# Modeling Semantic Relationship in Multi-turn Conversations with Hierarchical Latent Variables

**Lei Shen**[1,2]  **Yang Feng**[1,2*]  **Haolan Zhan**[2,3]

[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]State Key Laboratory of Computer Science,
Institute of Software, Chinese Academy of Sciences, Beijing, China

{shenlei17z, fengyang}@ict.ac.cn   zhanhl@ios.ac.cn

## Abstract

Multi-turn conversations consist of complex semantic structures, and it is still a challenge to generate coherent and diverse responses given previous utterances. It's practical that a conversation takes place under a background, meanwhile, the query and response are usually most related and they are consistent in topic but also different in content. However, little work focuses on such hierarchical relationship among utterances. To address this problem, we propose a Conversational Semantic Relationship RNN (CSRR) model to construct the dependency explicitly. The model contains latent variables in three hierarchies. The discourse-level one captures the global background, the pair-level one stands for the common topic information between query and response, and the utterance-level ones try to represent differences in content. Experimental results show that our model significantly improves the quality of responses in terms of fluency, coherence and diversity compared to baseline methods.

## 1 Introduction

Inspired by the observation that real-world human conversations are usually multi-turn, some studies have focused on multi-turn conversations and taken context (history utterances in previous turns) into account for response generation. How to model the relationship between the response and context is essential to generate coherent and logical conversations. Currently, the researchers employ some hierarchical architectures to model the relationship. Serban et al. (2016) use a context RNN to integrate historical information, Tian et al. (2017) sum up all utterances weighted by the similarity score between an utterance and the query, while Zhang et al. (2018) apply attention mechanism on history utterances. Besides, Xing et al.

(2018) add a word-level attention to capture fine-grained features.

In practice, we usually need to understand the meaning of utterances and capture their semantic dependency, not just word-level alignments (Luo et al., 2018). As shown in Table 1, this short conversation is about speaker A asks the current situation of speaker B. At the beginning, they talk about B's position. Then in the last two utterances, both speakers think about the way for B to come back. A mentions "*umbrella*", while B wants A to "*pick him/her up*". What's more, there is no "word-to-word" matching in query and response. Unfortunately, the aforementioned hierarchical architectures do not model the meaning of each utterance explicitly and has to summarize the meaning of utterances on the fly during generating the response, and hence there is no guarantee that the inferred meaning is adequate to the original utterance. To address this problem, variational autoencoders (VAEs) (Kingma and Welling, 2014) are introduced to learn the meaning of utterances explicitly and a reconstruction loss is employed to make sure the learned meaning is faithful to the corresponding utterance. Besides, more variations are imported into utterance level to help generate more diverse responses.

| |
|---|
| A: Where are you? |
| B: I'm stuck in my office with rain. |
| A: Didn't you bring your umbrella? |
| B: No. Please come and pick me up. |

Table 1: An example of the semantic relationship in a multi-turn conversation.

However, all these frameworks ignore the practical situation that a conversation usually takes place under a background with two speakers communicating interactively and query is the most relevant utterance to the response. Hence we need

---
*Corresponding Author

to pay more attention to the relationship between query and response. To generate a coherent and engaging conversation, query and response should be consistent in topic and have some differences in content, the logical connection between which makes sure the conversation can go on smoothly.

On these grounds, we propose a novel **Conversational Semantic Relationship RNN (CSRR)** to explicitly learn the semantic dependency in multi-turn conversations. CSRR employs hierarchical latent variables based on VAEs to represent the meaning of utterances and meanwhile learns the relationship between query and response. Specifically, CSRR draws the background of the conversation with a discourse-level latent variable and then models the consistent semantics between query and response, e.g. the topic, with a common latent variable shared by the query and response pair, and finally models the specific meaning of the query and the response with a certain latent variable for each of them to capture the content difference. With these latent variables, we can learn the relationship between utterances hierarchically, especially the logical connection between the query and response. What is the most important, the latent variables are constrained to reconstruct the original utterances according to the hierarchical structure we define, making sure the semantics flow through the latent variables without any loss. Experimental results on two public datasets show that our model outperforms baseline methods in generating high-quality responses.

