Date of publication MAR-29, 2021, date of current version MAY-25, 2020.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.20.1.2090

Improving Conversation Modelling using Attention Based Variational Hierarchical RNN

SANDEEP A. THORAT, KOMAL P. JADHAV
RIT Sakharale, Islampur, India
(e-mail: sandip.thorat@ritindia.edu, komalpjadhav22@gmail.com)

Corresponding author: Sandeep A. Thorat (e-mail: sandip.thorat@ritindia.edu).

ABSTRACT Conversation modeling is one of most important applications of natural language processing. Building response generation model for open domain conversation in a Chatbot is one of the hardest challenges in this area. The deep neural network architectures such as sequence to sequence models and its hierarchical variants provide a significant improvement in the field of conversation modeling. Although these models require large size corpus, they may cause huge data loss in training phase. Also, these models are unable to concentrate on important data in given context. It affects on generation of responses. To tackle these issues, this research work proposes a Variational Hierarchical Conversation RNN with Attention mechanism (VHCRA) model for response generation. The VHCRA uses the concept of latent variable representation to avoid data degeneracy and the attention mechanism to identify important data within context. The model is trained on large size benchmark dataset, i.e., Cornell Movie Dialog corpus which contains conversations from different movies. The model is evaluated using automatic evaluation metrics such as Negative Log-likelihood and Embedding-Based Metrics. The experimental result shows that the proposed model gains significant improvement in comparison with recently proposed approaches and generate meaningful responses according to the context.

KEYWORDS chatbot, conversation agent, response generation, variational hierarchical RNN, deep learning, natural language processing, attention mechanism.

I. INTRODUCTION

RECENT years have seen rapid growth of sequential conversational data on Internet which have become an essential part of natural language processing. The motivation behind designing a conversational agent is to conduct smooth and natural conversation between computer and a human. To keep the human-computer interaction going, responses must be relevant, persistent, properly understood and reacted by user. The job of conversation model is to be able to determine the best responses for any given message that it receives.

The field of conversational modeling has witnessed signicant improvements after rise of deep learning methods. Recent approaches such as Recurrent Neural Network(RNN) and its variants, Sequence-to-sequence model gives more promising results for response generation models [1, 2]. There are many RNN encoder decoder architectures such as Hierarchical RNN [3] and its different forms are useful for conversational modeling. These models have hierarchical structure, sequence of utterance and each utterance of tokens [4].

Although all existing models have made a signi- cant progress in response generation, they face data degeneration problem [5]. Due to this problem, large amount of data get lost while training the model. The Variational Auto-Encoder, i.e., VAE [6] which uses hierarchical RNN also provides a powerful architecture for language modeling. The VAE encodes data to latent variables and then decodes latent variable to reconstruct the input data [7]. Even if the model is dependent on latent variable, it suffers from data degeneration problem.The reason behind it is autoregressive power of decoder RNN [5]. Because of that the quality of generated responses becomes poor.

Another major drawback of neural response generation model is that they produce more generic responses such as “i don’t know”, which makes a conversation more unrealistic. Although all the data of contextual data is not equally important for response generation, there is need to find out important words within the context to generate appropriate output. The different state-of-the art models such as HRED and VHRED focus on hierarchical structure of context whereas they ignore the important data within the context which leads to generation of irrelevant responses [8]. The concept of attention mechanism [9] for training neural network allows a model to concentrate on important data in context.

The objective of this work is to propose a response generation model for open domain conversation in Chatbot which overcomes the data degeneration problem and produces high quality responses for given context. In this work we propose a Variational Hierarchical Con-versational RNN with Attention mechanism (VHCRA) which is the combination of variational hierarchical model and attention mechanism to solve the data degeneration problem as well as produce relevant response generation. The proposed model uses the concept of VHCR [5] where it uses a global conversation latent variable along with local utterance latent variable to produce a hier-archical structure. At the beginning, word level Gated Recurrent Unit(GRU) encodes each sentence into hidden encoder vector. This vector is then fed as input to context RNN level where it is updated along with global latent variable and produces hidden representation of context. Then attention mechanism [10] is applied to encoder vector calculating the attention weight to generate attention vector that attends important words in the sentence. Finally, the context vector along with attention vector is taken as input by decoder and it generates a response for input context.

