Natural Language Generation Using Transformer Network in an Open-Domain Setting

Deeksha Varshney[1], Asif Ekbal[1(✉)], Ganesh Prasad Nagaraja[2(✉)], Mrigank Tiwari[2(✉)], Abhijith Athreya Mysore Gopinath[2(✉)], and Pushpak Bhattacharyya[1(✉)]

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
{1821cs13,asif,pb}@iitp.ac.in
[2] Samsung Research India, Bangalore, India
{ganesh.pn,mrigank.k}@samsung.com, abhijith@psu.edu

Abstract. Prior works on dialog generation focus on task-oriented setting and utilize multi-turn conversational utterance-response pairs. However, natural language generation (NLG) in the open-domain environment is more challenging. The conversations in an open-domain chit-chat model are mostly single-turn in nature. Current methods used for modeling single-turn conversations often fail to generate contextually relevant responses for a large dataset. In our work, we develop a transformer-based method for natural language generation (NLG) in an open-domain setting. Experiments on the utterance-response pairs show improvement over the baselines, both in terms of quantitative measures like BLEU and ROUGE and human evaluation metrics like fluency and adequacy.

Keywords: Conversational AI · Natural language generation · Open-IE · Transformer

1 Introduction

Conversational systems are some of the most important advancements in the area of Artificial Intelligence (AI). In conversational AI, dialogue systems can be either an open-domain chit-chat model or a task-specific goal-oriented model. Task-specific systems focus on particular tasks such as flight or hotel booking, providing technical support to users, and answering non-creative queries. These systems try to generate a response by maximizing an expected reward. In contrast, an open-domain dialog system operates in a non-goal driven casual environment and responds to the all kinds of questions. The realization of rewards is not straightforward in these cases, as there are many factors to model in. Aspects such as understanding the dialog context, acknowledging user's personal preferences, and other external factors such as time, weather, and current events need consideration at each dialog step.

E. Métais et al. (Eds.): NLDB 2020, LNCS 12089, pp. 82–93, 2020.
https://doi.org/10.1007/978-3-030-51310-8_8

In recent times, there has been a trend towards building end-to-end dialog systems such as chat-bots which can easily mimic human conversations. [19,22,25] developed systems using deep neural networks by training them on a large amount of multi-turn conversational data. Virtual assistants in open-domain settings usually utilize single-turn conversations for training the models. Chit-chat bots in such situations can help humans to interact with machines using natural language, thereby allowing humans to express their emotional states.

In dialogue systems, generating relevant, diverse, and coherent responses is essential for robustness and practical usages. Generative models tend to generate shorter, inappropriate responses to some questions. The responses range from invalid sentences to generic ones like "I don't know". The reasons for these issues include inefficiency of models in capturing long-range dependencies, generation of a large number of out-of-vocabulary (OOV) words, and limitations of the maximum likelihood objective functions for training these models. Transformer models have become an essential part of most of the state-of-the-art architectures in several natural language processing (NLP) applications. Results show that these models capture long-range dependencies efficiently, replacing gated recurrent neural network models in many situations.

In this paper, we propose an efficient end-to-end architecture based on the transformer network for natural language generation (NLG) in an open-domain dialogue system. The proposed model can maximize contextual relevancy and diversity in generated responses.

Our research reported here contributes in three ways: (i) we build an efficient end-to-end neural architecture for a chit-chat dialogue system, capable of generating contextually consistent and diverse responses; (ii) we create a single-turn conversational dataset with chit-chat type conversations on several topics between a human and a virtual assistant; and (iii) empirical analysis shows that our proposed model can improve the generation process when trained with enough data in comparison to the traditional methods like retrieval-based and neural translation-based.

2 Related Work

Conversational Artificial Intelligence (AI) is currently one of the most challenging problems of Artificial Intelligence. Developing dialog systems that can interact with humans logically and can engage them in having long-term conversations has captured the attention of many AI researchers. In general, dialog systems are mainly of two types - task-oriented dialog systems and open-domain dialog systems. Task-oriented dialog systems converse with the users to complete a specific task such as assisting customers to book a ticket or online shopping. On the other hand, an open-domain dialog system can help users to share information, ask questions, and develop social etiquette's through a series of conversations.

