b} \in \mathbb{R}^{d \times 1}$ is a trainable parameter, $\hat{R}_t$ is the baseline reward, and R_* is an arbitrary reward. The baseline reward is used in order to reduce the variance of the actual reward and is a linear model trained by minimizing the Mean Squared Error (MSE) loss between R_* and $\hat{R}_t$:

$$\min_{\mathbf{W_b}} \frac{1}{T} \sum_{t=1}^{T} |R_* - \hat{R}_t|^2 \tag{12}$$

2.5. Hybrid Training

As the reward function from REINFORCE may deteriorate the model from generating fluent and relevant response, we follow [21, 13] to stabilize the training process by employing a mixed learning objective which augments the policy learning objective after pre-training with $\mathcal{L}_{\text{MLE}}$ as such:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{RL}} + (1 - \lambda) \mathcal{L}_{\text{MLE}} \tag{13}$$

, where λ is a hyper-parameter that interpolates each loss.

3. EXPERIMENT SETTING

We mainly conduct our experiments and evauations on the EmpatheticDialogue [7] dataset, which consists of 25k conversations between a *Speaker* and a *Listener* about a given emotional situation. The dataset provides evenly distributed 32 emotion labels that we map down to positive/negative sentiments. To continue, in order to promote diversity of the response generation, we pre-train the policy model on a combined dataset of EmpatheticDialogue, DailyDialog [22], and PersonaChat [23].

Before training the sentiment look-ahead reward functions, we first have to train the sentiment classifier $g(\cdot)$ from Section 2.2, which is also R_C. For training the sentiment classifier, we fine-tune the pre-trained BERT [24] on SST-2 [25] dataset and the situation texts from EmpatheticDialogue by mapping emotions to binary sentiment labels as mentioned earlier. We then use this to label the sentiments of each turn, generate forward sentiment and sentiment improvement labels, and train the GRU-based reward models R_F and R_I.

We use the pre-trained BERT *base* [24] model, 300 dimensional pre-trained FastText[1] word embeddings, hidden size $d = 300$ for GRU cells, and tie the embedding weights of

---

[1]https://fasttext.cc/

Automatic Evaluation Results

	Sentence Length	Distinct 1-gram	Distinct 2-gram	Distinct 3-gram	Average BLEU
Human	14.463	0.065	0.386	0.725	-
Seq2Seq	11.494 ± 0.0281	0.019 ± 0.0003	0.115 ± 0.0003	0.263 ± 0.0014	$\mathbf{6.558} \pm 0.0520$
MultiSeq	$\mathbf{13.264} \pm 0.0298$	0.015 ± 0.0003	0.102 ± 0.0007	0.246 ± 0.0012	6.250 ± 0.0433
Current	12.799 ± 0.0928	0.015 ± 0.0004	0.100 ± 0.0009	0.240 ± 0.0023	6.275 ± 0.0250
Forward	13.177 ± 0.0122	0.020 ± 0.0004	0.123 ± 0.0004	0.285 ± 0.0004	6.292 ± 0.0289
Improvement	13.098 ± 0.0368	$\mathbf{0.026} \pm 0.0004$	$\mathbf{0.170} \pm 0.0013$	$\mathbf{0.379} \pm 0.0025$	6.317 ± 0.0144
Simulation	12.154 ± 0.0432	0.011 ± 0.0001	0.080 ± 0.0004	0.202 ± 0.0015	6.192 ± 0.0629

Table 1. Automatic metrics measuring fluency of generated responses. *Improvement* model significantly outperforms others in all Distinct N-gram measures. It also achieves the second highest in Sentence Length and Avereage BLEU score.

the decoder's output layer as in [26]. We use learning rates of $[1e-3, 1e-4]$ and hybrid training ratio λ of 0.5. Finally, we use top-k sampling [27] using $k = 40$ along with softmax temperature τ of 0.5.[2]

3.1. Evaluation

Automatic Metrics We evaluate 4 metrics that represent how fluent and relevant the generated responses are:

- **Sentence Length**: Measure average length of generated utterances.

- **Distinct $\{1, 2, 3\}$-grams**: Count percentage of unique $\{1, 2, 3\}$-grams as in [9].

- **Average BLEU**: Average of BLEU- $\{1, 2, 3, 4\}$ as in [7].

- **Bag of Words Embedding Similarity**: Three different types of cosine similarities (**Extrema**, **Average**, **Greedy**) of Bag-of-Words Embedding between generation and target as in [28].