The proposed VHCRA algorithm is evaluated using Cornell movie dialog corpus which has large scale open domain conversation data. The algorithm is compared with existing models using automatic evaluation methods such as embedding based matrices and calculating Negative log likelihood. Experimental result shows that the proposed VHCRA model outperforms the various state-of-the art response generation model when evaluating automatically. It prevents model from data degeneration problem by giving stable KL-divergence. It successfully generates longlength meaningful sentence for ground-truth.

Following are the key research contributions of this work:

1. The research work proposes fully data driven response generation model.
2. The proposed model alleviates the data degeneration problem and produces more relevant responses.
3. The research work does experimental comparison of the proposed model with other state-of-the art models to verify the effectiveness of the proposed model.

This paper is structured as follows: Section II describes history of conversational agent and the literature survey related to proposed work. Section III discusses theoretical concept of Variational Auto-encoder, Variational Hierarchical RNN and Attention mechanism. The same section describes working of proposed model (VHCRA) with help of mathematical modeling. Section IV describes the experimental environment and results using automatic evaluation metrics. The paper ends by giving conclusion in the Section V.

II. LITERATURE SURVEY

The development of conversation agent started in 1960 whereas Eliza was first designed Chatbot by Joseph Weizenbhaum [11]. This started the journey of development of smart and effective Chatbot systems.

Conversational models are generally categorized into two types as retrieval-based and generative-based models. The Retrieval based model generate a response by selecting from predefined set of responses. ALICE [12] is popular example of retrieval based Chatbot system. Generative-based models are based on machine translation techniques, where there is no predefined set of responses. It generates the response based on previous information [13]. Various machine learning and deep learning algorithms show powerful improvement in designing a generative based conversation system.

In recent times, a sequence to sequence model which uses multilayer Long-Short-Term Memory network (LSTM) gained more popularity for conversational modeling [1]. The Sordoni *et al.* [4] introduced a neural network architecture to generate context-sensitive responses by using word embedding method. This architecture is completely data driven and has become a ground work for further research. Later different variants of neural network architectures were developed such as Hierarchical structure of RNN Encoder Decoder (HRED) [3] and then HRED with latent variable representation (VHRED) [2]. These models gave more promising result for response generation. Xiaoyu Shen *et al.* [14] presented a conditional variational framework inspired by semi supervised generative model, i.e., VAE model [6] for specific response creation, where the model outperformed HRED. To overcome the data degeneration problem related to hierarchical RNN structure Yookon Park *et al.* [5] proposed a novel Variational Heirarchical Conversation RNN model (VHCR) which successfully utilized latent variables for conversation generation. All these studies take the encoder-decoder framework to address response generation issues.

Later the attention mechanism which was firstly introduced for machine translation [9] was used for response generation [15] to incorporate topic related information into sequence-to-sequence model to produce informative and interesting responses.The Hai-Tao Zheng *et al.* [16] proposed a Gated attention neural network model to generate responses automatically, where for better utilization of contextual information an attention mechanism was used in encoder decoder framework. The Shang et al. [17] added a feedback attention mechanism in encoder decoder framework to generate a response to post. It uses attention mechanism in encoder decoder framework which differs from Serban *et al.* [2] model which adopts hierarchical structure for the same framework. Futher, Chen Xing *et al.* [8] extended the use of attention mechanism to hierarchical attention for multi-turn response generation. The hierarchical attention and knowledge matching network was successfully applied by Junqing He *et al.* [18] to deal with vital information deficiency problem and to produce logically correct responses.The Variational Encoder-Decoder(VED) model is combined with attention vector that develop variational attention mechanism to discover bypassing phenomena in VED [7].