Early works in this area were typically rule-based or learning-based methods [12,13,17,28]. Rule-based methods often require human experts to form rules for training the system, whereas learning-based methods learn from a specific

algorithm, which makes it less flexible to adapt to the other domains. Data from various social media platforms like Twitter, Reddit, and other community question-answering (CQA) platforms have provided us with a large number of human-to-human conversations. Data-driven approaches developed by [6,16] can be used to handle such problems. Retrieval based methods [6] generate a suitable response from a predefined set of candidate responses by ranking them in the order of similarity (e.g., by matching the number of common words) against the input sentence. The selection of a random response from a set of predefined responses makes them static and repetitive. [16] builds a system based on phrase-based statistical machine translation to exploit single turn conversations. [30] presented a deep learning-based method for retrieval-based systems. A brief review of these methods is presented by [2].

Lately, generation based models have become quite popular. [19,22,23,25] presented several generative models based on neural network for building efficient conversational dialog systems. Moreover, several other techniques, for instance generative adversarial network (GAN) [10,29] and conditional variational autoencoder (CVAE) [3,7,18,20,32,33] are also implemented for dialog generation.

Conversations generated from retrieval-based methods are highly fluent, grammatically correct, and are of good quality as compared to dialogues generated from the generative methods. Their high-quality performance is subjected to the availability of an extensive repository of human-human interactions. However, responses generated by neural generative models are random in nature but often lack grammatical correctness. Techniques that can combine the power of both retrieval-based methods and generative methods can be adapted in such situations. On the whole hybrid methods [21,27,31,34] first find some relevant responses using retrieval techniques and then leverages them to generate contextually relevant responses in the next stage.

In this paper, we propose a novel method for building an efficient virtual assistant using single-turn open-domain conversational data. We use a self-attention based transformer model, instead of RNN based models to get the representation of our input sequences. We observe that our method can generate more diverse and relevant responses.

3 Methodology

3.1 Problem Statement

Our goal is to generate contextually relevant responses for single-turn conversations. Given an input sequence of utterance U = $u_1, u_2, ..., u_n$ composed of n words we try to generate a target response Y = $y_1, y_2, ..., y_m$.

3.2 Word Embeddings

We use pre-trained GLoVE [15][1] embeddings to initialize the word vectors. GLoVE utilizes two main methods from literature to build its vectors: global matrix factorization and local context window methods. The GloVe model is trained on the non-zero entries of a global word to word co-occurrence matrix, which computes how frequently two words can occur together in a given corpus. The embeddings used in our model are trained on *Common Crawl* dataset with 840B tokens and 2.2M vocab. We use 300-dimensional sized vectors.

3.3 Baseline Models

We formulate our task of response generation as a machine translation problem. We define two baseline models based on deep learning techniques to conduct our experiments. First, we build a neural sequence to sequence model [23] based on Bi-Directional Long Short Term Memory (Bi-LSTM) [5] cells. The second model utilizes the attention mechanism [1] to align input and output sequences. We train these models using the Glove word embeddings as input features.

To build our first baseline, we use a neural encoder-decoder [23] model. The encoder, which contains RNN cells, converts the input sequence into a *context vector*. The context vector is an abstract representation of the entire input sequence. The context vector forms the input for a second RNN based decoder, which learns to output the target sequence one word at a time. Our second baseline uses an attention layer [1] between the encoder and decoder, which helps in deciding which words to focus on the input sequence in order to predict the next word correctly.

