# Towards Less Generic Responses in Neural Conversation Models: A Statistical Re-weighting Method

**Yahui Liu**[1]*, **Victoria Bi**[2]†, **Jun Gao**[3], **Xiaojiang Liu**[2], **Jian Yao**[1], **Shuming Shi**[2]
[1]Wuhan University, Wuhan, China
[2]Tencent AI Lab, Shenzhen, China
[3]Soochow University, Suzhou, China
{liuyahui, jian.yao}@whu.edu.cn, imgaojun@gmail.com,
{victoriabi, kieranliu, shumingshi}@tencent.com

## Abstract

Sequence-to-sequence neural generation models have achieved promising performance on short text conversation tasks. However, they tend to generate generic/dull responses, leading to unsatisfying dialogue experience. We observe that in conversation tasks, each query could have multiple responses, which forms a *1-to-n* or *m-to-n* relationship in the view of the total corpus. The objective function used in standard sequence-to-sequence models will be dominated by loss terms with generic patterns. Inspired by this observation, we introduce a statistical re-weighting method that assigns different weights for the multiple responses of the same query, and trains the standard neural generation model with the weights. Experimental results on a large Chinese dialogue corpus show that our method improves the acceptance rate of generated responses compared with several baseline models and significantly reduces the number of generated generic responses.

## 1 Introduction

Many recent works have been proposed to use neural networks to generate responses for open-domain dialogue systems (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015; Li et al., 2016a,c; Serban et al., 2017; Shen et al., 2017; Li et al., 2017; Yu et al., 2017; Xu et al., 2017). These methods are inspired by the sequence-to-sequence (Seq2Seq) framework (Sutskever et al., 2014), which is originally applied for Neural Machine Translation (NMT). They aim at maximizing the probability of generating a response given an input query, and generally use the *maximum likelihood estimation* (MLE) as their objective function. However, various problems occur when Seq2Seq models are used for dialogue generation tasks. One of the most important problems is that such models are inclined to generate generic and dull responses (e.g., I don't know), rather than meaningful and specific answers (Sordoni et al., 2015; Serban et al., 2016; Li et al., 2016a,c; Kannan et al., 2016; Li et al., 2017; Xie, 2017; Wei et al., 2017; Mou et al., 2017).

Until now, it has attracted increasing studies to address the issue of generating generic response. For example, Li et al. (2016a) used the mutual information theory to reconstruct MLE, but this model is easy to generate ungrammatical outputs. They further proposed a fast diverse decoding approach (Li et al., 2016b), which modifies the beam search to re-rank meaningful responses into higher positions. Similar works explore different ways to encourage response diversity for picking less generic responses in the decoding search (Vijayakumar et al., 2016; Li and Jurafsky, 2016). In the reinforcement learning framework (Li et al., 2016c), the reward function used in the decoding considers the ease of answering, which is measured by a distance towards a set of 8 generic responses. Thus, it can also alleviate the problem of generating generic responses to some extent. Lison and Bibauw (2017) proposed to add a weighting model to learn the "quality" of the query and response pair, but it relies heavily on additional inputs. All these works tried to add extra optimized terms in the encoding or decoding modules in Seq2Seq, making the training or prediction more complicated.

In this work, we consider the reason why Seq2Seq often generates generic responses by analyzing the MLE objective function directly. We notice that multiple responses are often associated with one single input query. As shown in Figure 1, the relationship between queries and responses is much looser in conversation models than that in NMT, since the space of possible responses is much

*This work was done while Yahui Liu was with Tencent AI Lab.
†Corresponding author


*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2769–2774
Brussels, Belgium, October 31 - November 4, 2018. 

larger than the space of possible translations for a given sentence. On one hand, the information of these responses is only required to be relevant to the input query but usually differs from it. On the other hand, a query accepts large semantic diversity among its responses. Hence, it is a *1*-to-*n* relationship between a query and its responses (Vinyals and Le, 2015; Zhou et al., 2017). Meanwhile, we can see there is a *m*-to-*n* relationship between all queries and responses in the training corpus. Then, we find that MLE, which learns a *1*-to-*1* mapping in response generation, naturally puts more emphasis on optimizing the frequent patterns. Thus, the converged local optimum is easy to output these patterns or their combinations, leading to generic responses.

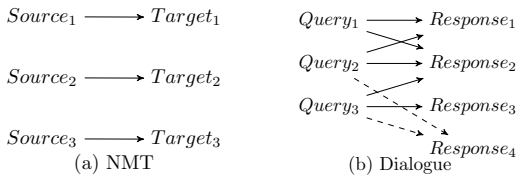


Figure 1: An illustration of the differences between NMT and dialogue generation. *Response*$_4$ is the potential cases that are not collected in corpus.

