Promoting Diversity for End-to-End Conversation Response Generation

Yu-Ping Ruan[1], **Zhen-Hua Ling**[1], **Quan Liu**[1,2], **Jia-Chen Gu**[1], **Xiaodan Zhu**[3]
[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
[2]iFLYTEK Research, Hefei, P.R. China
[3]Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada
{ypruan,gujc}@mail.ustc.edu.cn, {zhling,quanliu}@ustc.edu.cn, xiaodan.zhu@queensu.ca

Abstract

We present our work on Track 2 in the Dialog System Technology Challenges 7 (DSTC7). The DSTC7-Track 2 aims to evaluate the response generation of fully data-driven conversation models in knowledge-grounded settings, which provides the contextual-relevant factual texts. The Sequence-to-Sequence models have been widely used for end-to-end generative conversation modelling and achieved impressive results. However, they tend to output dull and repeated responses in previous studies. Our work aims to promote the diversity for end-to-end conversation response generation, which follows a two-stage pipeline: 1) Generate multiple responses. At this stage, two different models are proposed, i.e., a variational generative (VariGen) model and a retrieval based (Retrieval) model. 2) Rank and return the most related response by training a topic coherence discrimination (TCD) model for the ranking process. According to the official evaluation results, our proposed Retrieval and VariGen systems ranked first and second respectively on objective diversity metrics, i.e., Entropy, among all participant systems. And the VariGen system ranked second on NIST and METEOR metrics.

1 Introduction

Natural language conversation occupies one of the most challenging tasks in artificial intelligence (AI) and natural language processing (NLP), which can be categorized into task oriented dialog systems (Young et al. 2013) and non-task oriented chatbots. Previous studies in building dialog systems mainly focused on either rule-based or learning-based methods (Schatzmann et al. 2006; Young et al. 2013; Bordes and Weston 2016; Wen et al. 2016). These methods often require manual efforts in rule designing and feature engineering, which makes it difficult to develop an extensible open domain conversation system.

Recently due to the explosive growth of social media, there comes vast amount of conversation text available on the web. This makes the data-driven approach to settle the conversation problem possible. Thus, there has been a growing interest in applying encoder-decoder models (Sutskever, Vinyals, and Le 2014; Vinyals and Le 2015; Shang, Lu, and Li 2015; Serban et al. 2016) for conversation in a completely end-to-end data-driven fashion, which have produced impressive results.

The Dialog System Technology Challenges (DSTC) in its seventh edition offers a track (Track 2) (Galley et al. 2018) devoted exclusively to fully data-driven approaches for conversation modelling. Different from previous data-driven dialog systems which mostly focused on chitchat, DSTC7-Track 2 tries to push the data-driven conversation models beyond chitchat in order to produce system responses that are both substantive and "useful" which can contain factual contents. So DSTC7-Track 2 provides not only the social conversation corpus but also contextual-related factual texts to build a knowledge-grounded conversation settings.

Among those neural end-to-end models for data-driven conversation modelling, the encoder-decoder framework has been widely adopted and they principally learn the mapping from an input conversation context x to its target response y. However, previous studies on generating responses for chit-chat conversations (Serban et al. 2016; Li et al. 2015) have found that ordinary encoder-decoder models tend to generate dull, repeated and generic responses in conversations, such as *"i don't know", "that's ok"*, which are lack of diversity. One possible reason is the deterministic calculation of ordinary encoder-decoder models which constrains them from learning the 1-to-n mapping relationship, especially on semantic connections, between input sequence and potential multiple target sequences, which is common in social media conversation corpus.