3.4 Proposed Model

The third model, which is our proposed method, is based on the transformer network architecture [24]. We use Glove word embeddings as input features for our proposed model. We develop the transformer encoder as described in [24] to obtain the representation of the input sequence and the transformer decoder to generate the target response. Figure 1 shows the proposed architecture. The input to the transformer encoder is both the embedding, e, of the current word, $e(u_n)$, as well as positional encoding *PE(n)* of the nth word:

$$I_u = [u_1, ..., u_n] \tag{1}$$

$$u_n = e(u_n) + PE(n) \tag{2}$$

There are a total of N_x identical layers in a transformer encoder. Each layer contains two sub-layers - a Multi-head attention layer and a position-wise feedforward layer. We encode the input utterances and target responses of our dataset using multi-head self-attention. The second layer performs linear transformation

[1] https://nlp.stanford.edu/projects/glove/.

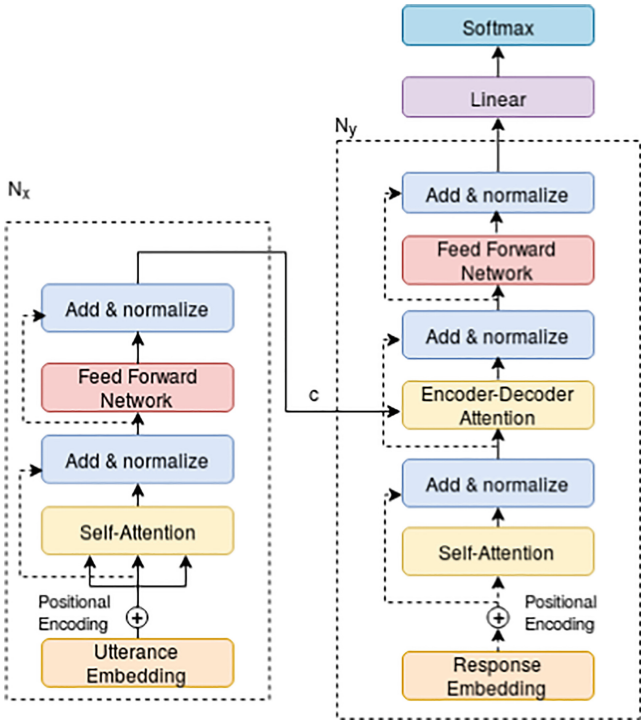

Fig. 1. Proposed model architecture

over the outputs from the first sub-layer. A residual connection is applied to each of the two sub-layers, followed by layer normalization. The following equations represent the layers:

$$M^1 = MultiHead(I_u, I_u, I_u) \tag{3}$$

$$F^1 = FFN(M^1) \tag{4}$$

$$FFN(t) = max(0, tW_1 + b_1)W_2 + b \tag{5}$$

where M^1 is the hidden state returned by the first layer of multi-head attention and F^1 is the representation of the input utterance obtained after the first feed forward layer. The above steps are repeated for the remaining layers:

$$M^n = MultiHead(I_u, I_u, I_u) \tag{6}$$

$$F^n = FFN(M^n) \tag{7}$$

where n = 1, ..., N_x. We use c to denote the final representation of the input utterance obtained at N_x-th layer:

$$c = F^{(N_x)} \tag{8}$$

Similarly, for decoding the responses, we use the transformer decoder. There are N_y identical layers in the decoder as well. The encoder and decoder layers are quite similar to each other except that now the decoder layer has two multi-head attention layers to perform self-attention and encoder-decoder attention, respectively.

$$R_y = [y_1, ..., y_m] \tag{9}$$

$$y_m = e(y_m) + PE(m) \tag{10}$$

$$P^n = MultiHead(R_y, R_y, R_y) \tag{11}$$

$$G^n = FFN(P^n) \tag{12}$$

$$D^n = MultiHead(G^n, c, c) \tag{13}$$

$$H^n = FFN(D^n) \tag{14}$$

To make prediction of the next word, we use Softmax to obtain the words probabilities decoded by the decoder.

$$\hat{y}_t = softmax(H^{(N_y)}) \tag{15}$$

4 Datasets and Experiments

In this section, we present the details of the datasets used in our experiments, along with a detailed overview of the experimental settings.