There are several studies presented in literature to enhance the conversation modelling approach. Many of the studies address the reason behind producing irrelevant responses is the degeneration of the information. This research work extends variational hierarchical encoder decoder model by using attention mechanism. The model is hierarchical in structure and allows to concentrate on important data within context by using attention mechanism. This helps model to produce relevant responses and to improve the measurement of generalization of data. The work is inspired by given literature and follows the idea of [5] to solve the data degeneration problem and to improve the accuracy of generated responses.

III. PROPOSED APPROACH FOR IMPLEMENTING CONVERSATION MODEL

In this section, the subsection A introduces various technical backgrounds to understand the proposed model. Later subsection B discusses architecture and mathematical modeling of proposed conversational model.

A. TECHNICAL BACKGROUND

Before introducing proposed model, let us first briefly review the Variational Auto-Encoder, Variational Hierarchical RNN and Attention Mechanism.

VARIATIONAL AUTO-ENCODER

Variational Auto-Encoder (VAE) is a type of generative model that uses deep learning neural network architecture to predict variational distribution parameters [6]. It is regularized version of standard auto-encoder which is able to learn meaningful representation from the data [19]. The VAE is encoder-decoder architecture which uses idea of re-parameterization of data.

In VAE, reconstruction and KL-Divergence are two important terms that measure the effectiveness of decoder to reconstruct the data and amount of information encoded in latent variable respectively. The model uses latent variable representation then also suffers from data degeneration. The reason behind it is that, the VAE model incorporates autoregressive power of decoder RNN which ignores latent variable. In other word, the KL divergence goes to vanish and the decoder is unable to learn relationship between latent variable and actual data.

VARIATIONAL HIERARCHICAL CONVERSATION RNN

By the inspiration of hierarchical latent variable RNN structure, named as VHRED [2], the VHCR model is developed [5] to overcome degeneration problem. Unlike VAE, the VHRED model is hierarchical in structure and continuously adds higher dimensional latent variable to each utterance which are inferred by maximizing variational lower bound. While training VHRED model it degenerates with the KL divergence term goes to zero. The degeneration problems occur because of expressive power of hierarchical decoder RNN and the VAE structure that induces data sparsity. To overcome this issue the key update done in VHCR model is, a global latent variable along with local latent variable is used to form hierarchical latent structure.

ATTENTION MECHANISM

As the attention mechanism is introduced in proposed work to improve the result of previous work and to generate relevant responses for context, this section tells about attention mechanism. One of the common problems with encoder-decoder is that at the time step t the encoder forces to encode the information which might not be fully relevant to generate target response[20]. It makes the model computationally expensive as well as produces irrelevant responses. To incorporate this problem and to improve the performance of encoder-decoder RNN the attention mechanism is firstly introduced by Bahdanau et al. [9]. It has both the properties of aligning and translating. Alignment process identifies more relevant part of input sequence to each word in the output. Translation is the process to produce appropriate output using relevant information. The attention-based models are classified into two broad categories global (Luong attention) and local (Bahdanau attention) [10]. The key difference between global and local attention is that, the global attention calculates attention weights using hidden state of decoder from the current time step whereas local attention requires all history of decoder states from previous time steps.

To train a deep learning models, two alignment like approaches are popular which make ease of latent alignment directly: soft attention which replaces probabilistic model with deterministic soft function and hard attention training a latent alignment by maximizing lower bound [21]. In this work attention mechanism is used at decoder level. It weakens the autoregressive power of decoder RNN which is responsible for data degeneration.

B. PROPOSED MODEL: VARIATIONAL HIERARCHICAL CONVERSATION RNN WITH ATTENTION MECHANISM

As discussed, the autoregressive power of hierarchical decoder RNN is main reason behind the degeneration problem. Normally the VAE models depend on local latent variable, which is insufficient to capture complex structure of sentence. At the end, auto-regressive decoder is used in response generation model which always tends to ignore latent variable [22]. This is the reason the KL divergence goes to vanish and the model collapse [23].

