IMPROVING CONTEXTUAL COHERENCE IN VARIATIONAL PERSONALIZED AND EMPATHETIC DIALOGUE AGENTS

Jing Yang Lee[†]   *Kong Aik Lee*[‡]   *Woon Seng Gan*[†]

[†] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[‡] Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

In recent years, latent variable models, such as the Conditional Variational Auto Encoder (CVAE), have been applied to both personalized and empathetic dialogue generation. Prior work have largely focused on generating diverse dialogue responses that exhibit persona consistency and empathy. However, when it comes to the contextual coherence of the generated responses, there is still room for improvement. Hence, to improve the contextual coherence, we propose a novel Uncertainty Aware CVAE (UA-CVAE) framework. The UA-CVAE framework involves approximating and incorporating the aleatoric uncertainty during response generation. We apply our framework to both personalized and empathetic dialogue generation. Empirical results show that our framework significantly improves the contextual coherence of the generated response. Additionally, we introduce a novel automatic metric for measuring contextual coherence, which was found to correlate positively with human judgement.

Index Terms— Dialogue Generation, Latent Variables, Uncertainty Estimation, Artificial Intelligence

1. INTRODUCTION

The ultimate goal of open-domain dialogue agents is to engage humans in seamless, natural conversation. To achieve this goal, dialogue agents are expected to generate responses which demonstrate empathy and persona consistency. Hence, the tasks of personalized [1, 2, 3] and empathetic (or emotional) dialogue generation [4, 5, 6, 7] has received significant research attention in recent years. Traditionally, neural personalized or empathetic open-domain dialogue agents demonstrate a tendency to generate generic and repetitive responses. To address the lack of diversity in the generated responses, numerous approaches have explored the application of latent variable models, particularly the Conditional Variational Auto Encoder (CVAE), to personalized and empathetic dialogue generation. Empirical results from prior works [8, 9, 10, 11, 12, 6] have shown that CVAE-based models have largely succeeded in generating diverse dialogue responses. Despite the gains in response diversity, there is still

much room for improvement, in particular, when it comes to contextual coherence.

To improve the contextual coherence of the responses generated by CVAE-based agents, we propose approximating and incorporating uncertainty during the generation process. There are two types of uncertainty: aleatoric and epistemic uncertainty [13]. While aleatoric uncertainty captures the uncertainty in the input data, epistemic uncertainty quantifies the uncertainty due to the model. Typically, aleatoric uncertainty is predicted as a model output, and the epistemic uncertainty is obtained via ensemble methods [14] or monte carlo simulations [15]. Aleatoric uncertainty can be further categorized as either homoscedastic or heteroscedastic. Homoscedastic, or task-dependant, uncertainty refers to input invariant noise which differs from task to task. On the other hand, heteroscedastic, or data-dependant, uncertainty quantifies the uncertainty (or noise) inherent in each input and is thus unique to each input.

Hence, we introduce an Uncertainty-Aware CVAE (UA-CVAE) framework, which involves approximating and incorporating the heteroscedastic aleatoric uncertainty via an uncertainty aware latent variable. We also propose an Utterance Entailment (UE) score that quantifies the contextual coherence of an agent's responses. Experimental results show that our framework significantly improves overall response coherence. The remainder of this paper is organized as follows: An overview of aleatoric uncertainty approximation in CVAE-based dialogue agents is provided in Section 2. In Section 3, the UA-CVAE framework and UE score are described in detail. Experimental results and the conclusion are provided in sections 4 and 5 respectively.

2. UNCERTAINTY IN DIALOGUE GENERATION

For our discussion, let $X = \{x_{0,0}, ..., x_{M-1,N-1}\}$ refers to the dialogue context (M refers to the number of utterances in the context, and N denotes the number tokens in a given utterance), and $Y = \{y_0, ..., y_{L-1}\}$ refers to the generated dialogue response (L represents the number of tokens). Similarly, $\bar{Y}$ will refer to the reference dialogue response. Additionally, we use z and c to denote the sampled latent variable
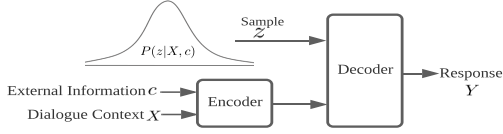
Fig. 1. Overview of a CVAE-based dialogue agent.

and any external information (persona description or emotion label) respectively.