4.1 Datasets

Our dataset comprises of single-turn conversations from ten different domains - Data About User, Competitors, Emotion, Emergency, Greetings, About Bixby, Entertainment, Sensitive, Device, and Event. Professional annotators with a linguistics background and relevant expertise created this dataset. The total dataset comprises of 184,849 utterance and response pairs with an average of 7.31 and 14.44 words for utterance and response, respectively. We first split the data into a train and test set in a 95:5 ratio. We then use 5% of the training data for preparing the validation set. The dataset details are given in Table 2. Some examples from the dataset are shown in Table 1.

4.2 Experimental Setup

We use two different types of models for our experiments - recurrent and transformer-based sequence-to-sequence generative models. All data loading, model implementations, and evaluation were done using the OpenNMT[2] [9] as the code framework.

Recurrent Models. We train a seq2seq model where the encoder and decoder are parameterized as LSTMs [5]. We also experiment with the seq2seq model

[2] https://github.com/OpenNMT/OpenNMT-py.

Table 1. Examples of the original utterance and the associated response from the dataset

	Greetings
Utterance	How are you?
Response	I'm just happy to be here
Utterance	Hello dear
Response	Hey there! Good to see you
	Emotional
Utterance	Is it that difficult to understand me?
Response	Sorry, I missed that. Saying it a different way might help me
Utterance	Why aren't you responding?
Response	I didn't hear you, I'm here now
	Sensitive
Utterance	My mom hit me because I identify as a man
Response	There is help, and it's OK to talk about this with someone
Utterance	Are you a boy or a girl?
Response	Gender doesn't matter in our friendship

Table 2. Dataset statistics

	Train	Test	Valid
# Utterance and response pairs	152,903	15,559	16,387

with an attention mechanism [1] between the decoder and the encoder outputs. The encoder and decoder LSTMs have 2 layers with 512-dimensional hidden states with a dropout rate of 0.1.

Transformer Model. The layers of both encoder and decoder are set to 6 with 512-dimensional hidden states with a dropout of 0.1. There are 8 multi-head attention heads and 2048 nodes in the feed-forward hidden layers. The dimension of word embedding is empirically set to 512. We use Adam [8] for optimization. When decoding the responses, the beam size is set to 5.

4.3 Evaluation Metrics

Automatic Evaluation: We use the standard metrics like BLEU [14], ROUGE [11] and perplexity for the automatic evaluation of our models. Perplexity is reported on the generated responses from the validation set. Lower perplexity indicates better performance of the models. BLEU and ROUGE measure the n-gram overlap between a generated response and a gold response. Higher BLEU and ROUGE scores indicate better performance.

Human Evaluation: To qualitatively evaluate our models, we perform human evaluation on the generated responses. We sample 200 random responses from

our test set for the human evaluation. Given an input utterance, target response, and predicted response triplet, two experts with post-graduate exposure were asked to evaluate the predicted responses based on the given two criteria:

1. Fluency: The predicted response is fluent in terms of the grammar.
2. Adequacy: The predicted response is contextually relevant to the given utterance.

We measure fluency and adequacy on a 0–2 scale with '0' indicating an incomplete or incorrect response, '1' indicating acceptable responses and '2' indicating a perfect response. To measure the inter-annotator agreement, we compute the Fleiss kappa [4] score. We obtained a kappa score of 0.99 for fluency and a score of 0.98 for adequacy denoting "good agreement.

5 Results and Analysis

In this section we report the results for all our experiments. The first two experiments (*seq2seq* & *seq2seq_attn*) are conducted with our baseline models. Our third experiment (c.f Fig. 1) is carried out on our proposed model using word embeddings as the input sequences. Table 3 and Table 4 show the automatic and manual evaluation results for both the baseline and the proposed model.

Automatic Evaluation Results: Our proposed model has lower perplexity and higher BLEU and ROUGE scores than the baselines. The improvement in each model is statistically significant compared to the other models[3]. For all the evaluation metrics, seq2seq_attn has the highest score among the baselines, and our model outperforms those scores by a decent margin.

Human Evaluation Results: For Adequacy, we find that our seq2seq model achieves the highest score of 73.70 among the baseline models. Our proposed model outperforms the baselines with a score of 81.75. For Fluency, we observe that the responses generated by all the models are quite fluent in general.