## 2 Approach

Given $n$ input messages $\{\mathbf{u}_t\}_{t=0}^{n-1}$, we consider the last one $\mathbf{u}_{n-1}$ as query and others as context. $\mathbf{u}_n$ denotes corresponding response.

The proposed model is shown in Figure 1. We add latent variables in three hierarchies to HRED (Serban et al., 2016). $\mathbf{z}^c$ is used to control the whole background in which the conversation takes place, $\mathbf{z}^p$ is for the consistency of topic between query and response pair, $\mathbf{z}^q$ and $\mathbf{z}^r$ try to model the content difference in each of them, respectively. For simplicity of equation description, we use $n-1$ and $n$ as the substitution of $q$ and $r$.

### 2.1 Context Representation

Each utterance $\mathbf{u}_t$ is encoded into a vector $\mathbf{v}_t$ by a bidirectional GRU (BiGRU), $f_\theta^{utt}$:

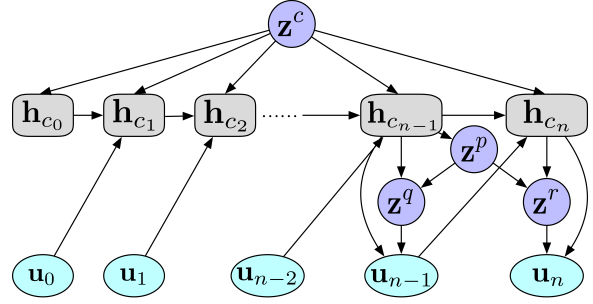$$\mathbf{v}_t = f_\theta^{utt}(\mathbf{u}_t) \tag{1}$$



Figure 1: Graphical model of CSRR. $\mathbf{u}_t$ is the $t$-th utterance, $\mathbf{h}_{c_t}$ encodes context information up to time $t$.

For the inter-utterance representation, we follow the way proposed by Park et al. (2018), which is calculated as:

$$\mathbf{h}_{c_t} = \begin{cases} \text{MLP}_\theta(\mathbf{z}^c), & \text{if } t = 0 \\ f_\theta^{ctx}(\mathbf{h}_{c_{t-1}}, \mathbf{v}_{t-1}, \mathbf{z}^c), & \text{otherwise} \end{cases} \tag{2}$$

$f_\theta^{ctx}(\cdot)$ is the activation function of GRU. $\mathbf{z}^c$ is the discourse-level latent variable with a standard Gaussian distribution as its prior distribution, that is:

$$p_\theta(\mathbf{z}^c) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \tag{3}$$

For the inference of $\mathbf{z}^c$, we use a BiGRU $f^c$ to run over all utterance vectors $\{\mathbf{v}_t\}_{t=0}^n$ in training set. ($\{\mathbf{v}_t\}_{t=0}^{n-1}$ in test set):

$$q_\phi(\mathbf{z}^c|\mathbf{v}_0, ..., \mathbf{v}_n) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}^c, \boldsymbol{\sigma}^c \mathbf{I}) \tag{4}$$

$$\text{where } \mathbf{v}^c = f^c(\mathbf{v}_0, ..., \mathbf{v}_n) \tag{5}$$

$$\boldsymbol{\mu}^c = \text{MLP}_\phi(\mathbf{v}^c) \tag{6}$$

$$\boldsymbol{\sigma}^c = \text{Softplus}(\text{MLP}_\phi(\mathbf{v}^c)) \tag{7}$$

$\text{MLP}(\cdot)$ is a feed-forward network, and Softplus function is a smooth approximation to the ReLU function and can be used to ensure positiveness (Park et al., 2018; Serban et al., 2017; Chung et al., 2015).

### 2.2 Query-Response Relationship Modeling

According to VAEs, texts can be generated from latent variables (Shen et al., 2017). Motivated by this, we add two kinds of latent variables: pair-level and also utterance-level ones for query and response.