2.1. CVAE-based Dialogue Generation

In a CVAE-based dialogue agent, the generated response Y is conditioned on the the dialogue context X, the external information c (persona description or emotion label), as well as latent variable z, which is randomly sampled from latent Gaussian distribution $p(z|X,c)$. An overview is provided in Fig 1. The stochasticity introduced by z improves the diversity of the generated dialogue.

2.2. Aleatoric Uncertainty Approximation

In this section, we show that for a CVAE-based dialogue agent, the variance σ^2 of the latent Gaussian distribution is an approximation of the *aleatoric uncertainty* of the generated dialogue response. For a CVAE-based dialogue agent, the latent variable z is assumed to follow an isotropic multivariate Gaussian distribution:

$$z \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I}) \tag{1}$$

where μ and σ^2 refer to the mean and variance vectors of the latent distribution. The CVAE-based dialogue generation process can be represented by the following conditional distribution:

$$p(Y, z|X, c) = p(Y|z, X, c)p(z|X, c) \tag{2}$$

In a CVAE-based dialogue agent, both $p(Y|z,X,c)$ and $p(z|X,c)$ will be defined by neural networks parameterized by π and θ respectively. In particular, as z is a Gaussian random variable and $p(z|X,c) = \mathcal{N}(\mu, \sigma^2\mathbf{I})$, the network $p_\theta(z|X,c)$ will be tasked with generating a mean μ and a variance σ^2. Hence, since the variance σ^2 predicted by the neural network $p_\theta(z|X,c)$ is dependant on the input data, σ^2 would capture the uncertainty inherent in the input X and c i.e., the heteroscedastic aleatoric or data-dependant uncertainty. This is similar to the uncertainty estimation approach introduced in [15], where a data dependant function of the uncertainty is learned. From (2), we observe that the response generation process $p(Y, z|X, c)$ and $p(z|X, c)$ are both conditioned only on X and c. Hence, we conclude that the aleatoric uncertainty of the generated dialogue $p(Y, z|X, c)$ can be approximated by the variance σ^2, which captures the aleatoric uncertainty of $p(z|X, c)$.
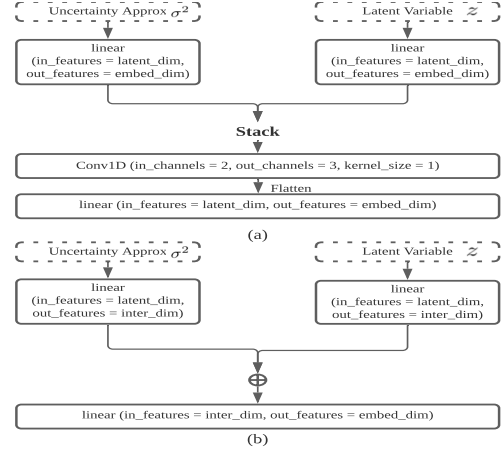

Fig. 2. (a) Overview of the combination network in UA-CVAE(C). (b) Overview of the combination network in UA-CVAE(M). $\oplus$ indicates element-wise addition. Both networks output z_u. The embedding dimension (embed_dim), intermediate dimension (inter_dim) and the latent dimension (latent_dim) are fixed at 768, 384 and 256 respectively.