Human Evaluation We modify the human evaluation method from [7] as 5-point scale does not directly compare between models. Instead of 5-point ratings, we conduct Multiple Choice Testing against Baselines and task 5 human judges with 100 dialogue samples. We ask them to choose among 6 different models the best responses for each criterion (**Fluency**, **Relevance**, and **Empathy**). We also gave them options to choose 'Tie - All' and 'Tie - None' which indicate all answers were good or all answers were bad.

Bag of Words Embedding Similarity

	Extrema	Average	Greedy
Seq2Seq	0.908 ± 0.0015	0.961 ± 0.0005	0.949 ± 0.0007
MultiSeq	0.904 ± 0.0013	0.962 ± 0.0011	0.948 ± 0.0003
Current	0.903 ± 0.0022	0.961 ± 0.0009	0.948 ± 0.0004
Forward	0.901 ± 0.0011	0.961 ± 0.0006	0.947 ± 0.0014
Improvement	$\mathbf{0.919} \pm 0.0008$	$\mathbf{0.966} \pm 0.0003$	$\mathbf{0.952} \pm 0.0006$
Simulation	$\mathbf{0.919} \pm 0.0020$	$\mathbf{0.966} \pm 0.0011$	0.949 ± 0.0003

Table 2. Automatic metrics for measuring relevance of generated responses using Bag-of-Words Embedding Cosine Similarity with human reference. *Improvement* and *Simulation* models show highest scores in all three types.

4. RESULTS & DISCUSSION

4.1. Discussion of Automatic Evaluation Results

Because of the stochastic nature of using top-k sampling [27] as a decoding strategy, we run each evaluation 3 times and report the mean and standard deviation.

From Table 1, it is clearly visible that *Improvement* model significantly outperforms other proposed models and baselines in terms of Distinct N-grams. This indicates that this model can generate more diverse sentences than the others. *Improvement* model having the second highest Sentence Length and Average BLEU also points out that our *Improvement* model generates more fluent utterances.

We evaluate how relevant the generated sentences are using three different types of Bag-of-Words Embedding Similarity as in [28], namely Extrema, Average, and Greedy. For all metrics, we use FastText embeddings and measure the cosine similarity between generated utterances and human responses in order to check how semantically close our responses are to the references. For all three metrics, Table 2 shows us that *Improvement* and *Simulation* achieves higher scores than other baselines, which indicates that these responses could be

Multiple Choice Testing against Baselines

	Empathy	Relevance	Fluency
Seq2Seq	0.11	0.14	0.05
MultiSeq	0.12	0.09	0.03
Current	0.11	0.12	0.03
Forward	0.09	0.1	0.07
Improvement	**0.24**	**0.26**	**0.12**
Simulation	0.07	0.04	0.01
Tie - All	0.08	0.06	0.68
Tie - None	0.18	0.19	0.01
Total	1.0	1.0	1.0
Kappa	0.27	0.2	0.08

A/B Testing against Human

	Empathy	Relevance	Fluency
Improvement	0.14	0.05	0.05
Human	**0.4**	0.42	0.36
Tie - All	0.37	**0.44**	**0.52**
Tie - None	0.09	0.09	0.07
Total	1.0	1.0	1.0
Kappa	0.24	0.31	0.07

Table 3. *Improvement* model clearly outperforms all other models in all three tested categories, where Kappa is annotator agreement score. For A/B testing, *Human* outperforms *Improvement* mode, but the ratios of *Tie - All* are also significantly high.

considered more semantically closer to the human responses.

4.2. Discussion of Human Evaluation Results

We conduct two types of multiple-choice human surveys: one against other models and another testing the best model against human responses.

Multiple-choice Testing Note that for Table 3, *Kappa* indicates Fleiss-Kappa multi-annotator agreement score [29]. Heuristically, a score between 0.2 and 0.4 is within range of "fair agreement." We chose to do a multiple choice test so that we could conduct a more direct comparison with other models and find the best, whereas 5-point scale as in [7] would result in an independent set of results that are not directly comparable. From Table 3, we could easily see that *Improvement* model results the highest in all three categories, which is consistent with the automatic results. For Fluency, it is clear that the vast majority of the human judges selected *Tie All*. This is because most answers were often grammatically accurate.