As depicted in Figure 1, $\mathbf{h}_{\mathbf{c}_{n-1}}$ encodes all context information from utterance $\mathbf{u}_0$ to $\mathbf{u}_{n-2}$. We use $\mathbf{z}^p$ to model the topic in query and response pair. Under the same topic, there are always some differences in content between query and response, which is represented by $\mathbf{z}^q$ and $\mathbf{z}^r$, respectively. We first define the prior distribution of $\mathbf{z}^p$

5498

as follows:

$$p_\theta(\mathbf{z}^p|\mathbf{u}_{<n-1}, \mathbf{z}^c) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{n-1}, \boldsymbol{\sigma}_{n-1}\mathbf{I}) \quad (8)$$

$\mathbf{u}_{<n-1}$ denotes utterances $\{\mathbf{u}_i\}_{i=0}^{n-2}$, $\boldsymbol{\mu}_{n-1}$ and $\boldsymbol{\sigma}_{n-1}$ are calculated as:

$$\boldsymbol{\mu}_{n-1} = \text{MLP}_\theta(\mathbf{h}_{\mathbf{c}_{n-1}}, \mathbf{z}^c) \quad (9)$$

$$\boldsymbol{\sigma}_{n-1} = \text{Softplus}(\text{MLP}_\theta(\mathbf{h}_{\mathbf{c}_{n-1}}, \mathbf{z}^c)) \quad (10)$$

Since $\mathbf{z}^q(\mathbf{z}^{n-1})$ and $\mathbf{z}^r(\mathbf{z}^n)$ are also under the control of $\mathbf{z}^p$, we define the prior distributions of them as:

$$p_\theta(\mathbf{z}^i|\mathbf{u}_{<i}, \mathbf{z}^c, \mathbf{z}^p) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\mathbf{I}) \quad (11)$$

Here, $i = n-1$ or $n$. The means and the diagonal variances are computed as:

$$\boldsymbol{\mu}_i = \text{MLP}_\theta(\mathbf{h}_{\mathbf{c}_i}, \mathbf{z}^c, \mathbf{z}^p) \quad (12)$$

$$\boldsymbol{\sigma}_i = \text{Softplus}(\text{MLP}_\theta(\mathbf{h}_{\mathbf{c}_i}, \mathbf{z}^c, \mathbf{z}^p)) \quad (13)$$

The posterior distributions are:

$$q_\phi(\mathbf{z}^p|\mathbf{u}_{\leq n-1}, \mathbf{z}^c) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}'_{n-1}, \boldsymbol{\sigma}'_{n-1}\mathbf{I}) \quad (14)$$

$$q_\phi(\mathbf{z}^i|\mathbf{u}_{\leq i}, \mathbf{z}^c, \mathbf{z}^p) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}'_i, \boldsymbol{\sigma}'_i\mathbf{I}) \quad (15)$$

$q_\phi(\cdot)$ is a recognition model used to approximate the intractable true posterior distribution. The means and the diagonal variances are defined as:

$$\boldsymbol{\mu}'_{n-1} = \text{MLP}_\phi(\mathbf{v}_{n-1}, \mathbf{v}_n, \mathbf{h}_{\mathbf{c}_{n-1}}, \mathbf{z}^c) \quad (16)$$

$$\boldsymbol{\sigma}'_{n-1} = \text{Softplus}(\text{MLP}_\phi(\mathbf{v}_{n-1}, \mathbf{v}_n, \mathbf{h}_{\mathbf{c}_{n-1}}, \mathbf{z}^c)) \quad (17)$$

$$\boldsymbol{\mu}'_i = \text{MLP}_\phi(\mathbf{v}_i, \mathbf{h}_{c_i}, \mathbf{z}^c, \mathbf{z}^p) \quad (18)$$

$$\boldsymbol{\sigma}'_i = \text{Softplus}(\text{MLP}_\phi(\mathbf{v}_i, \mathbf{h}_{c_i}, \mathbf{z}^c, \mathbf{z}^p)) \quad (19)$$

Note that in Equation 16 and 17, both $\mathbf{v}_{n-1}$ and $\mathbf{v}_n$ are taken into consideration, while Equation 18 and 19 use $\mathbf{z}^p$ and corresponding $\mathbf{v}_i$.