To cope with the problem of lack diversity, we built our systems following a two-stage processing pipeline: 1) The first module (module-1) outputs multiple candidate responses according to input context and related textual facts. 2) The second module (module-2) ranks the output responses from module-1 and returns the most topic relevant response. Two different models are designed for module-1, i.e., one Variational generative model and one retrieval based model. Variational generative models are suitable for learning the 1-to-n mapping relationship due to their variational sampling mechanism for deriving latent representations, and recently they have been applied to dialog response generation (Zhao, Zhao, and Eskenazi 2017; Serban et al. 2017). For module-2, a topic coherence discrimination (TCD) model is designed and trained based on ESIM (Chen et al. 2017) for the ranking process. According to the official evaluation results,

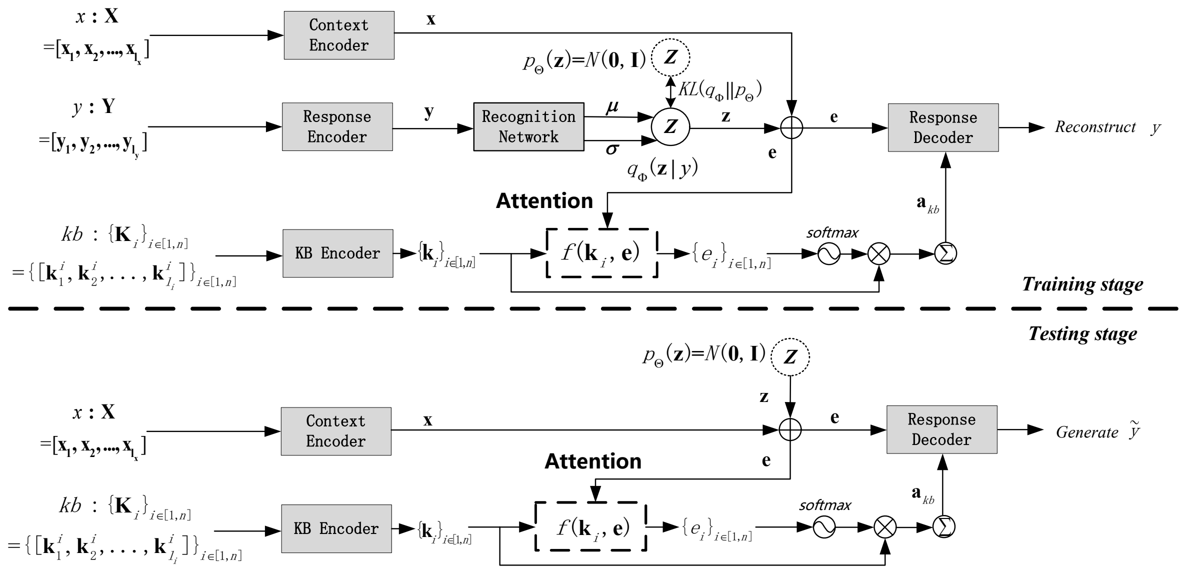

Figure 1: The model architecture of the variational generative model (VariGen) implemented in this paper. $\bigoplus$ denotes the concatenation of input vectors. All the encoders and decoders are 1-layer GRU-RNNs, the recognition network is a two-layer MLP.

our proposed Retrieval and VariGen system ranked first and second respectively on objective diversity metrics, i.e., Entropy, among all participant systems, and the VariGen system ranked second on NIST and METEOR metrics.

2 System Description

As we have described in Section 1, we designed our systems following a two-stage processing pipeline: 1) The first module (module-1) outputs multiple possible responses according to input context and related textual facts. 2) The second module (module-2) ranks the output responses from module-1 and return the most topic relevant response. Here, we will describe our models in detail.

Candidate response generation

At this stage, a generative model based on Variational AutoEncoder (VAE) and a retrieval model based on Bag-of-Words (BoW) representation are designed as module-1 in order to output multiple responses with diversity.