Inspired by this observation, we propose a statistical re-weighting method which modifies MLE by re-weighting the multiple responses for each query such that MLE will not be dominated by the frequent patterns or their combinations. The proposed method calculates the weights of a response with the consideration of two statistical features: similarity frequency and sentence length. Our model is simple and efficient to optimize without adding additional terms into the original Seq2Seq objective function. We validate the performance of our proposed method on a large Chinese dialogue corpus. Results show that it can improve the acceptance rate of the generated responses and significantly suppress the number of generic responses.

## 2 Proposed Method

Standard Seq2Seq models for NMT and dialogue generation aim at estimating the conditional probability $p(\mathbf{y}|\mathbf{x})$ where $\mathbf{x} = (x_1, \ldots, x_T)$ is an input sequence and $\mathbf{y} = (y_1, \ldots, y_{T'})$ is its corresponding output sequence whose length $T'$ may differ from $T$. During training, we learn all the model parameters $\boldsymbol{\theta}$ by summing the negative log likelihood of each sample pair $(\mathbf{x}, \mathbf{y})$ in the training corpus $\mathbb{C}$:

$$\ell(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = -\sum_{t=1}^{T'} \log p(y_t|\mathbf{x}, \mathbf{y}_{[t-1]}; \boldsymbol{\theta}), \quad (1)$$

$$\mathcal{L}(\mathbb{C}, \boldsymbol{\theta}) = \sum_{(\mathbf{x},\mathbf{y})\in\mathbb{C}} \ell(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}). \quad (2)$$

Recall that generic responses are those that are safe and universal for many queries and thus frequently appear in the training corpus. Hence, if we have two responses of $\mathbf{x}$ in which one is generic and the other one contains more meaningful content, using $\mathcal{L}(\mathbb{C}, \boldsymbol{\theta})$ in Eq. 1 will put the same emphasis on optimizing each of their loss terms. Therefore, $\mathcal{L}(\mathbb{C}, \boldsymbol{\theta})$ contains a large amount of patterns from the generic responses, thus it is not surprised to see that the trained models are stuck into local optimum that are inclined to generate these patterns or their combinations.

Based on this observation, we argue that a good loss function of Seq2Seq for dialogue generation should not be dominated by the patterns from generic responses. Here, we propose a re-weighting method for responses of a query $\mathbf{x}$. Specifically, $\ell(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$ in Eq. 1 is modified to be:

$$\ell_w(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = w(\mathbf{y}|\mathbf{x})\ell(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}), \quad (3)$$

where $w(\mathbf{y}|\mathbf{x}) \in (0, 1]$ is a soft weight for a response $\mathbf{y}$ of a query $\mathbf{x}$. In the implementation, we make the normalization of this loss at the mini-batch level for better computational efficiency. Hence, the loss of Eq. 2 for a mini-batch $\mathcal{L}(\mathbb{B}, \boldsymbol{\theta})$ takes the form:

$$\mathcal{L}(\mathbb{B}, \boldsymbol{\theta}) = \frac{\sum_{\mathbf{x},\mathbf{y}\in\mathbb{B}} \ell_w(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})}{\sum_{\mathbf{x},\mathbf{y}\in\mathbb{B}} w(\mathbf{y}|\mathbf{x})}. \quad (4)$$

We summarize two common properties for the responses:

- Responses with the patterns of frequently appearing in the training corpus tend to be generic. Here, the patterns refer to both the whole sentence or n-grams which can be described by similarities among responses.

- Very short and long responses should be avoid. Owing to the MLE objective function, the Seq2Seq frameworks are inclined to generate short responses that are universal replies. While long responses usually contain more specific information which may not be generalized to most conversation scenarios. Hence,

high-quality responses tend to be with moderate length.

We propose an estimator by considering these two properties:

$$w(\mathbf{y}|\mathbf{x},\mathbb{R},\mathbb{C}) = \frac{\Phi(\mathbf{y})}{\max_{\mathbf{r}\in\mathbb{R}}\{\Phi(\mathbf{r})\}}, \quad (5)$$

where $\mathbb{R}$ denotes all collected responses of $\mathbf{x}$ in $\mathbb{C}$. For each response, the estimator gives a weight by:

$$\Phi(\mathbf{y}) = \alpha\mathcal{E}(\mathbf{y}) + \beta\mathcal{F}(\mathbf{y}). \quad (6)$$

Here, $\mathcal{E}(\mathbf{y})$ and $\mathcal{F}(\mathbf{y})$ correspond to the mentioned two properties respectively:

