Dirichlet Latent Variable Hierarchical Recurrent Encoder-Decoder in Dialogue Generation

Min Zeng Yisen Wang† **Yuan Luo**
Department of Computer Science and Engineering
Shanghai Jiao Tong University
`eewangyisen@gmail.com; {min_zeng, yuanluo}@sjtu.edu.cn`

Abstract

Variational encoder-decoders have achieved well-recognized performance in the dialogue generation task. Existing works simply assume the Gaussian priors of the latent variable, which are incapable of representing complex latent variables effectively. To address the issues, we propose to use the Dirichlet distribution with flexible structures to characterize the latent variables in place of the traditional Gaussian distribution, called Dirichlet Latent Variable Hierarchical Recurrent Encoder-Decoder model (Dir-VHRED). Based on which, we further find that there is redundancy among the dimensions of latent variable, and the lengths and sentence patterns of the responses can be strongly correlated to each dimension of the latent variable. Therefore, controllable responses can be generated through specifying the value of each dimension of the latent variable. Experimental results on benchmarks show that our proposed Dir-VHRED yields substantial improvements on negative log-likelihood, word-embedding-based and human evaluations.

1 Introduction

Recurrent neural networks (RNNs) (Bengio et al., 2003) have achieved great success on many natural language processing tasks. For the dialogue generation task, a RNN-based Hierarchical Recurrent Encoder-Decoder (HRED) was first proposed in (Serban et al., 2016). It consists of three RNNs: encoder RNN, context RNN and decoder RNN. Firstly, the encoder RNN encodes each utterance into a fixed-size real-valued vector through word embedding. Then, the last hidden state of encoder RNN is feed into the context RNN to summarize the dialogue information. Finally, the decoder RNN takes the last state of context RNN as input and produces the probability distribution of the word in the next utterance.

Although it outperforms n-gram and other neural network language models, HRED only produces one word at a time, which is unable to fully grasp the holistic high-level syntactic properties (*e.g.*, topics, tones or sentiment) (Bowman et al., 2015). When the sentence grows longer, it has the drawback of tending to generate short and generic responses (Vinyals and Le, 2015). Thus, Serban et al. (2017) proposed the Variational Hierarchical Recurrent Encoder-Decoder (VHRED) by combining HRED with Variational Autoencoders (VAEs) (Kingma and Welling, 2013) that introduced a latent variable to characterize the sentence-level representation for learning holistic properties. However, VHRED imposes a symmetric distribution (*i.e.*, Gaussian distribution) to the latent variable, which, though facilitating analyzing, are incapable of representing complex latent variables effectively. Therefore, in order to generate more meaningful and expressive responses, a more flexible and tractable prior distribution of the latent variable is needed.

In this paper, we propose to use the Dirichlet distribution to characterize the latent variable in VHRED, named Dir-VHRED. Dirichlet distribution is a popular conjugate prior for Multinomial distributions in Bayesian statistics. It can be concave or convex, monotonously rising or decreasing, symmetrical or asymmetrical, which makes it more flexible for better capturing the sentence-level properties. Our main contributions are summarized as follows:

- We introduce Dirichlet distribution to VAE-based dialogue generation model and propose the Dir-VHRED model for better grasping the sentence-level properties and generating more meaningful and expressive responses.

---

†Corresponding author: Yisen Wang.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1267–1272,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

- We find that the lengths and sentence patterns of the responses can be strongly correlated to each dimension of the latent variable.
- Experiments on three kinds of evaluation metrics demonstrate the superiority of our proposed Dir-VHRED model.

2 Preliminaries

2.1 Variational Autoencoders

The key idea of Variational autoencoders (VAEs) is to reconstruct the input x through the latent variable z (Kingma and Welling, 2013). As the log-likelihood $\log p_\theta(x)$ is intractable, its lower bound *Evidence Lower BOund* (ELBO), denoted as $\mathcal{L}(\theta, \phi; x)$, is involved to make the maximization tractable:

$$\begin{aligned}\log p_\theta(x) &\geq \mathcal{L}(\theta, \phi; x)\\ &= -KL(q_\phi(z|x)||p(z)) \\ &\quad + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)],\end{aligned} \tag{1}$$

where $p(z)$ denotes the prior distribution of z, and $q_\phi(z|x)$ is used for approximating the intractable true posterior distribution $p_\theta(z|x)$. Note that the total loss is the negative of ELBO.