In existing VHCR model to alleviate the data degeneration problem a global latent variable along with local latent variable is used for conversation and for each utterance. It still makes the model to ignore latent structure of conversation and it does not stop from KL- vanishing.

In proposed approach, along with global and local latent variable, we add attention mechanism, to see the effect on decoder RNN. VHCRA extends the VHCR [5] model, it uses the hierarchical RNNs, VHCRA uses an additional attention layer to extract the important information from the context and to model conversation which is not part of VHCR.VHCRA model treats both global and latent variables and additional attention vector as random variable.

PROPOSED ARCHITECTURE

Suppose in a dataset there are N samples of conversation (c_1,c_2,…,c_n) where each c_i is sequence of sentence (Y_{i1}, Y_{i2}, …, Y_{in}). The Hierarchical RNN accepts variable length sentence as input and generates response sequence as output. Fig. 1 shows the architecture of proposed system. It integrates attention vector along with global and latent variable into VHRED structure. It consists of three layers: Encoder RNN, Context RNN, and Decoder RNN. The following section gives detailed explanation of how the input sequence is processed in these three layers to get appropriate output.

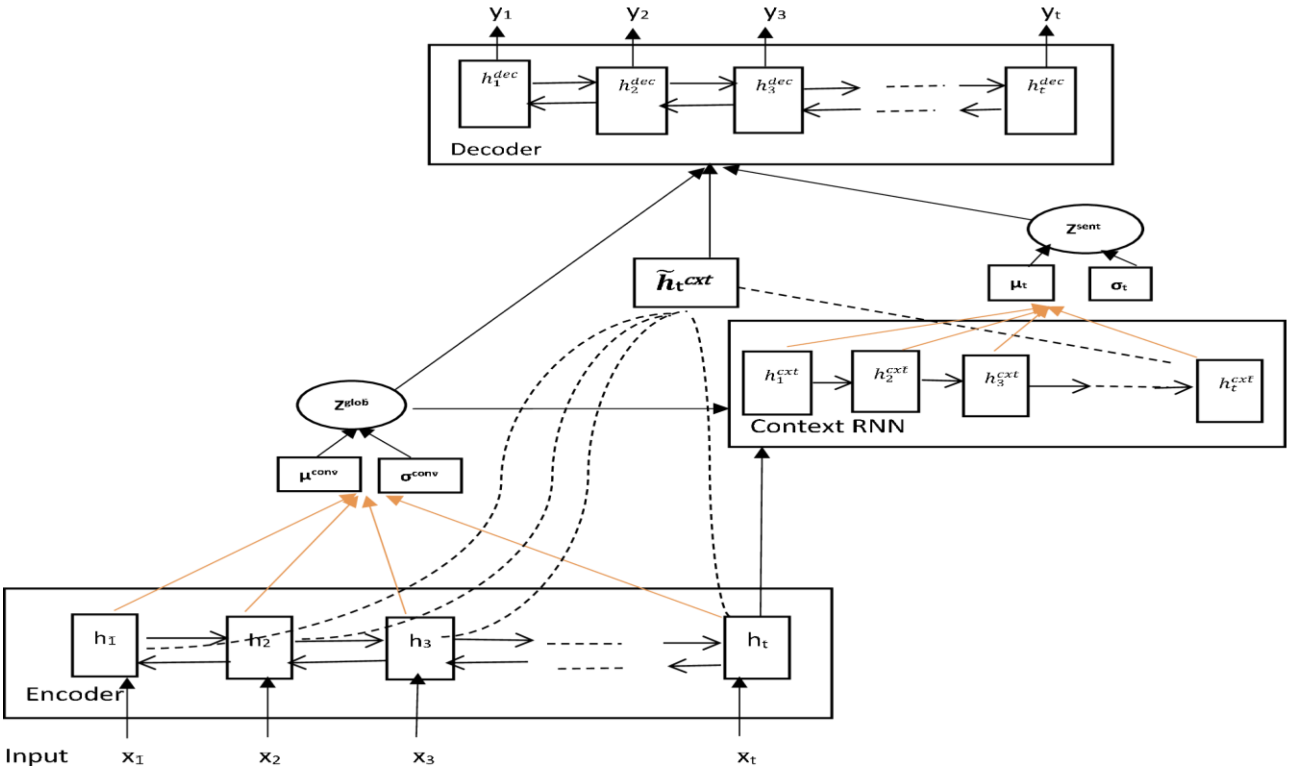

Figure 1. Architecture of VHCRA System

The proposed work introduces attention vector along with global and local latent variable for generating a sequence of sentence of conversation. As discussed earlier, the proposed VHCRA model extends VHCR response generation model. Following equations to find out local latent variable and global latent variable are inspired by the VHCR model. Both the encoder and decoder use Gated Recurrent Unit (GRU) type of recurrent neural network.

Initially encoder RNN f_θ^{enc} at the time step t encodes each utterance into encoder's hidden vector i.e. h_t^{enc} by using word embedding (see Eq. 1):

$$h_t^{enc} = f_\theta^{enc}\,(y_t), \qquad (1)$$

where f is non linear transformation which is Gated Recurrent Unit (GRU).