3. UNCERTAINTY-AWARE CVAE

As mentioned, the conditional distributions $p(Y|z, X, c)$ and $p(z|X, c)$ are defined by neural networks parametrized by π and θ respectively. A GPT-2 pretrained model [16] is used to obtain the sequence embeddings of X, c and $\bar{Y}$, denoted by X', c' and $\bar{Y}'$ respectively. The prior network $p_\theta(z|X,c)$ is composed of a single-layer MLP tasked with generating the mean μ_p and the log-variance $log(\sigma_\mathrm{p}^2)$ of the latent Gaussian distribution:

$$\begin{bmatrix} \mu_\mathrm{p} \\ log(\sigma_\mathrm{p}^2) \end{bmatrix} = W_\mathrm{p} \begin{bmatrix} c' \\ X' \end{bmatrix} + b_\mathrm{p} \tag{3}$$

where W_p and b_p refer to the weights and bias of the prior network. Before being fed into the prior network, the sequence embedding of the dialogue context and the external information are obtained via a GPT-2 pretrained model. The variance, is attained by computing the exponential of the log-variance $log(\sigma_\mathrm{p}^2)$. To obtain z, the latent Gaussian distribution is randomly sampled via the reparametrization trick to ensure differentiability. During training, the true posterior distribution $q(z|X, c, \bar{Y})$ is approximated by another MLP parameterized by ϕ. This network, termed the recognition network $q_\phi(z|X, c, \bar{Y})$, also consists of a single layer MLP:

$$\begin{bmatrix} \mu_\mathrm{r} \\ log(\sigma_\mathrm{r}^2) \end{bmatrix} = W_\mathrm{r} \begin{bmatrix} c' \\ X' \\ \bar{Y}' \end{bmatrix} + b_\mathrm{r} \tag{4}$$

where W_r and b_r refer to the weights and bias of the recognition network. Similarly, Before being fed into the prior network, the sequence embedding of the dialogue context and
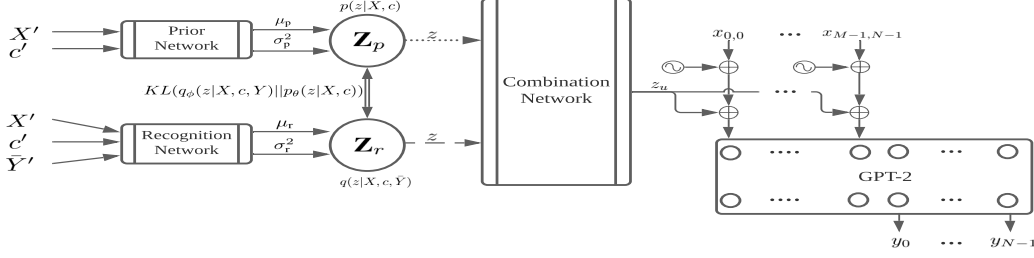
Fig. 3. Overview of the UA-CVAE framework. The dashed and dotted connection occurs in only the training and inference stages respectively.

the external information are obtained via a GPT-2 pretrained model [16]. The KL divergence between the Gaussian distributions generated by the prior and recognition network will be incorporated in the final loss function. Then, the latent variable z and the variance vector σ_p^2 (inference) or σ_r^2 (training), an approximation of the *aleatoric uncertainty* of the generated dialogue response, is fed to the combination network which outputs the uncertainty aware latent variable z_u. By conditioning the decoder on the uncertainty aware latent variable z_u, the decoder would possess an awareness of the uncertainty or noise inherent in X and c. This allows the decoder to recognize and adapt to uncertain or noisy dialogue contexts and persona description/emotion labels, reducing the number of contextually incoherent responses. We implement 2 variants of the combination network (Fig 2), resulting in 2 variants of the UA-CVAE framework: UA-CVAE(M) and UA-CVAE(C). While the combination network in UA-CVAE(M) constitutes a several linear layers, the combination network in UA-CVAE(C) involves a 1D convolution layer. Then, z_u is added to the positional encoding and token-level embedding at each decoding step, and fed to the response decoder $p_\pi(Y|z, X, c)$ which constitutes a GPT-2 pretrained language model [16].

Since training a CVAE would require an intractable marginalization over z, Stochastic Gradient Variational Bayes (SGVB) is used to train the UA-CVAE. Additionally, due to the vanishing latent variable problem, we include the Bag-of-Words (BoW) loss in the loss function [12]. The final loss function is as follows:

$$
\begin{aligned}
L(X,Y,c,z,z_u) = &-KL(q_\phi(z|X,c,Y)||p_\theta(z|X,c)) \\
&+ E_{q_\phi(z|X,c,Y)}[log\,p(Y|X,c,z_u)] \\
&+ E_{q_\phi(z|X,c,Y)}[log\,p(Y_{bow}|X,c,z_u)]
\end{aligned} \tag{5}
$$

where Y_{bow} refers to the generated dialogue response BoW. A overview is provided in Fig 3.