2.2 Dialogue Model

Assuming dialogue $\mathcal{D}$ consisting of N utterances, the VAE-based dialogue model generates the responses by utilizing the latent variable z. The generation of the next utterance is defined by:

$$\begin{aligned}p_\theta(z_n|W_{<n}) &= \mathcal{N}(\mu_{prior(W_{<n}),\Sigma_{prior}(W_{<n})}),\\ p_\theta(W_n|z_n, W_{<n}) &= \prod_m p_\theta(w_{n,m}|z_n, w_{n,<m}, W_{<n}),\end{aligned} \tag{2}$$

where W_n is the n-th utterance of the dialogue and $w_{n,m}$ indicates the m-th word of the n-th utterance. As the log-likelihood is intractable, instead of pursuing the exact maximum of the log-likelihood, we maximize its lower bound ELBO:

$$\begin{aligned}&\log p_\theta(W_1, W_2, \cdots, W_N) \geq \\ &\sum_{n=1}^{N} -KL[q_\phi(z_n|W_1, W_2, \cdots, W_N)||p_\theta(z_n|W_{<n})] \\ &\quad + \mathbb{E}_{q_\phi(z_n|W_1,W_2,\cdots,W_N)}[\log(p_\theta(W_N|z_n, W_{<n}))],\end{aligned} \tag{3}$$

where $p_\theta(z_n|W_{<n})$ denotes the prior distribution of z_n, and $q_\phi(z_n|W_1, W_2, \cdots, W_N)$ is used for approximating the intractable true posterior distribution $p_\theta(z_n|W_1, W_2, \cdots, W_N)$.

2.3 Dirichlet Distribution

Dirichlet distribution, a family member of continuous multivariate probability distribution, is regarded as a multivariate generalization of the Beta distribution. In case of the Dirichlet distribution, it is a conjugate prior for the Multinomial distribution. The probability density function of Dirichlet distribution is given by:

$$f(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{4}$$

where parameter $\alpha_i > 0$ is a K dimension vector and $B(\alpha)$ is a Beta function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}, \quad \alpha = (\alpha_1, \cdots, \alpha_K), \tag{5}$$

When $\alpha_1 = \alpha_2 = \cdots = \alpha_K$, Dirichlet distribution is a symmetric distribution. Particularly, when all the factors in α equal 1, Dirichlet distribution becomes a uniform distribution.

3 The Proposed Dir-VHRED Model

Mathematically, Dirichlet distribution owns a flexible structure. With different settings of parameter α, Dirichlet distribution may have various forms, which can be concave or convex, monotonously rising or decreasing, symmetrical or asymmetrical. Therefore, equipped with the diverse structure, Dirichlet distribution is capable of modeling the complex latent variable. Furthermore, Dirichlet distribution has paved its way in many natural language processing tasks like topic model, text classification and so on. For example, it was introduced in the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model the topic distribution and word distribution, and experimentally presents a well-recognized performance.

Motivated by the above facts, we thus introduce Dirichlet distribution to model the latent variable z in order to better grasp the sentence-level properties which further encourages the better responses. With the distribution of z changing from Gaussian to Dirichlet, we may face a tough but important problem: the reparameterization of z. The variables with any distribution can be generated through the transformation of the samples from standard Uniform distribution if the corresponding inverse Cumulative Distribution Function (CDF) is known. For the Gaussian latent variable, it can make use of the inverse CDF to do the reparameterization (Kingma and Welling, 2013). While for the Dirichlet latent variable, unlike the known CDF of Gaussian distribution, the CDF of Dirichlet distribution is too complex to obtain. Inspired from (Jankowiak and Obermeyer, 2018), we use the reject sampling to solve the Dirichlet reparameterization problem. Furthermore, to mitigate KL-vanishing problem, in contrast to the static weight scheme in (Bowman et al., 2015), the imposed weight on KL-divergence in our model is dynamic. Since the KL-divergence and reconstruction loss antagonize each other, by setting the weight as the reciprocal of the reconstruction loss, the KL and reconstruction loss enable to keep a dynamic balance, which leads to a more stable KL-divergence.