A global latent variable z^{Glob} is generated by using bi-directional RNN where input for it is encoder vector produced by encoder RNN (see in Eq. 2-5):

$$q_\phi(z^{Glob}|y_{1,}\ ...\ y_n) = \mathcal{N}(z|\mu^{conv}, \sigma^{conv}, \mathrm{I})\,, \qquad (2)$$

$$\mu^{conv} = MLP_\phi(h^{conv}), \qquad (3)$$

$$\sigma^{conv} = Softplus(MLP_\phi(h^{conv})), \qquad (4)$$

where h^{conv} is a vector generated by bi-directional RNN f^{conv} which runs over the utterance vector generated by encoder RNN

$$h^{conv} = f^{conv}\ (h_1^{enc}, \ldots., h_n^{enc}\ ). \tag{5}$$

The context RNN f_θ^{cxt} models the context of conversation by updating its hidden state using encoder vector along with global latent variable (see in Eq. 6):

$$h_t^{cxt} = \begin{cases} MLP_\theta\ (z^{Glob}\ ) & t = 0 \\ f_\theta^{cxt}(h_{t-1}^{cxt},\ \ h_{t-1}^{enc}, z^{Glob}\ ) & otherwise \end{cases} . \tag{6}$$

The context vector h_t^{cxt} defines conditional prior $p_\theta(z_t^{sent}|y < t)$ which is a factorized Guassian Distribution. A global attention mechanism is used at the decoder level by modifying the hidden state of context vector with attention score. In global attention mechanism, the score is content-based function which has different forms: dot, general, concat [10].The dot score function is used to modify the context vector (see Eq. 7 and 8). Finally the attentional context vector $\tilde{h}_{\mathrm{t}}{}^{cxt}$ is generated by calcualting the score function of context vector and all encoder vectors. It is a vector of values of more attentional words in the utterance

$$\tilde{h}_{\mathrm{t}}{}^{cxt} = \frac{\exp\ (\text{score}\ (h_t^{cxt}, h_t^{enc}))}{\sum_{h_{t'}^{enc}} \exp\ (\text{score}\ (h_t^{cxt}, h_t^{enc}))}, \tag{7}$$

where score function is calculated by using dot function

$$\text{score}\ (h_t^{cxt}\ , h_t^{enc}) = h_t^{cxt\top} h_t^{enc} \tag{8}$$

Finally, at the decoder level, the attentional decoder RNN $f_{\theta att}^{dec}$ generates the utterance x conditioned on attentional context vector $\tilde{h}_{\mathrm{t}}{}^{cxt}$, local latent variable z_t^{sent} along with global variable z^{Glob} (see Eq. 9)

$$p_\theta\ (y_t|y_{<t}) = f_{\theta\ att}^{dec}(y|\tilde{h}_{\mathrm{t}}{}^{cxt}, z_t^{sent}\ z^{Glob}\ ). \tag{9}$$