3.1. Utterance Entailment Score

Prior work have shown that the task of Natural Language Inference, which involves evaluating if a given premise entails or contradicts a predefined hypothesis, correlates with human judgement with regard to contextual coherence. Our metric, termed the Utterance Entailment (UE) score, requires finetuning a BERT model on the SNLI corpus and applying the resultant model to the dialogue context and generated response. Specifically, given a single utterance in the dialogue context x_i and a generated response Y, the model judges if the premise x_i and the hypothesis Y are contradictory, entailing or neutral. Then, a rating of 1,-1 or 0 is assigned when the premise and hypothesis are entailing, contradictory or neutral respectively. For each response, the final Utterance Entailment score is computed by summing all assigned ratings. Hence, for each dialogue example, the UE score can be computed as follows:

$$
UE(X,Y) = \sum_{i=1}^{M} BERT_{NLI}(x_i, Y) \tag{6}
$$

where X and Y refer to the dialogue context and generated response respectively. The final UE-score for a given test set is obtained by averaging the UE-scores of each dialogue example.

4. EXPERIMENTS & RESULTS

4.1. Corpora

We benchmark our framework on the ConvAI2 [17] and EMPATHETICDIALOGS [18] corpora. For EMPATHETICDIALOGS, the human speaker describes a personal situation and the dialogue agent is expected to generate an empathetic response. Each dialogue example is provided an emotion label (e.g., 'excited'). ConvAI2 consists of dialogues where two interlocutors are trying to get to know each other. Each interlocutor is assigned a persona description consisting of multiple persona statements (e.g., 'I love dogs', 'I like eggs').

4.2. Implementation & Baselines

In our implementation of UA-CVAE(M) and UA-CVAE(C), the response decoder consists of a GPT-2 pretrained language model [16] which is finetuned during training. We utilize the GPT-2 base model from HuggingFace consisting of 12 layers,

Table 1. Results of automatic evaluation.

	Model	PPL	Avg Length	Rouge-l (F1)	Rouge-l (R)	Rouge-l (P)	METEOR	Distinct-1	Distinct-2	Distinct-3	UE-score
ConvAI2	seq2seq	42.5615	10.3481	0.0115	0.0672	0.0064	0.3109	0.3927	0.6870	0.7509	0.0002
	HRED	33.0022	10.3594	0.0126	0.0732	0.0069	0.3023	0.4134	0.7531	0.8204	0.0006
	VHRED	39.6422	10.5012	0.0110	0.0636	0.0061	**0.3154**	0.4339	0.8142	0.8817	0.0008
	VHCR	32.1812	10.8626	0.0114	0.0660	0.0063	0.3099	0.4364	0.7641	0.8183	0.0004
	GPT-2	15.5766	9.5820	0.1434	**0.1846**	0.1236	0.2294	0.4496	0.8160	0.8705	0.0540
	CVAE	14.2457	11.0069	0.1172	0.1406	0.1073	0.2631	0.4773	**0.8729**	**0.9245**	0.0406
	UA-CVAE(M)	14.5423	**11.7421**	**0.1291**	0.1517	0.1200	0.2576	0.4562	0.8554	0.9149	0.0648
	UA-CVAE(C)	**14.1253**	11.3658	0.1169	0.1475	**0.1245**	0.2507	**0.4981**	0.8648	0.9106	**0.0906**
Empathetic Dialogues	seq2seq	52.9023	10.9731	0.0054	0.0355	0.0030	0.3243	0.4468	0.7561	0.8130	0.0103
	HRED	68.2843	11.0345	0.0053	0.0353	0.0029	**0.3311**	0.4444	0.8333	0.8988	0.0095
	VHRED	55.4275	11.0139	0.0045	0.0297	0.0025	0.3305	0.4396	0.8149	0.8798	0.0089
	VHCR	63.4752	11.2758	0.0034	0.0416	0.0034	0.3017	0.4884	0.7668	0.8112	0.0090
	GPT-2	16.6456	7.3657	0.0390	0.0676	0.0299	0.2525	0.4848	0.7902	0.8315	0.1081
	CVAE	16.3177	8.0514	0.0745	0.1087	0.0651	0.2333	0.5139	0.8715	0.9328	0.1074
	UA-CVAE(M)	**15.2021**	9.8180	**0.0867**	**0.1132**	0.0817	0.2913	**0.5526**	**0.9149**	**0.9370**	0.1198
	UA-CVAE(C)	17.1932	**12.1111**	0.0859	0.1045	**0.0825**	0.2794	0.4922	0.8877	0.9239	**0.1257**