## 2.3 Training

Because of the existence of latent variables in query-response pair, we use decoder $f_\theta^{dec}$ to generate $\mathbf{u}_{n-1}$ and $\mathbf{u}_n$:

$$p_\theta(\mathbf{u}_i|\mathbf{u}_{<i}) = f_\theta^{dec}(\mathbf{u}_i|\mathbf{h}_{\mathbf{c}_i}, \mathbf{z}^c, \mathbf{z}^p, \mathbf{z}^i) \quad (20)$$

The training objective is to maximize the following variational lower-bound:

$$\begin{aligned}
\log p_\theta(\mathbf{u}_{n-1}, \mathbf{u}_n|\mathbf{u}_0, ..., \mathbf{u}_{n-2}) \geq \\
\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{u}_i|\mathbf{z}^c, \mathbf{z}^p, \mathbf{z}^i, \mathbf{u}_{<i})] \\
- D_{KL}(q_\phi(\mathbf{z}^c|\mathbf{u}_{\leq n})||p_\theta(\mathbf{z}^c)) \\
- D_{KL}(q_\phi(\mathbf{z}^p|\mathbf{u}_{\leq n})||p_\theta(\mathbf{z}^p|\mathbf{u}_{<n-1})) \\
- \sum_{i=n-1}^n D_{KL}(q_\phi(\mathbf{z}^i|\mathbf{u}_{\leq i})||p_\theta(\mathbf{z}^i|\mathbf{u}_{<i}))
\end{aligned} \quad (21)$$

Equation 21 consists of two parts: the reconstruction term and KL divergence terms based on three kinds of latent variables.

## 3 Experiment

### 3.1 Experimental Settings

**Datasets:** We conduct our experiment on **Ubuntu Dialog Corpus** (Lowe et al., 2015) and **Cornell Movie Dialog Corpus** (Danescu-Niculescu-Mizil and Lee, 2011). As Cornell Movie Dialog does not provide a separate test set, we randomly split the corpus with the ratio 8:1:1. For each dataset, we keep conversations with more than 3 utterances. The number of multi-turn conversations in train/valid/test set is 898142/19560/18920 for Ubuntu Dialog, and 36004/4501/4501 for Cornell Movie Dialog.

**Hyper-parameters:** In our model and all baselines, Gated Recurrent Unit (GRU) (Cho et al., 2014) is selected as the fundamental cell in encoder and decoder layers, and the hidden dimension is 1,000. We set the word embedding dimension to 500, and all latent variables have a dimension of 100. For optimization, we use Adam (Kingma and Ba, 2015) with gradient clipping. The sentence padding length is set to 15, and the max conversation length is 10. In order to alleviate degeneration problem of variational framework (Bowman et al., 2016), we also apply KL annealing (Bowman et al., 2016) in all models with latent variables. The KL annealing steps are 15,000 for Cornell Movie Dialog and 250,000 for Ubuntu Dialog.

**Baseline Models:** We compare our model with three baselines. They all focus on multi-turn conversations, and the third one is a state-of-the-art variational model. 1) Hierarchical recurrent encoder-decoder (HRED) (Serban et al., 2016). 2) Variational HRED (VHRED) (Serban et al., 2017) with word drop (w.d) and KL annealing (Bowman et al., 2016), the word drop ratio equals to 0.25. 3) Variational Hierarchical Conversation RNN (VHCR) with utterance drop (u.d) (Park et al., 2018) and KL annealing, the utterance drop ratio equals to 0.25.