The variational posterior is a factorized Guassian distribution where the mean and variance are predicted from the target utterance by using feed forward neural network (see Eq. 10-11).

$$p_\theta\ (z^{Glob}) = \mathcal{N}(z|0, \mathrm{I}), \tag{10}$$

$$p_\theta\ (z_t^{sent}|y_{<t}, z^{Glob}) = \mathcal{N}(z|\mu_t\, , \sigma_t, \mathrm{I}), \tag{11}$$

where,

$$\mu_t = MLP_\phi \qquad (\ \tilde{h}_{\mathrm{t}}{}^{cxt}, z^{Glob}\ )$$

$$\sigma_t = Softplus(MLP_\phi\,(\ \tilde{h}_{\mathrm{t}}{}^{cxt}, z^{Glob}\ ))$$

The proposed model VHCRA solves the degeneration problem by modifying the context vector with attention mechanism at the decoder level. Because of attention, the model concentrates on important part within the context and produces more accurate response. It produces the local latent variables more related to the context. Because of that it achieves better and stable KL divergence, which prevents the model from collapse.

IV. EXPERIMETAL ENVIRONMENT AND SETUP

Section A discusses experimental environment, such as dataset and parameters used for comparison. Section B provides information about experimental results and analysis of the same. The subsection also provides information about a qualitative comparison of generated responses.

A. EXPERIMENTAL ENVIRONMENT

VHCRA model is evaluated using Cornell movie dialog corpus [24] which is a benchmark dataset for NLP models. It contains 220579 conversations from 617 movies. The dataset is randomly separated into training, validation and testing sets with ratio of 80:10:10. The conversations in dataset are kept like each utterance is no more than 30 words. Before training the model, data is preprocessed. In this experiment the tokenization and data preprocessing are carried by using Spacy. A GRU with hidden dimension 1000 is used for encoder and decoder RNN. The bidirectional RNN is used for latent variable distribution. The model is trained with Adam optimizer with learning rate of 0.0001. We use batch size 80 and models are trained for 30 epochs. The KL annealing is used from 0 to1 over 25k steps on dataset.

We use Pytorch Framework which is open source library for deep learning. As the dataset is large in size, it faces trouble to train the deep learning models by utilizing the CPU.

The proposed model is trained on Google Collaboratory. The proposed VHCRA model is compared with two recently developed state-of-the art response generation models, i.e., VHRED with word drop (VHRED + w.d) [2] and VHCR with utterance drop (VHCR + u.d) [5].

VHRED + w.d is a modification of HRED which uses latent variable for generation along with the word drop regularization of probability 0.25.

VHCR + u.d is similar to VHRED which uses global latent variable along with local latent variable with newly developed utterance drop regularization technique by setting probability ratio 0.25.

One of the challenges for developing response generation system is that there is no satisfying metrics for the purpose of evaluating final output of the system [25]. The following two metrics are mostly used for evaluation of conversation response generation models in literature [2, 14, 15, 26].

Negative Log-Likelihood: A Negative Log-Likelihood (NLL) is loss function widely used in neural network, it measures accuracy of the model.

Embedding-Based Metrics: learns the word vector representation and computes cosine similarity between ground truth and generated response [25].

The experimental results using above two parameters are discussed in detailed below.

B. EXPERIMENTAL RESULT

While decoding the data, it may get lost which can be measured by reconstruction term and KL divergence term (KL-Div). The reconstruction term (Recon-loss) measures how effectively the decoder has learned to reconstruct input y given its latent representation z.

The KL-divergence measures the amount of information encoded in latent variable. Table 1 shows the result of Negative Log-likelihood. The KL divergence obtained by proposed model is better and validation loss is lower than other two models. Table 1 also shows that the VHCRA model obtains higher and stable KL divergence. It proves that by using additional attention vector, the model can alleviate data degeneration problem and produce more relevant responses for given context.