Table 2. Results of human evaluation. Kappa scores generally range from 0.5 to 0.6, indicating moderate agreement.

	Model	UA-CVAE(M)				UA-CVAE(C)			
		Win	Tie	Loss	Kappa	Win	Tie	Loss	Kappa
ConvAI2	seq2seq	53%	45%	2%	0.54	55%	39%	6%	0.59
	HRED	49%	42%	9%	0.49	58%	38%	4%	0.46
	VHRED	43%	41%	16%	0.52	43%	51%	6%	0.50
	VHCR	42%	38%	20%	0.64	51%	41%	8%	0.59
	GPT-2	37%	58%	5%	0.57	41%	45%	14%	0.51
	CVAE	34%	49%	17%	0.43	44%	45%	11%	0.47
	UA-CVAE(M)	-	-	-	-	29%	48%	23%	0.63
	UA-CVAE(C)	23%	48%	29%	0.63	-	-	-	-
Empathetic Dialogues	seq2seq	50%	42%	8%	0.42	58%	32%	10%	0.66
	HRED	53%	43%	4%	0.49	64%	29%	7%	0.58
	VHRED	57%	38%	5%	0.53	57%	38%	5%	0.55
	VHCR	59%	33%	8%	0.54	61%	37%	2%	0.60
	GPT-2	33%	52%	15%	0.64	52%	29%	19%	0.49
	CVAE	39%	47%	14%	0.56	34%	53%	13%	0.54
	UA-CVAE(M)	-	-	-	-	31%	56%	13%	0.64
	UA-CVAE(C)	13%	56%	31%	0.64	-	-	-	-

12 attention heads as well as a hidden dimension of 768. The size of the latent variable z and z_u was fixed at 256. The Adam optimizer (learning rate = 0.0001, batch size = 16) was used. The maximum utterance length and maximum number of turns were fixed at 40 and 4 respectively.

In addition to **UA-CVAE(C)** and **UA-CVAE(M)**, we implement the following baselines: **Seq2seq** with attention applied to the GRU decoder, Hierarchical Recurrent Encoder-Decoder (**HRED**) [19], Variational VHRED (**VHRED**) [20], Variational Hierarchical Conversation RNN (**VHCR**) [21], **GPT-2**, and the **CVAE** with a GPT-2 decoder.

5. RESULTS & DISCUSSION

5.1. Results

We evaluate the generated dialogue with the following automatic metrics (Table 1): PPL, ROUGE [22], METEOR [23], intra-response Distinct-n (1,2,3) [24], and the UE-Score. The average response length is also provided. For human evaluation (Table 2), we engage 4 graduates to evaluate the contextual coherence of the generated responses. Specifically, each graduate was told to compare a response generated by a UA-CVAE variant with a response generated by the baselines, and select the more contextually coherent response.

5.2. Discussion

Based on the human evaluation results and the automatic metrics, we can conclude that there is a positive correlation between the proposed UE-score and human judgement. UA-CVAE(M)/(C) generally achieve a lower percentage of wins and higher percentage of losses against baselines with higher UE-scores such as GPT-2 and CVAE.