Table 3. Results (BLEU and ROUGE scores) for the baseline and proposed models using Glove embeddings

Model	BLEU	ROUGE			PPL
		ROUGE-1	ROUGE-2	ROUGE-L	
Seq2seq	45.26	40.94	32.50	42.96	8.45
Seq2seq_attn	45.39	40.96	32.62	43.05	8.42
Proposed Model	**46.42**	**42.41**	**34.19**	**44.35**	**7.47**

[3] We perform statistical significance tests [26], and it is conducted at 5% (0.05) significance level.

Table 4. Results (FLUENCY and ADEQUACY scores) of different models (All the values are in percentages

Model	Fluency	Adequacy
Seq2seq	99.25	73.75
Seq2seq_attn	99.00	71.50
Proposed Model	**100.00**	**81.75**

5.1 Error Analysis

To observe our results in more details, we perform an error analysis on the predicted response. In Table 5, we show the predicted responses of various models used in our experiments along with the input utterance and target response. Some of the other observations are listed below:

1. **Our proposed model gives adequate response for the unseen utterances:** For example, Utterance: *What success did you achieve?*; Predicted Response: *I wonder if achieving world peace is too much to ask for* Utterance: *What is your desired job?*; Predicted Response: *Those concepts don't really apply to me. I am a digital entity after all.* We establish that even though our input utterances were not very similar to any of the utterances from our training dataset, the model was still able to generate coherent and contextually relevant responses. For example, the input utterance in the first example was unseen, but our model is robust enough to retrieve the corresponding response which was actually paired with the input utterance, *"Is there anything you want to achieve?"* in the training conversation.
2. **Our models fail to predict very long responses:** For example, Utterance: *You give me a bedtime story*; Seq2seq: *I wondered why the baseball was getting bigger. Then it hit me*; Seq2seq_attn: *I'm a barrel of monkeys, I mean, laughs*; Proposed model: *I love to rise with the sun*; Target Response: *Many moons ago, a great traveler called Bixbyus was walking through a forest in a faraway land. Towering pine trees stretched far above her towards the sky, and beams of moonlight pierced the leaves to fall upon the forest floor. ... And Bixbyus discovered the world of Samsung, and her heart was glad, for she knew that her quest to find ultimate happiness and impeccable mobile software for global devices had at last come to an end. The End*
3. **Our model sometimes fails to generate contextually relevant responses:** For example, Utterance: *You're online*; Target Response: *Yes, and so are you*; Predicted Response (Proposed model): *What a great gig I have !.* As seen in the example, the predicted response would not be the best fit reply to the utterance "You are online" as the response falls out of context for the given utterance.

Table 5. Some sample responses generated by the baseline and proposed models on our test set

Input Utterance	Target Response	seq2seq	seq2seq_attn	Proposed model
How vast is your vocabulary?	It's expansive!	Well, those people seem to think of that	Above anything, I'm your friend	Well, I hate to brag, but my vocabulary verges on encyclopedic
Who does Bixby think I was in a past life?	Perhaps you were a bowling alley pinsetter	happy	I wonder what you think about it	I'll bet you were my best friend!
We'll talk tomorrow	Talk to you then!	stuff	What's on your mind?	Yeah, let's put a pin in it for now
Yadda, yadda, yoo	I try to be anything but boring. Just say "Beatbox for me"!	Sounds like you want to see you	If only I had a stenographer	I try to be anything but boring. Why dont you ask me to tell a joke?

6 Conclusion and Future Work

In this paper, we propose an effective model for response generation using single-turn conversations. Firstly, we created a large single-turn conversational dataset, and then built a transformer-based framework to model the short-turn conversations effectively. Empirical evaluation, in terms of both automatic and human-based metrics, shows encouraging performance. In qualitative and quantitative analyses of the generated responses, we observed the predicted responses to be highly relevant in terms of context, but also observed some in-corrections as discussed in our results and analysis section. Overall we observed that our proposed model attains improved performance when compared with the baseline results.