In this experiment we used three embedding based evaluation metrics, i.e., Average, Greedy, Extrema. The average metrics takes mean of word embedding of each token in sentence and computes cosine similarity between response vector and ground truth vector. The extrema metrics calculates utterance level embedding by taking extreme value amongst all word vectors in sentence and computes cosine similarity between response and ground truth.

Table 1. Result of Negative log likelihood

	NLL	Recon. Loss	KL_Div
VHRED+ w.d	≤ 3.962	3.642	0.321
VHCR+ u.d	≤ 3.964	3.498	0.455
VHCRA	≤ 3.936	3.429	0.510

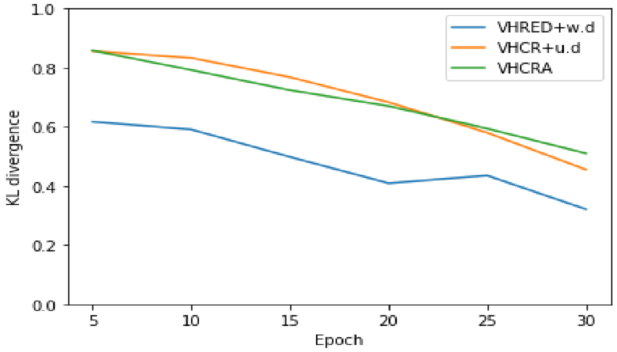

Figure 2. Decomposition of KL Divergence

Instead of computing sentence level embedding, the greedy metrics greedily matches tokens of ground truth and generates responses based on cosine similarity. The greedy approach concentrates on response that is more likely to be ground truth. Table 2 shows comparative results of three different models with embedding-based metrics.

Table 2. Results of Embedding based metrics

	Average	Extrema	Greedy
VHRED+ w.d	0.434	0.198	0.307
VHCR+ u.d	0.445	0.205	0.308
VHCRA	0.450	0.211	0.310

The proposed VHCRA achieves the better textual similarity according to all of three embedding based metrics which prove that the proposed model significantly bring improvement for both word level and utterance level similarity.

Table 3 compares the generated responses of algorithms. Overall the proposed VHCRA generates more relevant response within a context of given conversation.

Table 3. Qualitative Comparison of Generated Responses

Context	VHRED	VHCR	VHCRA
chin, you have been ordered by the ministry of public security to assist the fbi and secret service with our investigation, to provide us with whatever we	why, mr. scott, i'm not the only one who can do to the <unk> of the world.	of course, sir.	that's right. i didn't know what to say.
tim, i have to tell you something.	hey, don't you want to go to the party?	what do you want me?	wait a minute. you're going to have to wait for me
we'll see how wonderful it is after you spend the next twenty years making agent orange.	i'm not going to let you get this. i'm not going to let you get this.	and what do you mean?	that's right, it's a good idea though

V. CONCLUSION

We observed that, most of the VAE based models face the data degeneration problem, which produces irrelevant responses to context. As well as many hierarchical RNN's only concentrate on its hierarchical structure and ignore the important data within context. The research work introduced a Variational Hierarchical RNN along with Attention (VHCRA) mechanism for conversation response generation.

The proposed VHCRA model tries to alleviate both problems by combining the variational hierarchical structure with attention mechanism. The VHCRA model is applied to response generation task and it is observed that there is enough improvement over previous models in several ways. It obtains higher KL-divergence which prevents the model from data degeneration. The experimental result shows that, proposed model produces responses much closure to natural conversation.

References

[1] I. Sutskever, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, pp. 3104–3112, 2014.

[2] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, and J. Pineau, "A hierarchical latent variable encoder-decoder model for generating dialogues," *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17),* February 2017, pp. 3295–3301.

[3] I. V Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2014, pp. 3776–3783.

[4] A. Sordoni, M. Galley, M. Auli, and C. Brockett, "A neural network approach to context-sensitive generation of conversational responses," *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 196–205. https://doi.org/10.3115/v1/N15-1020.