Variational generative model (VariGen) The model architecture of the variational generative model (VariGen) implemented in this paper is shown in Figure 1. For an input conversation context $x = [x_1, x_2, ..., x_{l_x}]$ with l_x words, we can derive the corresponding output hidden states $[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{l_x}]$ by sending its word embedding sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{l_x}]$ into the *Context Encoder*. Then, the mean pooling of hidden states $[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{l_x}]$ is used to present the input context, denoted as $\mathbf{x}$. Similarly, for its correlated textual facts $kb = [k_1, k_2, ..., k_n]$ with n textual facts in it, and each k_i in kb is also a word sequence, i.e., $k_i = [k_1^i, k_2^i, ..., k_{l_i}^i]$ with l_i words, we can derive vector representation $\mathbf{k_i}$ for each k_i by inputting its word embedding sequence $\mathbf{K}_i = [\mathbf{k}_1^i, \mathbf{k}_2^i, ..., \mathbf{k}_{l_i}^i]$ into the *KB Encoder*. Also, we can derive vector representation $\mathbf{y}$ for response y by inputting $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{l_y}]$ into the *Response Encoder*.

The *Recognition Network* is a multi-layer perceptron (MLP), which has a hidden layer with *softplus* activation and a linear output layer in our implementation. The recognition network predicts $\boldsymbol{\mu}$ and $\log(\boldsymbol{\sigma}^2)$ from $\mathbf{y}$, which gives $q_\phi(\mathbf{z}|y) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$. Then samples of latent variable $\mathbf{z}$ are generated from the $q_\phi(\mathbf{z}|y)$ at training stage or directly from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ at testing stage. To guarantee the feasibility of error backpropagation for model training, reparametrization (Kingma and Welling 2013) is performed to generate the samples of $\mathbf{z}$.

Then we derive the encoding vector $\mathbf{e}$ by concatenating the $\mathbf{z}$ and $\mathbf{x}$, which combines the information of conversation context x and response y. And we use $\mathbf{e}$ to extract the key information in correlated kb by using attention on $[\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_n]$ as follows,

$$e_i = \mathbf{v}^T \tanh(\mathbf{W}[\mathbf{e}; \mathbf{k}_i]), \tag{1}$$

$$\mathbf{a}_{kb} = \sum_{i=1}^{n} \frac{\exp e_i}{\sum_{j=1}^{n} \exp e_j} \mathbf{k}_i, \tag{2}$$

where $\mathbf{a}_{kb}$ is the final attentive output vector from input kb. Finally, for the $ResponseDecoder$, its initial hidden state is $\mathbf{x}$. In each time step, its input is the concatenation of the word embedding from the previous time step, the encoding vector $\mathbf{e}$, and the attentive kb vector $\mathbf{a}_{kb}$.

Our VariGen model can be efficiently trained with the stochastic gradient variational Bayes (SGVB) (Kingma and Welling 2013) framework by maximizing the lower bound of the conditional log likelihood $\log p(y|x, kb)$ as follows,

$$\mathcal{L}(\theta, \phi; x, y, kb) = -KL(q_\phi(\mathbf{z}|y)||\mathcal{N}(\mathbf{0}, \mathbf{I})) + \mathbf{E}_{q_\phi(\mathbf{z}|y)}[\log p_\theta(y|x, \mathbf{z}, kb)] \leq \log p(y|x, kb). \quad (3)$$

in which the summation of the log-likelihood of reconstructing y from the $ResponseDecoder$ and the negative K-L divergence between $q_\phi(\mathbf{z}|y)$ and prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is used as the objective function for training.

When generating responses at testing stage, k samples of $\mathbf{z}$ are generated from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, for each $\mathbf{z}$ sample, a beam-search is adopted to return the top-1 result. The final k generated results will be input to TCD model for subsequent ranking.