Following the techniques described above, ELBO then has the form:

$$\begin{aligned}\log p_\theta(x) &\geq \mathcal{L}(\theta, \phi; x)\\ &= -\lambda KL(q_\phi(z|x)||p(z))\\ &\quad + \mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x|z))],\end{aligned} \tag{6}$$

where

$$\lambda = -1/\mathbb{E}_{q_\phi(z|x)}[\log(p_\theta(x|z))], \tag{7}$$

and the KL-divergence term can be derived as*:

$$\begin{aligned}&KL(q_\phi(z|x)||p(z)) =\\ &\log\Gamma(\sum_{k=1}^{K}\alpha_k) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) - \log\Gamma(\sum_{k=1}^{K}\beta_k)\\ &+\sum_{k=1}^{K}\log\Gamma(\beta_k) + \sum_{k=1}^{K}(\alpha_k - \beta_k)(\psi(\alpha_k) - \psi(\sum_{k=1}^{K}\alpha_k)),\end{aligned} \tag{8}$$

where α, β are parameters of Dirichlet distribution $q_\phi(z|x)$ and $p(z)$ respectively, K is the dimension of z, and ψ is the Digamma function.

4 Experiments

4.1 Datasets

All experiments are conducted on the following two dialogue datasets:

Ubuntu Dialogue Corpus (Ubuntu) (Lowe et al., 2015): Ubuntu Dialogue Corpus is a large dataset for research in unstructured multi-turn dialogue systems. It contains 1 million two-person conversation and the conversation consists of average 8 turns.

Cornell Movie Dialogs Corpus (Movie) (Danescu-Niculescu-Mizil and Lee, 2011): Cornell Movie Dialogs Corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts. It contains 220,579 conversations from 617 movies.

4.2 Experimental Setup

We conduct a series of experiments to compare our proposed Dir-VHRED with two baselines: HRED (Serban et al., 2016) and VHRED (Serban et al., 2017). All the RNN modules (*i.e.*, the encoder, context and decoder RNNs) adopt a single-layer GRU (Cho et al., 2014) with 1000 hidden units. The word drop rate is set to 0.25, the dimensionality of latent variable z is 3, and the word embedding size is 200†.

For VHRED, the weight parameter λ of KL-divergence is initialized as 0 and gradually increased to 1 at the $20,000$-th/$100,000$-th training steps for Movie/Ubuntu datasets. While, for Dir-VHRED, the parameter λ is adaptively determined by Eq. (7).

For the dataset, we truncate utterances longer than 30 words and split the train/validation/test sets by 0.8/0.1/0.1 respectively.

All models are trained by Adam optimizer (Kingma and Ba, 2014) with batch size 40 and learning rate 1×10^{-4}. At the training time, we stop the training when the loss on the validation set does not decrease within 5 epochs. At the evaluating time, beam search (Wiseman and Rush, 2016) with beam size 5 is used for generating output responses. Both the single response (1-turn) and the three consecutive responses (3-turn) are evaluated for each model. Our code is available at `https://github.com/cloversjtu/dir-vhred`.

4.3 Performance Evaluation

In order to comprehensively evaluate the model performance on dialogue generation, we adopt the following three evaluation metrics: 1) negative

---

[*] Detailed derivation can be found in Appendix A.1.

[†] The word embedding is obtained from Word2Vec embeddings trained on the Google News Corpus.

log-likelihood metric to measure the loss of the generating procedure; 2) word-embedding metric to measure the cosine distance between the generated responses and the ground truth responses; and 3) human evaluation.

4.3.1 Negative Log-likelihood Evaluation

Table 1 reports the comparisons of the per-word negative log-likelihood (NLL) composition of different models. NLL consists of the reconstruction loss (*reconstruction*) and the KL-divergence (*KL-div.*). KL-divergence indicates the information encoded in z, and reconstruction loss represents the loss of reconstructing x through z. As can be seen from Table 1, with almost the same KL-divergence, Dir-VHRED achieves the lowest NLL on both datasets, implying better performance compared with VHRED. Dir-VHRED also has the lower reconstruction loss, which indicates that Dirichlet prior is better than the Gaussian prior for reconstructing the responses.