- $\mathcal{E}(\mathbf{y}) = e^{-af(\mathbf{y})}$, where $f(\mathbf{y})$ is a function related to the frequency of response $\mathbf{y}$. It could be formulated as

$$f(\mathbf{y}) = \max\{0, \text{Count}(D(\mathbf{y},\mathbf{y}_j) \geq \tau) - b\} \quad \forall j \in |\mathbb{C}|,$$

  where $D(\cdot)$ refers to the similarity between two sentences, $a$ is a scale factor, $b$ is bias and $\tau \in [0,1]$ is a threshold specifying the similarity that two responses will be considered identical. For instance, it could be the simplest strictly matching, which is used in our experiments. Other methods like cosine distance of TF-IDF (token or n-grams) can also be applied, but may encounter computational issues for large corpus. A response with a higher frequency will be assigned with a smaller $\mathcal{E}(\mathbf{y})$.

- $\mathcal{F}(\mathbf{y}) = e^{-c||\mathbf{y}|-|\hat{\mathbf{y}}||}$, where $|\mathbf{y}|$ denotes the number of tokens in $\mathbf{y}$, $|\hat{\mathbf{y}}| = \frac{1}{|\mathbb{C}|}\sum_{\mathbf{r}\in\mathbb{C}}|\mathbf{r}|$ refers to the average length of responses in the total training corpus, and $c$ is a scale factor. Here, the "moderate length" is set to the average length of responses of the total training corpus. In practice, we have tried to use long responses (longer than average length) to fine-tune the Seq2Seq model. Though it slightly increases the average length of generated responses, the generated responses suffer from more ungrammatical and influent issues. Hence, if a response is too short or long, it will receive a low score of $\mathcal{F}(\mathbf{y})$.

Mentioned hyper-parameters $\{\alpha, \beta, a, b, \tau, c\}$ are constant values in the following experiments, which are set to $\{0.5, 0.5, 0.33, 3, 1.0, 0.33\}$. When we performed our experiments, we tried several hyper-parameter settings and found that our method is not sensitive to different hyper-parameters and achieves stable results in general. Hence, we do not spend many efforts to specifically tune these hyper-parameters.

| Response (frequency/length) | $w(\mathbf{y}\|\mathbf{x})$ | $\mathcal{E}(\mathbf{y})$ | $\mathcal{F}(\mathbf{y})$ |
|---|---|---|---|
| 一直单身 (70/2)<br>Single till now. | 0.047 | 0.014 | 0.066 |
| 求交往 (173/2)<br>Would you like to have a date with me? | 0.039 | 0.000 | 0.066 |
| 等你看到别人甜蜜就知道一个人的痛苦了 (2/13)<br>Once seeing the happy lovers, you will feel sad. | 0.769 | 1.000 | 0.560 |
| 应该还在哭吧 (1/5)<br>You may be crying now. | 0.665 | 1.000 | 0.176 |
| 单身狗的自我安慰 (2/4)<br>Morale-boosting of singles. | 0.622 | 1.000 | 0.128 |
| 我也是这么觉得 (30/6)<br>I think so. | 0.149 | 0.052 | 0.248 |
| 单身不好,单身23年,孤单太久了 (1/10)<br>Given being single for 23 years, it's not so good. | 1.000 | 1.000 | 0.924 |
| 多么心塞的领悟 (1/4)<br>What a painful understanding! | 0.623 | 1.000 | 0.128 |

Table 1: Weights of the responses for a query "其实单身也挺好的 (It's pretty good to be single)".

To validate that our design function in Eq. 5 and Eq. 6 are effective to weight the responses, Table 1 shows the weights of 8 responses for a query "其实单身也挺好的 (It's pretty good to be single)". As can be seen, the weights are reasonable, in which the higher-ranked responses are more informative ones with low similarity frequency and moderate length.

## 3 Experiments

### 3.1 Corpus and Evaluation

We crawl conversation pairs from some popular Chinese social media websites[1], and select 7M high-quality pairs as our training corpus. Conventional metrics such as BLEU (Papineni et al., 2002) and perplexity, are improper to be used for response generation tasks. Following previous works (Li et al., 2016c, 2017), we apply human annotations. We randomly sample 500 queries (not used in training) as our test samples, and recruit 3 annotators to evaluate each generated response from two aspects:

- *Fluency*: 0 (unreadable), 1 (readable but with some grammar mistakes), 2 (fluent);

- *Relevance*: 0 (not relevant at all), 1 (relevant at a distant level), 2 (relevant, including the

[1]Weibo: www.weibo.com, Baidu Tieba: tieba.baidu.com, and Zhihu: www.zhihu.com

generic responses), 3 (relevant as well as interesting).