Retrieval model The retrieval model we built in this work is mainly based on Bag-of-Word (BoW) representation. Specifically, for each input conversation context $x = [x_1, x_2, ..., x_{l_x}]$ with l_x words, we derive the vector representation $\mathbf{x}$ as a weighted sum of its word embedding sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{l_x}]$, in which each word's inverse-document-frequency (idf) is used as its sum weight and the GloVe embeddings (Pennington, Socher, and Manning 2014) are used. To cope with variable length of x, we normalize the final weighted BoW vector $\mathbf{x}$ with the sum of all corresponding idf weights. Then we treat input conversation context x_{query} as a query and retrieve related conversation context x_{pool} in training pool according to cosine similarity between weighted BoW representation $\mathbf{x}_{query}$ and $\mathbf{x}_{pool}$. We do not use the kb in our retrieval model. Finally, we return corresponding responses of top-k related x_{pool} as outputs.

Response reranking

The multiple candidate responses output from module-1 are reranked using a topic coherence discrimination (TCD) model, which is designed and trained based on the ESIM model (Chen et al. 2017). Specifically, we replace all BiLSTMs in the ESIM with 1-layer LSTMs and define the objective of the TCD model as judging whether a response is a valid response to a given conversation context. In order to train the TCD model, all context-response pairs in the training set are used as positive samples and negative samples are constructed by randomly shuffling the mapping between contexts and responses. Finally, ranking scores are adopted to rerank all responses generated for one context. The scores are calculated as:

$$score = \log p_\theta(\tilde{y}|x, kb) + \lambda * \log p_{TCD}(true|x, \tilde{y}), \quad (4)$$

where the first term is the log-likelihood of generating response $\widetilde{y}$ using the decoder network in the VariGen models (when module-1 is the retrieval model, this term will be omitted). The second term is the log-likelihood of the output probability of the TCD model. λ represents the weight between these two terms.

3 Experiments

Datasets

The training, development, and final test datasets of DSTC7-Track 2 were built using the official scripts[1], we downloaded the formatted Reddit conversation data and its corresponding textual facts. We finally collected $1,918,146$, $108,600$, and $10,808$ samples for training, development and final test respectively.

To alleviate the problem of out-of-vocabulary (oov), we replaced all *date, time, email, numbers, link* and other data words with special tokens like $\langle data \rangle$, $\langle time \rangle$ and so on using *CommonRegex*[2] tool. We lowercased all words and filtered the *kb* with no more than 5 words. Finally, the corresponding *kb* of each *context-response* pair has averagely 152 items in it and each kb item has averagely 33 words. We set the vocabulary size as $51,996$ and all the oov rates of *contexts*, *responses*, and *kb* in training dataset are no more than 3%.

Parameter setting

We trained the VariGen model in our experiments with the following hyperparameters. All word embeddings, hidden layers of the recognition network and hidden state vectors of the encoders and decoders had 300 dimensions. The latent variables $\mathbf{z}$ had 100 dimensions. All encoder and decoder in VariGen shared one set of embeddings, and the vocabulary size was $51,996$. All model parameters were initialized randomly with Gaussian-distributed samples except for word embeddings were initialized with Glove embeddings (Pennington, Socher, and Manning 2014). The method of Adam(Kingma and Ba 2014) was adopted for optimization. The initial learning rate was $1e-04$. Gradient clipping was set to 1, and the batch size was 16. We generated multiple responses for each post, the number of $\mathbf{z}$ samples was set to 50, the beam search size was 5. When training the TCD model, we adopted the same parameter settings above for training VariGen model, and the weight λ for reranking was heuristically set to 10.

Baseline models

We listed three baseline systems provided by the official organizer of DSTC7-Track2 and their evaluation results will be used to help us analyse the performance of our proposed systems.

- *Constant:* This system always return "i don't know what you mean." as responses.
- *Random:* This system always randomly picks up a response from training dataset.
- *Seq2Seq:* This is a GRU-based sequence-to-sequence (Seq2Seq) generative system. The model does not use grounding information ("facts"), attention or beam search. It uses greedy decoding (unkown token disabled).