Table 1: Results of NLL on Ubuntu and Movie datasets with different models. The $\leq$ symbol denotes the variational bound.

Ubuntu dataset			
Model	NLL	*reconstruction*	*KL-div.*
HRED	3.844	-	-
VHRED	$\leq$ 4.132	3.765	0.367
Dir-VHRED	$\leq$ 3.999	3.614	0.385
Movie dataset			
Model	NLL	*reconstruction*	*KL-div.*
HRED	3.944	-	-
VHRED	$\leq$ 4.233	3.904	0.330
Dir-VHRED	$\leq$ 4.073	3.741	0.332

4.3.2 Word-embedding-based Evaluation

Word-embedding metric is designed to measure the similarity between words, and can be divided into three categories: *average* (Foltz et al., 1998), *greedy* (Rus and Lintean, 2012) and *extrema* (Forgues et al., 2014). The *average* metric calculates sentence-level embeddings, while the *greedy* and *extrema* ones compute the word-to-word cosine similarity. Their difference lies on that *greedy* takes the average word vector in the sentence as the sentence embedding while *extrema* adopts the extreme value of these word vectors. Tables 2 and 3 demonstrate the results of word-embedding metric on 1-turn and 3-turn responses from different models. Embedding metrics first map the generating responses to a vector space and then compute the cosine similarly with corresponding ground responses. Therefore, it can be used to measure the sentence-level semantic similarity to a certain extent. Dir-VHRED achieves the best performance among the three word-embedding metrics on 1-turn and 3-turn responses of the two datasets. That is, the responses generated by Dir-VHRED own the highest topic similarity with the ground truth.

Table 2: Results of three embedding metrics on Ubuntu dataset with 1-turn and 3-turn responses.

Model	*Average*	*Greedy*	*Extrema*
1-turn			
HRED	0.448	0.297	0.271
VHRED	0.506	0.355	0.286
Dir-VHRED	**0.538**	**0.387**	**0.303**
3-turn			
HRED	0.426	0.275	0.260
VHRED	0.524	0.366	0.292
Dir-VHRED	**0.552**	**0.393**	**0.313**

Table 3: Results of three embedding metrics on Movie dataset with 1-turn and 3-turn responses.

Model	*Average*	*Greedy*	*Extrema*
1-turn			
HRED	0.461	0.316	0.329
VHRED	0.519	0.376	0.356
Dir-VHRED	**0.554**	**0.409**	**0.358**
3-turn			
HRED	0.483	0.334	0.337
VHRED	0.530	0.382	0.356
Dir-VHRED	**0.566**	**0.412**	**0.363**

4.3.3 Human Evaluation

Moreover, we invite 7 qualified volunteers to do the human evaluation on 700 randomly sampled dialogues from Ubuntu dataset. Volunteers are required to rate the generated responses from different models with scores from 1 to 3 (3 is the best), in term of the sentence coherence and contextual

Table 4: Responses generated by different models.

# 1 Input sentence:
that is a upstream bug and red hat marked it <unk>.
# 1 Generated response:
HRED: i know it's not a bug, but i'll try it out.
VHRED: can you provide a link for that?
Dir-VHRED: so i should report it to the bug report? i don't see any bug reports.

# 2 Input sentence:
i tried but still its not workin.
# 2 Generated response:
HRED: what kind of laptop?
VHRED: is it a laptop? or the fn key is not set?
Dir-VHRED: what is the model of your laptop? what model?

# 3 Input sentence:
i tried that and even rebooting the pc.
# 3 Generated response:
HRED: what does ifconfig say?
VHRED: check your etc network interfaces.
Dir-VHRED: what's the error message?

topic. The average human evaluation scores for HRED, VHRED and Dir-VHRED are 2.50, 2.65 and 2.70 respectively. We also conduct Friedman test (Friedman, 1940) on the human evaluation score to check its statistical significance, and the value of Chi-square test statistics is 5696.25 which indicates that the P-value is less than 0.05. Thus, Dir-VHRED is significantly better than others.