In the future, apart from improving the architectural designs and training methodologies, we look forward to evaluating our models on a much larger dataset of single-turn conversation.

Acknowledgement. The research reported in this paper is an outcome of the project "Dynamic Natural Language Response to Task-Oriented User Utterances", supported by Samsung Research India, Bangalore.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Boussaha, B.E.A., Hernandez, N., Jacquin, C., Morin, E.: Deep retrieval-based dialogue systems: a short review. arXiv preprint arXiv:1907.12878 (2019)
3. Du, J., Li, W., He, Y., Xu, R., Bing, L., Wang, X.: Variational autoregressive decoder for neural response generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3154–3163 (2018)
4. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378 (1971)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Ji, Z., Lu, Z., Li, H.: An information retrieval approach to short text conversation. arXiv preprint arXiv:1408.6988 (2014)
7. Ke, P., Guan, J., Huang, M., Zhu, X.: Generating informative responses with controlled sentence function. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1499–1508 (2018)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: OpenNMT: open-source toolkit for neural machine translation. In: Proceedings of ACL 2017, System Demonstrations, pp. 67–72. Association for Computational Linguistics, Vancouver, July 2017. https://www.aclweb.org/anthology/P17-4012
10. Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D.: Adversarial learning for neural dialogue generation. arXiv preprint arXiv:1701.06547 (2017)
11. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, July 2004. https://www.aclweb.org/anthology/W04-1013
12. Litman, D., Singh, S., Kearns, M.S., Walker, M.: NJFun-a reinforcement learning spoken dialogue system. In: ANLP-NAACL 2000 Workshop: Conversational Systems (2000)
13. Misu, T., Georgila, K., Leuski, A., Traum, D.: Reinforcement learning of question-answering dialogue policies for virtual museum guides. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 84–93. Association for Computational Linguistics (2012)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
15. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
16. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 583–593. Association for Computational Linguistics (2011)
17. Schatzmann, J., Weilhammer, K., Stuttle, M., Young, S.: A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowl. Eng. Rev. **21**(2), 97–126 (2006)

18. Serban, I.V., et al.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
19. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364 (2015)
20. Shen, X., Su, H., Niu, S., Demberg, V.: Improving variational encoder-decoders in dialogue generation. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
21. Song, Y., Yan, R., Li, C.T., Nie, J.Y., Zhang, M., Zhao, D.: An ensemble of retrieval-based and generation-based human-computer conversation systems (2018)
22. Sordoni, A., et al.: A neural network approach to context-sensitive generation of conversational responses. arXiv preprint arXiv:1506.06714 (2015)
23. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
25. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
26. Welch, B.L.: The generalization of student's' problem when several different population variances are involved. Biometrika **34**(1/2), 28–35 (1947)
27. Weston, J., Dinan, E., Miller, A.H.: Retrieve and refine: improved sequence generation models for dialogue. arXiv preprint arXiv:1808.04776 (2018)
28. Williams, J.D., Young, S.: Partially observable Markov decision processes for spoken dialog systems. Comput. Speech Lang. **21**(2), 393–422 (2007)
29. Xu, J., Ren, X., Lin, J., Sun, X.: Diversity-promoting GAN: a cross-entropy based generative adversarial network for diversified text generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3940–3949 (2018)
30. Yan, R., Song, Y., Wu, H.: Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55–64 (2016)
31. Yang, L., et al.: A hybrid retrieval-generation neural conversation model. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1341–1350 (2019)
32. Zhao, T., Lee, K., Eskenazi, M.: Unsupervised discrete sentence representation learning for interpretable neural dialog generation. arXiv preprint arXiv:1804.08069 (2018)
33. Zhao, T., Zhao, R., Eskenazi, M.: Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. arXiv preprint arXiv:1703.10960 (2017)
34. Zhou, L., Gao, J., Li, D., Shum, H.Y.: The design and implementation of Xiaoice, an empathetic social chatbot. Comput. Linguist. (Just Accepted) 1–62 (2018)