### 3.2 Evaluation Design

Open-domain response generation does not have a standard criterion for automatic evaluation, like BLEU (Papineni et al., 2002) for machine translation. Our model is designed to improve the co-

| Model | Average | Extrema | Greedy | Dist-1 | Dist-2 | Coherence | Fluency | Informativeness |
|---|---|---|---|---|---|---|---|---|
| **Ubuntu Dialog** | | | | | | | | |
| HRED | 0.570 | 0.329 | 0.415 | 0.494 | 0.814 | 2.96 | 3.64 | 2.89 |
| VHRED+w.d | 0.556 | 0.312 | 0.405 | 0.523 | 0.856 | 2.52 | 3.35 | 3.24 |
| VHCR+u.d | 0.572 | 0.330 | 0.416 | 0.512 | 0.837 | 2.42 | 3.48 | 2.99 |
| CSRR | **0.612** | **0.345** | **0.457** | **0.561** | **0.882** | **3.39** | **3.91** | **3.75** |
| **Cornell Movie Dialog** | | | | | | | | |
| HRED | 0.547 | 0.370 | 0.387 | 0.489 | 0.801 | 3.02 | 3.65 | 2.85 |
| VHRED+w.d | 0.556 | 0.365 | 0.405 | 0.512 | 0.850 | 3.05 | 3.76 | 3.24 |
| VHCR+u.d | 0.587 | 0.378 | 0.434 | 0.507 | 0.837 | 3.13 | 3.73 | 3.06 |
| CSRR | **0.620** | **0.395** | **0.462** | **0.522** | **0.873** | **3.43** | **3.82** | **3.78** |

Table 2: Automatic and human evaluation results on Ubuntu Dialog Corpus and Cornell Movie Dialog Corpus.

herence/relevance and diversity of generated responses. To measure the performance effectively, we use 5 automatic evaluation metrics along with human evaluation.

**Average, Greedy and Extrema:** Rather than calculating the token-level or n-gram similarity as the perplexity and BLEU, these three metrics are embedding-based and measure the semantic similarity between the words in the generated response and the ground truth (Serban et al., 2017; Liu et al., 2016). We use word2vec embeddings trained on the Google News Corpus [1] in this section. Please refer to Serban et al. (2017) for more details.

**Dist-1 and Dist-2:** Following the work of Li et al. (2016), we apply *Distinct* to report the degree of diversity. *Dist-1/2* is defined as the ratio of unique uni/bi-grams over all uni/bi-grams in generated responses.

**Human Evaluation:** Since automatic evaluation results may not be fully consistent with human judgements (Liu et al., 2016), human evaluation is necessary. Inspired by Luo et al. (2018), we use following three criteria. **Fluency** measures whether the generated responses have grammatical errors. **Coherence** denotes the semantic consistency and relevance between a response and its context. **Informativeness** indicates whether the response is meaningful and good at word usage. A general reply should have the lowest *Informativeness* score. Each of these measurement scores ranges from 1 to 5. We randomly sample 100 examples from test set and generate total 400 responses using models mentioned above. All generated responses are scored by 7 annotators, who are postgraduate students and not involved in other parts of the experiment.

[1] https://code.google.com/archive/p/word2vec/

### 3.3 Results of Automatic Evaluation

The left part of Table 2 is about automatic evaluation on test set. The proposed CSRR model significantly outperforms other baselines on three embedding-based metrics on both datasets. The improvement of our model indicates our semantic relationship modeling better reflects the structure of real-world conversations, and the responses generated by our models are more relevant to context. As for diversity, CSRR also gets the highest Dist-1 and Dist-2 scores.

For Ubuntu Dialog dataset, VHRED+w.d is the worst. With the help of discourse-level latent variable and utterance drop, VHCR+u.d leads to better performance. However, HRED is the worst on the Cornell Movie dataset. Park et al. (2018) empirically explained the difference based on that Cornell Movie Dialog dataset is small in size, but very diverse and complex in content and style, and models like HRED often fail to generate appropriate responses for the context.

### 3.4 Results of Human Evaluation

The right part of Table 2 is about human evaluation results on 400 (100×4) responses. First, it is clear that CSRR model receives the best evaluation on three aspects, which proves the effectiveness of CSRR on generating high quality responses. Second, because of the existence of discourse-level and pair-level latent variables, responses are more coherent. Since these two kinds of variables learn high level semantic information, utterance-level ones serve better on expression diversion, also improve sentence fluency and informativeness.