[1]https://github.com/DSTC-MSR-NLP/DSTC7-End-to-End-Conversation-Modeling

[2]https://github.com/madisonmay/CommonRegex

Models	nist1	nist2	nist3	nist4	bleu1	bleu2	bleu3	bleu4	Meteor
Constant	0.175	0.183	0.184	0.184	39.7%	12.8%	6.06%	2.87%	7.48%
Random	1.573	1.633	1.637	1.637	26.4%	6.7%	2.24%	0.86%	5.91%
Seq2Seq	0.849	0.910	0.915	0.916	45.2%	14.8%	5.23%	1.82%	6.96%
Retrieval	1.938	2.034	2.039	2.040	29.2%	8.2%	2.81%	1.05%	7.48%
VariGen	2.181	2.312	2.322	2.322	34.9%	10.6%	3.67%	1.21%	7.18%
Human	2.424	2.624	2.647	2.650	34.1%	12.4%	5.72%	3.13%	8.31%
*Best system performance**	2.341	2.510	2.522	2.523	41.3%	14.4%	5.01%	1.94%	8.07%

Table 1: The official *objective evaluation* results of three baseline systems, our two proposed systems, and human results on 2,208 samples from test set. The *Best system performance** represent the best results on each metric among all participant systems.

Models	entropy1	entropy2	entropy3	entropy4	div1	div2	avg_len
Constant	2.079	1.946	1.792	1.609	0.000	0.000	8.000
Random	6.493	9.670	10.403	10.467	0.160	0.647	19.192
Seq2Seq	3.783	5.017	5.595	5.962	0.014	0.048	10.604
Retrieval	6.360	9.374	10.009	10.057	0.108	0.449	22.336
VariGen	5.320	8.080	9.487	10.016	0.034	0.265	16.570
Human	6.589	9.742	10.410	10.445	0.167	0.670	18.757
*Best system performance**	6.360	9.374	10.009	10.057	0.121	0.449	—

Table 2: The official *objective diversity evaluation* results of three baseline systems, our two proposed systems, and human results on 2,208 samples from test set. The *Best system performance** represent the best results on each metric among all participant systems.

Models	**(A) Relevance**		**(B) Interest**		**(C) Overall**	
	Mean Score	95% CI	Mean Score	95% CI	Mean Score	95% CI
Constant	2.60	(2.560, 2.644)	2.32	(2.281, 2.364)	2.46	(2.424, 2.500)
Random	2.32	(2.269, 2.371)	2.35	(2.303, 2.401)	2.34	(2.288, 2.384)
Seq2Seq	2.91	(2.858, 2.963)	2.68	(2.632, 2.730)	2.80	(2.748, 2.844)
Retrieval	2.82	(2.771, 2.870)	2.57	(2.525, 2.619)	2.70	(2.650, 2.742)
VariGen	—	—	—	—	—	—
Human	3.61	(3.554, 3.658)	3.49	(3.434, 3.539)	3.55	(3.497, 3.596)

Table 3: The official *subjective evaluation* results of three baseline systems, our two proposed systems, and human results on 1,000 samples from test set.

Results and analysis

Objective evaluation As Table 1 shows, we present official evaluation results on NIST, BLEU, and METEOR metrics of three baseline systems, our *Retrieval* and *VariGen* systems, human results, and best system performance among all participants. These objective evaluations were performed on 2,208 samples from final test dataset. For BLEU metrics, we can find *Constant* system performed best on the whole among all systems though this system always output "i don't know what you mean.", which indicates the limitations of BLEU metrics. For METEOR metric, we can find that both our proposed *VariGen* and *Retrieval* systems outperformed *Seq2Seq* and *Random* systems. However, *Constant* system also obtained best performance among all systems except for *Human* and best system performance*, which also indicates the limitations of METEOR metric. NIST[3] is an variant of BLEU which instead of giving equal weight to each n-gram match but calculates how informative each particular n-gram is (Doddington 2002). We can find that our *Retrieval* and *VariGen* systems outperformed all three baseline systems, and *VariGen* performed better than *Retrieval* system. Also, we can find that *Random* system outperformed *Seq2Seq*, which may be attribute to that the *Seq2Seq* system always produces blank and non-sense responses with very low informativeness. According to the official evaluation results, our proposed *VariGen* system ranked second on automatic evaluation metrics NIST and METEOR among all participant systems.