*Acceptance* is then automatically calculated as a metric reflecting whether the response is acceptable to real users. A response will be assigned 1 when it gets *Fluency*$\geq$1 and *Relevance*$\geq$2, otherwise it will be assigned 0.

We implement our baseline Seq2Seq model using its standard objective function in Eq. 1 with two LSTM layers for encoding/decoding and a standard beam search with a beam size of 5 (the best setting), termed as *Seq2Seq*. We also compare several Seq2Seq variants:

- *Seq2Seq-RS*: training with a subset by randomly sampling only one from the multiple responses for each query;
- *Seq2Seq-MMI*: applying the maximum mutual information (Li et al., 2016a) (only the MMI-bidi);
- *Seq2Seq-DD*: applying the diverse decoding algorithm (Li et al., 2016b);
- *Ours-RW*: calculating weights via our re-weighting method proposed in Section 2. Without applying any other tricks, we implement three versions of our method by using $\mathcal{E}(\cdot)$ only, $\mathcal{F}(\cdot)$ only, a linear combination of $\mathcal{E}(\cdot)$ and $\mathcal{F}(\cdot)$ in Eq 6, termed as *Ours-RW*$_{\{\text{E,F,EF}\}}$.

### 3.2 Results and Discussion

Human annotation results are shown in Table 2. Several observations can be made. First, *Seq2Seq-RS* performs slightly worse than the baseline model. This means that it does not work to simply discard a large amount of training data to construct a 1-to-1 query-response subset for training. Second, *Seq2Seq-MMI* not only provides no improvement for the baseline but also inclines to generate generic response. Third, *Seq2Seq-DD* obtains higher relevance and acceptance scores than the baseline, which shows its effectiveness by re-ranking more meaningful responses into higher positions in beam search. Fourth, our method achieves the best performance on almost all metrics. When we use strictly matched frequency of each response, *Ours-RW*$_{\text{E}}$ does not perform better than the baseline model because that the percentage of responses with frequency higher than 3 is about 0.5% in our training corpus. However, it still enhances the performance in *Ours-RW*$_{\text{EF}}$, which performs the best and increases the acceptance of the baseline model from 0.42 to 0.55. This validates that the properties about similarity frequency and sentence length play important roles in generating better responses.

| Model | Evaluation Metrics | | |
|---|---|---|---|
| | Fluency | Relevance | Acceptance |
| Seq2Seq | 1.96±3.8e-5 | 1.31±5.3e-3 | 0.42±4.7e-4 |
| Seq2Seq-RS | **1.97**±8.1e-5 | 1.30±2.1e-3 | 0.42±9.9e-4 |
| Seq2Seq-MMI | 1.94±7.2e-5 | 1.19±4.0e-3 | 0.41±1.9e-4 |
| Seq2Seq-DD | 1.86±2.8e-3 | 1.40±1.5e-2 | 0.49±2.4e-3 |
| Ours-RW$_{\text{E}}$ | 1.95±1.9e-4 | 1.30±5.1e-3 | 0.42±4.2e-4 |
| Ours-RW$_{\text{F}}$ | **1.97**±1.5e-4 | 1.47±2.1e-3 | 0.51±6.3e-4 |
| Ours-RW$_{\text{EF}}$ | 1.96±8.3e-5 | **1.59**±1.9e-2 | **0.55**±4.4e-3 |

Table 2: Human annotation results.

Specifically, the average percentage of the generated responses that are assigned to relevance rating 2 (relevant, including the generic responses) and 3 (relevant as well as interesting) are presented in Table 4. It shows that our method achieves higher relevance score owing to generating more high-quality responses with rating 3.

To validate that our method is effective to reduce the number of generated generic responses, we calculated the *distinct-1* and *distinct-2* (Li et al., 2016a) for the compared methods respectively, which are the number of distinct unigrams and bigrams divided by total number of generated words respectively. As shown in Table 5, *Ours-RW*$_{\text{EF}}$ achieves the best performance on the two metrics. This indicates that our model often outputs more meaningful and relevant responses than the other compared methods.

We further randomly sample another 100K queries (not used in training) and use the various models to generate responses. We compare the frequencies of several common generic responses appearing in the generated results, as shown in Table 6. It shows that our method can significantly reduce the number of generic responses. For instance, we reduce about 75% of the case "我也不知道 (I don't know, either.)" and 77% of the case "我也想知道 (I want to know, too)" to be generated.