### 3.5 Case Study and Ablation Study

Table 3 shows the examples generated by CSRR model and other baseline models. For some

| Example 1: |
| --- |
| **Context**: how's sleeping beauty? $\mapsto$ still sleeping. $\mapsto$ kiss him, maybe he'll turn into a frog. |
| **HRED**: he's a good man. <br> **VHRED+w.d**: what? <br> **VHCR+u.d**: maybe he's dead. <br> **CSRR**: Really? I won't try. |
| Example 2: |
| **Context**: tried installing the driver from nvidia, but funny thing is i've installed $\mapsto$ what is the problem you are having exactly $\mapsto$ at this point driver activated but not in use |
| **HRED**: what card do you have? <br> **VHRED+w.d**: are you using the nvidia driver? or from the command line? <br> **VHCR+u.d**: i have no idea what you are talking about, i just know that <br> **CSRR**: you need to install the nvidia driver from the nvidia site. |
| Example 3: |
| **Context**: hi wondering whether to use 32 or 64 bit $\mapsto$ how much ram does the pc have? more than 2 gb you might $\mapsto$ definitely? because, i' ve heard that ubuntu recommends 32 bit and |
| **HRED**: it's a bit more than a 32 bit version of a 32bit system. <br> **VHRED+w.d**: i have no idea , but it doesn't matter. <br> **VHCR+u.d**: i don't know, but i haven't used it much. <br> **CSRR**: yes, but you can use 64 bit or 32 bit. |
| Example 4: |
| **Context**: how can i narrow down an issue i'm having? i am running $\mapsto$ check out the system logs in var logs ... in particular kern.log, syslog $\mapsto$ thank you very much. |
| **HRED**: no problem. <br> **VHRED+w.d**: i'm glad to hear that, don't know why. <br> **VHCR+u.d**: i'm not sure what you mean by ' <unk> '. <br> **CSRR**: you're welcome. i don't know if that's the case. |

Table 3: Examples generated by CSRR model and other baselines. The first example is from Cornell Movie Dialog, while the bottom three rows are from Ubuntu Dialog.

easy questions, like greeting (Example 4), both HRED and CSRR perform well. In contrast, VHRED+w.d and VHCR+u.d tend to generate general and meaningless responses. For hard questions, like some technical ones (Example 1 to 3), the proposed CSRR obviously outperforms other baselines. Note that VHCR is to show the effectiveness of $z^c$ and it can also be considered as the ablation study of CSRR to illustrate the validity of $z^p$. From above cases, we empirically find that with the help of $z^p$, response generated by CSRR are not only relevant and consistent to context, but also informative and meaningful.

## 4 Conclusion and Future Work

In this work, we propose a Conversational Semantic Relationship RNN model to learn the semantic dependency in multi-turn conversations. We ap-

ply hierarchical strategy to obtain context information, and add three-hierarchy latent variables to capture semantic relationship. According to automatic evaluation and human evaluation, our model significantly improves the quality of generated responses, especially in coherence, sentence fluency and language diversity.

In the future, we will model the semantic relationship in previous turns, and also import reinforcement learning to control the process of topic changes.

## References

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

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. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, pages 2980–2988.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *the 3rd International Conference on Learning Representations*.

Diederik P Kingma and Max Welling. 2014. Autoencoding variational bayes. In *the 2nd International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Ryan Lowe, Nissan Pow, Iulian V Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Liangchen Luo, Jingjing Xu, Junyang Lin, Qi Zeng, and Xu Sun. 2018. An auto-encoder matching model for learning utterance-level semantic dependency in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 702–707.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(Volume 1: Long Papers)*, pages 1792–1801.

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 *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3776–3783.

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 *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3295–3301.

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5610–5617.

Weinan Zhang, Yiming Cui, Yifa Wang, Qingfu Zhu, Lingzhi Li, Lianqiang Zhou, and Ting Liu. 2018. Context-sensitive generation of open-domain conversational responses. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2437–2447.