It is also apparent that UA-CVAE(M) and UA-CVAE(C) outperformed all implemented baselines in terms of contextual coherence. Furthermore, the Distinct-n, Rouge and METEOR scores achieved by UA-CVAE(M) and UA-CVAE(C) were comparable to that of the CVAE baseline, suggesting comparable response diversity and fluency. In terms of UE-score, compared to the CVAE baseline, UA-CVAE(M) and UA-CVAE(C) achieved a 60% and 121% improvement respectively on ConvAI2. However, on EMPATHETICDI-ALOGS, UA-CVAE(M) and UA-CVAE(C) achieved a 11% and 17% gain. The difference in gains could be attributed to the informative nature and shorter length of the dialogue utterances in ConvAI2, which would result in more positive NLI predictions ('entailment') when computing the UE-score. Also, we notice that UA-CVAE(C) has a slight edge over UA-CVAE(M) in terms of coherence. In addition to attaining higher UE-scores, the UA-CVAE(C) model achieved a larger percentage of wins on both corpora. This shows that a 1D convolution layer allows for a slightly more effective combination of the variance, or uncertainty, vector and latent variable.

6. CONCLUSION

We have proposed two UA-CVAE frameworks: UA-CVAE(M) and UA-CVAE(C), which improved coherence on both ConvAI2 (60% and 121%) and EMPATHETICDIALOGS (11% and 17%). Additionally, we introduced the UE-score, which correlated positively with human judgement in terms of contextual coherence. Future work could involve exploring the possibility of enhancing the dialogue quality by leveraging the epistemic or homoscedastic aleatoric uncertainty.

7. REFERENCES

[1] Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan, "Generating personalized dialogue via multi-task meta-learning," in *SEMDIAL*, Sept. 2021, pp. 88–97.

[2] Feng-Guang Su, Aliyah R. Hsu, Yi-Lin Tuan, and Hung-Yi Lee, "Personalized Dialogue Response Generation Learned from Monologues," in *INTERSPEECH*, 2019, pp. 4160–4164.

[3] Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang, "You impress me: Dialogue generation via mutual persona perception," in *ACL*, July 2020, pp. 1417–1427.

[4] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang, "Generating responses with a specific emotion in dialog," in *ACL*, July 2019, pp. 3685–3695.

[5] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen, "EmpDG: Multi-resolution interactive empathetic dialogue generation," in *COLING*, Dec. 2020, pp. 4454–4466.

[6] Xianda Zhou and William Yang Wang, "MojiTalk: Generating emotional responses at scale," in *ACL*, July 2018, pp. 1128–1137.

[7] Chenyang Huang, Osmar Zaïane, Amine Trabelsi, and Nouha Dziri, "Automatic dialogue generation with expressed emotions," in *NAACL-HLT*, June 2018, pp. 49–54.

[8] Jing Yang Lee, Kong Aik Lee, and Woon-Seng Gan, "Dlvgen: A dual latent variable approach to personalized dialogue generation," *CoRR*, vol. abs/2111.11363, 2021.

[9] Yu-Ping Ruan and Zhenhua Ling, "Emotion-regularized conditional variational autoencoder for emotional response generation," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[10] Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu, "Exploiting persona information for diverse generation of conversational responses," in *IJCAI*, 2019.

[11] Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang, "Guiding variational response generator to exploit persona," in *ACL*, July 2020, pp. 53–65.

[12] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *ACL*, July 2017, pp. 654–664.

[13] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *NeurIPS*, 2017, vol. 30.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, 2017, vol. 30.

[15] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016, pp. 1050–1059.

[16] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[17] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston, "The second conversational intelligence challenge (convai2)," 2019.

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

[19] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," 2016.

[20] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," *AAAI*, vol. 31, no. 1, 2017.

[21] Yookoon Park, Jaemin Cho, and Gunhee Kim, "A hierarchical latent structure for variational conversation modeling," in *NAACL-HLT*, June 2018, pp. 1792–1801.

[22] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, July 2004, pp. 74–81.

[23] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, June 2005, pp. 65–72.

[24] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," in *NAACL-HLT*, June 2016, pp. 110–119.