Table 4 lists some examples of the generated sentences of different models. It can be seen that the responses generated by Dir-VHRED best match the context given the input sentence, compared with other models. Dir-VHRED experimentally demonstrates to be able to generate fluent and diverse responses which match the human daily speaking style smoothly.

4.4 Interpretability of Dir-VHRED

In our above experiments, we find an interesting phenomenon for Dir-VHRED that there are many non-active dimensions of z and the number of active ones is almost always around 3. Therefore, we conduct additional ablation study experiments to analyze each dimension of z. To be specific, we circularly set one dimension of z to 1 and the others to 0, like "100", "010" and "001".

We demonstrate some generated responses through ablation study on each dimension of z on Ubuntu and Movie datasets in Table 5. We can observe that the **dimension 1** (*dim1*) is inclined to generate the shortest responses which, to a certain extent, shows the attitude of the speaker. Specifically, it prefers to generate the responses like "thanks", "no" or "yes". The **dimension 2** (*dim2*) usually generates longer and informative responses. For example on Movie dataset, the generated response is prone to end with "are you?" or "aren't you?", which shows a preference for rhetorical question. Other examples include "you're not going to be here, are you?" and "well, you're a good man, aren't you?". Intuitively, since the Movie dataset is a collection of conversations extracted from movie scripts, the generated responses are closer to daily dialogues. While for Ubuntu, the responses do not have strong individual emotion and rhetorical question rarely appear, as the conversations in this dataset are professional dialogue about Ubuntu system knowledge. The **dimension 3** (*dim3*) focuses on generating the longest and most comprehensive responses that are usually declarative sentences and are highest close to the dialogue topic.

Table 5: Responses generated by ablation study on different dimension of z in Dir-VHRED.

# 1 Input sentence:
does grub stay into hdd after installing another os?
# 1 Generated response:
dim1: yes.
dim2: yes, it will be in the mbr of the drive.
dim3: yes, it will overwrite grub, but it will overwrite the mbr, and then you can use the grub menu to boot the os.

# 2 Input sentence:
that leads me to the next question lol ... i do not know how to create new volumes exactly in cmdline, usually i use a gui.
# 2 Generated response:
dim1: gksudo.
dim2: alt f2 gksudo nautilus and gksudo nautilus will work.
dim3: <unk> is a gui app for managing users, you can use gksudo nautilus and you can use gksudo nautilus toget a root terminal and you can use.

# 3 Input sentence:
i see this use at your own risk. http askubuntu.com questions <unk> ubuntu 12 04 nvidia gtx 460 video card installation.
# 3 Generated response:
dim1: thanks.
dim2: thanks, i 'll give that a try. thanks !
dim3: i have the nvidia drivers installed, but i can't see the nvidia logo on the screen. i have to install the drivers for the graphics card.

# 4 Input sentence:
she isn't my cousin.
# 4 Generated response:
dim1: yes.
dim2: she 's a nurse, isn't she ?
dim3: yes, she 's a good friend of mine, she 's a nurse.

Therefore, we can easily generate the responses with the preferred style and length through changing the value of each dimension of z.

5 Conclusion

In this paper, we proposed to use Dirichlet distribution in place of traditional Gaussian distribution in VHRED for dialogue generation, called Dir-VHRED, to well capture the sentence-level properties. We also provided a new way on the setting of the weight of KL-divergence to alleviate KL-vanishing problem. Moreover, we found that the lengths and sentence patterns of the generated responses are correlated to the value of each dimension of the latent variable, which can be used for generating the required responses. Experiments on benchmark datasets show the superior of our proposed Dir-VHRED model.

Acknowledgments

We would like to thank Junqing Gu and Xiaoyu Shen for their constructive discussions. We also appreciate the code shared in (Park et al., 2018).

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *the 2nd Workshop on Cognitive Modeling and Computational Linguistics on ACL*.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Modern Machine Learning and Natural Language Processing workshop on NeurIPS*.

Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Martin Jankowiak and Fritz Obermeyer. 2018. Pathwise derivatives beyond the reparameterization trick. *arXiv preprint arXiv:1806.01851*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. *arXiv preprint arXiv:1804.03424*.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *the 7th Workshop on Building Educational Applications Using NLP on ACL*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*.

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

Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*.