As for the objective diversity evaluation, we present official evaluation results as Table 2 shows. The div1 and div2, a crude diversity measure of distinct 1-gram and 2-gram respectively as explained in (Li et al. 2015), are also known as distinct-1 and distinct-2. Entropy is believed to be a more principled measure of diversity as explained in (Zhang et al. 2018). We can found that our *Retrieval* and

[3] `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v14c.pl`

VariGen outperformed all baseline models on div1/2 and Entropy metrics except for the *Random* system. It's reasonably that *Random* system has a good performance on diversity metrics since it always randomly select a human response from training dataset. Our *Retrieval* system outperformed *VariGen*, but in Table 1 *VariGen* performed better than *Retrieval* on NIST and BLEU metrics. According to the official evaluation results, our *Retrieval* system ranked first on diversity metrics and *VariGen* system ranked second on Entropy metrics.

Subjective evaluations Table 3 presents the official subjective evaluation results which were performed on 1,000 samples from final test dataset. According to official organizer's description, three crowdsourced judges were asked to whether to select one of Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree in relation to the following statements: 1) The response is *relevant* and appropriate. 2)The response is *interesting* and informative. The final results were converted into a numerical score of 1 (Strongly Disagree) through 5 (Strongly Agree), and 95% confidence intervals were computed using 10,000 iterations of bootstrap sampling. Because official organizer can only evaluate one system for each participant, we have only our *Retrieval* system evaluated. From the results of Table 3, we can find that *Seq2Seq* baseline system outperformed our *Retrieval* system on all three metrics, which is not consistent with the objective comparison results in Table 1 and Table 2. The mismatching results between objective and subjective results indicate the difficulty of conversation response evaluation. According to the official evaluation results, our *Retrieval* system ranked third among all participant systems, which is not equivalent to its performance on objective evaluations. We think it may be attribute to that our *Retrieval* system is not as flexible as generative models and it does not utilize the textual facts, which can be confirmed in some way by the objective evaluation results in Table 1, i.e, the generative *VariGen* system performed better than *Retrieval* system on both NIST and BLEU metrics.

4 Conclusion

In this paper, we present our conversation response generation systems for DSTC7-Track 2. According to the official evaluation results, our proposed systems obtained best diversity performance on Entropy metrics among all participant systems, and at the same time, our systems also ranked second on NIST metrics, which shows that our proposed systems are promising to be developed further for conversation response generation.

5 Acknowledgments

This work was partially funded by the National Nature Science Foundation of China (Grant No. U1636201).

References

Bordes, A., and Weston, J. 2016. Learning end-to-end goal-oriented dialog. *CoRR* abs/1605.07683.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1657–1668.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145. Morgan Kaufmann Publishers Inc.

Galley, M.; Brockett, C.; Gao, X.; Dolan, B.; and Gao, J. 2018. End-to-end conversation modeling: Dstc7 task 2 description.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Schatzmann, J.; Weilhammer, K.; Stuttle, M. N.; and Young, S. J. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Eng. Review* 21(2):97–126.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, 3776–3784.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 3295–3301.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.

Wen, T.; Gasic, M.; Mrksic, N.; Rojas-Barahona, L. M.; Su, P.; Ultes, S.; Vandyke, D.; and Young, S. J. 2016. A network-based end-to-end trainable task-oriented dialogue system. *CoRR* abs/1604.04562.

Young, S.; Gašić, M.; Thomson, B.; and Williams, J. D. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5):1160–1179.

Zhang, Y.; Galley, M.; Gao, J.; Gan, Z.; Li, X.; Brockett, C.; and Dolan, B. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *arXiv preprint arXiv:1809.05972*.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 654–664.