Humans vs Improvement We then conducted an A/B test between the best performing model (*Improvement*) against human responses in order to see to what extent we can compete

Dialogue History
SPEAKER: Not all of my high-school memories are fond ones. Like the time I drank too much and passed out at a party.

Seq2Seq
LISTENER: I'm sorry to hear that. I'm sure you'll find a nice place to be a party.

MultiSeq
LISTENER: I know how you feel. I don't have kids, but I don't blame you.

Current
LISTENER: That is awesome. I bet you were so proud.

Forward
LISTENER: I love surprises, especially when you're a kid. I know it has to feel really bad about it.

Improvement
LISTENER: That's too bad. What was your favorite part?

Simulation
LISTENER: I'm so sorry to hear that. Did you have fun?

Table 4. Generated dialogue responses from different models. Our method (*Improvement*) produces the most empathetic and relevant response.

with humans. Unsurprisingly, as shown in Table 6, human responses are chosen approximately 40% of the time for all three categories. However, it is notable that *Tie - All* was chosen similarly as well, which means that the judges were not sure which one was better as both were good responses.

4.3. Generation Examples

We then look at some generation examples to verify the insights from previous discussions and discoveries made in the above correlation analysis. From the first dialogue in Table 4, we can see that all responses show positive sentiments and are fluent in terms of grammar, but *Improvement* exhibits the longest and most relevant response by copying the word "beach" from the input. Furthermore, for the second dialogue example, *Seq2Seq* is the only model to directly address "party", but it is not really relevant or fluent. On the other hand, *Improvement* not only properly addresses the SPEAKER's sentiment, but also asks what were the good memories instead of further discussing the bad ones.

5. CONCLUSION

In this paper, we propose a novel perspective on empathy, *sentiment look-ahead*, as the key to generating empathetic responses. We implement and train three different sentiment look-ahead reward functions to model how the user would feel

towards a generated dialogue response, and use such to encourage more empathetic responses. The empirical results on both automatic and human evaluations in terms of Empathy, Fluency, and Relevance confirm that our proposed approach is an effective way to generate more empathetic responses compared to other models that condition on current user emotions.

6. REFERENCES

[1] Bilyana Martinovski and David Traum, "Breakdown in human-machine interaction: the error is the clue," in *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, 2003, pp. 11–16.

[2] Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka, "Recognizing, modeling, and responding to users' affective states," in *International Conference on User Modeling*. Springer, 2005, pp. 60–69.

[3] Emmanuelle Zech and Bernard Rimé, "Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits," *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, vol. 12, no. 4, pp. 270–287, 2005.

[4] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *arXiv preprint arXiv:1812.08989*, 2018.

[5] Thomas S Polzin and Alexander Waibel, "Emotion-sensitive human-computer interfaces," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[6] Marcin Skowron, "Affect listeners: Acquisition of affective states by means of conversational systems," in *Development of Multimodal Interfaces: Active Listening and Synchrony*, pp. 169–181. Springer, 2010.

[7] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019.

[8] Oriol Vinyals and Quoc Le, "A neural conversational model," in *International Conference on Machine Learning*, 2015.

[9] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "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*, 2016, pp. 110–119.

[10] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao, "Deep reinforcement learning for dialogue generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.

[11] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing, "Toward controlled generation of text," in *International Conference on Machine Learning*, 2017, pp. 1587–1596.

[12] Ke Wang and Xiaojun Wan, "Sentigan: generating sentimental texts via mixture adversarial networks," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 4446–4452.

[13] Xianda Zhou and William Yang Wang, "Mojitalk: Generating emotional responses at scale," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1128–1137.

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

[15] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura, "Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[16] 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.

[17] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[18] Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[19] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," *International Conference on Learning Representations*, 2016.

[20] Ronald J Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[21] Romain Paulus, Caiming Xiong, and Richard Socher, "A deep reinforced model for abstractive summarization," *International Conference on Learning Representations*, 2018.

[22] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017.

[23] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston, "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)*, 2018, pp. 2204–2213.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "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)*, 2019, pp. 4171–4186.

[25] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1631–1642, Association for Computational Linguistics.

[26] Stephen Merity, Nitish Shirish Keskar, and Richard Socher, "Regularizing and optimizing lstm language models," *International Conference on Learning Representations*, 2018.

[27] Angela Fan, Mike Lewis, and Yann Dauphin, "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 889–898.

[28] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau, "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*, 2016, pp. 2122–2132.

[29] Joseph L Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, pp. 378, 1971.