[5] Y. Park, J. Cho, G. Kim, "A hierarchical latent structure for variational conversation modeling," *arXiv:1804.03424v2*, 2018. https://doi.org/10.18653/v1/N18-1162.

[6] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," *arXiv:1406.5298v2*, pp. 1–9, 2009.

[7] H. Bahuleyan, L. Mou, O. Vechtomova, and P. Poupart, "Variational attention for sequence-to-sequence models,"arXiv preprint arXiv:1712.08207, 2017.

[8] M. Zhou, C. Xing, Y. Wu, W. Wu, Y. Huang, "Hierarchical recurrent attention network for response generation," Proceedings of the *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018, pp. 5610–5617.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, pp. 1–15, 2015.

[10] M. Luong and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421. https://doi.org/10.18653/v1/D15-1166.

[11] J. Weizenbaum, "ELIZA – A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966. https://doi.org/10.1145/365153.365168.

[12] B. Abushawar and E. Atwell, "ALICE chatbot : Trials and outputs," *Computación y Sistemas*, vol. 19, no. 4, pp. 625–632, 2015. https://doi.org/10.13053/cys-19-4-2326.

[13] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 2, pp. 25–35, 2017. https://doi.org/10.1145/3166054.3166058.

[14] X. Shen and H. Su, "A conditional variational framework for dialog generation," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 504–509. https://doi.org/10.18653/v1/P17-2080.

[15] C. Xing, W. Wu, Y. Wu, and J. Liu, "Topic aware neural response generation," *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2016, pp. 3351–3357.

[16] H. Zheng, W. E. I. Wang, W. Chen, and A. K. Sangaiah, "Automatic generation of news comments based on gated attention neural networks," *IEEE Access*, vol. 6, pp. 702–710, 2018. https://doi.org/10.1109/ACCESS.2017.2774839.

[17] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 1577–1586. https://doi.org/10.3115/v1/P15-1152.

[18] J.He, B.Wang, M. Fu, "Hierarchical attention and knowledge matching networks with information enhancement for end-to-end task-oriented dialog systems," *IEEE Access*, vol. 7, pp. 18871–18883, 2019. ttps://doi.org/10.1109/ACCESS.2019.2892730.

[19] C. Mellon and U. C. Berkeley, "Tutorial on variational autoencoders," *arXiv Prepr. arXiv1606.05908*, pp. 1–23, 2016.

[20] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, 2018. https://doi.org/10.1109/MCI.2018.2840738.

[21] Y. Deng, Y. Kim, and A. M. Rush, "Latent alignment and variational attention," *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9735–9747.

[22] X. Shen, H. Su, S. Niu, and V. Demberg, "Improving variational encoder-decoders in dialogue generation," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 5456–5463.

[23] X. Shen and H. Su, "Towards Better Variational Encoder-Decoders in Seq2Seq Tasks," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) Towards*, 2018, no. 1, pp. 8155–8156.

[24] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations : A new approach to understanding coordination of linguistic style in dialogs," *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2011, pp. 76-87.

[25] C. Liu, R. Lowe, I. V Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 2122–2132. https://doi.org/10.18653/v1/D16-1230.

[26] Z. Xie, "Neural text generation: A practical guide," in *arXiv preprint arXiv:1711.09534*, 2018, pp. 1–21.

***Sandeep A. THORAT** is an Associate Professor at Department of Computer Science and Engineering, Rajarambapu Institute of Technology, Sangli, India. He did PhD in Computer Science from Shivaji University, Kolhapur and M.Tech from IIIT Hydrabad with specialization in Information Security. His area of interest are Machine Learning, Wireless Networks and Information Security. He has authored a book on C programming and designed a MOOCs course on Udemy platform.*

*Mrs. **KOMAL JADHAV** is a M. Tech Scholar at Department of Computer Science and Engineering, Rajarambapu Institute of Technology, Sakharale, India. Her area of interest are Machine Learning, Deep learning, and Natural Language Processing. Presently she is working on Chatbot design, implementations and deployment.*