## 4 Conclusion

In this paper, we propose a statistical re-weighting method to weight multiple responses differently and optimize the MLE objective function. The weight of each response is calculated based on

| Query | Response | | | |
|---|---|---|---|---|
| | Seq2Seq | Seq2Seq-MMI | Seq2Seq-DD | Ours-RW$_{EF}$ |
| 我要生气了<br>I'm going to be angry. | 我也生气了<br>So do I. | 生气了<br>I'm angry. | 我也生气了<br>So do I. | 不要生气<br>Don't be angry. |
| 最近有什么好书推荐<br>Would you like to recommend me some good books? | 看你喜欢什么类型的<br>It depends on what kind of books you like? | 我也想去看电影<br>I want to see the movie, too. | 你喜欢什么类型的<br>What kind of books do you like? | 《红楼梦》<br>Dream of the Red Chamber |
| 做人不一定要风风光光<br>It's not necessary to be a successful and grand man. | 一定要<br>Must be | 我也不知道<br>I don't know, either. | 是啊。。。<br>Yes... | 做人一定要善良<br>Everyone should try to be a kind man. |
| 这几天嗓子疼<br>I have a sore throat recently. | 我也不知道<br>I don't know, either. | 我也想去<br>I want to go, too | 我也是。。。<br>So do I... | 我这两天也是<br>So do I these days . |
| 二号煤矿放几天假<br>How many days did people in No.2 coal mine take off? | 假的<br>False | 我也想放假<br>I want to have a holiday, too | 假的，假的<br>False, False | 我放了3天<br>I took three days off. |
| 孟记普洱茶专营店一帖<br>An advertising of Mengji's Pu'er tea shop. | 我也想吃<br>I want to eat, too. | 我也想喝<br>I want to drink, too. | 铁观音，茶，茶，茶<br>Tat-Kuan-Yin Tea, tea, tea, tea | 我喜欢喝茶<br>I like drinking tea. |

Table 3: Examples of responses generated by Seq2Seq, Seq2Seq-MMI, Seq2Seq-DD and Ours-RW$_{EF}$.

| Model | rating 2(%) | rating 3(%) |
|---|---|---|
| Seq2Seq | 31.3 | 10.6 |
| Seq2Seq-MMI | 35.8 | 4.4 |
| Seq2Seq-DD | **37.1** | 11.7 |
| Ours-RW$_{EF}$ | 36.8 | **18.1** |

Table 4: Comparisons on the average percentage of the generated responses that are assigned to relevance rating 2 and 3.

| Model | distinct-1 | distinct-2 |
|---|---|---|
| Seq2Seq | 0.170 | 0.307 |
| Seq2Seq-MMI | 0.140 | 0.259 |
| Seq2Seq-DD | 0.131 | 0.170 |
| Ours-RW$_{EF}$ | **0.173** | **0.359** |

Table 5: Performances on the metrics *distinct-1* and *distinct-2*.

two terms according to the similarity frequency and its length. Experiments show that our approach improves the performance over the baseline models and reduces the number of generated generic responses significantly. It indicates that mismatching issue of objective function can be alleviated through such similar re-weighting methods, by which current encoder-decoder architectures can take full use of the *m*-to-*n* training corpus and model the dialogue generation tasks better.

| Response | Model | | | |
|---|---|---|---|---|
| | Seq2Seq | MMI | DD | RW$_{EF}$ |
| 我也不知道<br>I don't know, either. | 3296 | 4404 | 941 | 847 |
| 我也想知道<br>I want to know, too. | 3211 | 13764 | 795 | 738 |
| 我也觉得<br>I think so. | 883 | 632 | 1505 | 59 |
| 我也想问<br>I want to ask, too. | 515 | 51 | 441 | 12 |
| 不喜欢<br>I don't like it. | 337 | 131 | 214 | 95 |
| 看什么<br>What do you look at? | 254 | 52 | 377 | 87 |
| 不会<br>Will/can not | 109 | 92 | 54 | 24 |

Table 6: Frequencies of several generic responses.

## References


Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. *The North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Conference on Empirical Methods in Neural Language Processing (EMNLP)*.

Pierre Lison and Serge Bibauw. 2017. Not all dialogues are created equal: instance weighing for neu-

ral conversational models. In *Conference on Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2017. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence (AAAI)*.

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

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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. *arXiv preprint arXiv:1705.00316*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

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

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2017. Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. *arXiv preprint arXiv:1712.02250*.

Ziang Xie. 2017. Neural text generation: A practical guide. *arXiv preprint arXiv:1711.09534*.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, Xiaolong Wang, Zhouran Wang, and Chao Qi. 2017. Neural response generation via gan with an approximate embedding layer. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *AAAI Conference on Artificial Intelligence (AAAI)*.