Multi-turn Dialogue Generation Using Self-attention and Nonnegative Matrix Factorization

Chen Hu1*, Neng Wan1, Songtao Cai1 and Guangping Zeng1

1Beijing Key Laboratory of Knowledge Engineering for Materials Science, University of Science and Technology Beijing, Beijing 100083, China
*E-mail: s20190677@xs.ustb.edu.cn

Abstract. Recently, neural generative models have shown significant potential in the field of human-computer interaction, especially the dialogue system. In this paper, we propose a new model for multi-turn dialogue generation, which uses the self-attention mechanism to extract relevant information in the dialogue history, and utilizes the nonnegative matrix factorization (NMF) to learn topic vectors from an external corpus. The response generated by the model is not only affected by the dialogue context, but also by the corresponding topic vectors. We conduct experiments on a public dataset to evaluate the performance of our model. Experimental results demonstrate that our model has the ability to generate more diverse and context-sensitive responses.

1. Introduction

Research on how to generate diverse and context-sensitive responses for multi-turn dialogue is a valuable task of dialogue systems, which has aroused wide concern in the industry and academia. Traditional sequence-to-sequence frameworks [1, 2] fail in leveraging the dialogue history and tend to generate a large number of general responses. To make use of the dialogue history, many models have been proposed. HRED [3] and its variants [4, 5] adopt a widely accepted method which utilizes a hierarchical recurrent neural network (RNN) [6] based architecture to encode each utterance as well as the entire dialogue history. The main disadvantage of this method is that it ignores the importance of different information in the context, which may bring a lot of irrelevant noise when generating responses from the model. HRAN [7] addresses the shortcomings of HRED based models by using conventional attention mechanism at the token level and utterance level to extract important parts of tokens and utterances, respectively. However, it may lose some relevant information due to the position bias problem [8]. Recently, [9] uses the famous self-attention mechanism [10] for capturing the relevant context and responses, which achieves better performance than previous models.

To further increase the diversity of the generated responses, some studies [11–13] consider leveraging topic modeling methods such as Latent Dirichlet Allocation (LDA) [14] and nonnegative matrix factorization (NMF) [15] to introduce topic information to the process of generating responses. In this work, we attempt to leverage the self-attention mechanism to extract relevant information in the dialogue history and apply the NMF algorithm for topic modeling to generate more diverse and context-sensitive replies. We also evaluate the model on a public multi-turn dialogue dataset and make a comparison with several baseline models.
The contributions of this work are as follow: 1) We propose a model which could effectively use the context and topic information for response generation. 2) Empirical studies on automatic evaluation metrics and human judgment indicate the effectiveness of our model.

2. Related work
Recently, with the rapid development of deep neural networks, dialogue systems have made significant progress. However, how to make improvement in the informativeness and diversity of the generated responses is still a challenging work. Many methods have been explored to utilize the context information to increase the quality of the responses. HRED [3] conceived a hierarchical architecture which recurrently encoded each utterance in the context to an utterance vector. Then, these vectors are processed by another RNN to extract the context information. VHRED [5] aimed to improve the performance and robustness of the HRED by introducing latent variables to the middle state. HRAN used a hierarchical attention mechanism [7] to learn the importance of relevant utterances and tokens in the context. ReCoSa [9] toke the advantage of the self-attention mechanism [10] in capturing long-distance dependence to extract the relevance of the context and responses.

Some research focused on utilizing topic information to enrich the content of the responses. Topic aware sequence to sequence(TA-Seq2Seq) [11] model is proposed for single-turn dialogue generation which leveraged a pretrained Twitter LDA [16] to acquire topic words of the context and decoded with a joint attention mechanism. THRED [12] acquired topic words form a Latent dirichlet allocation(LDA) [14] model and proposed a hierarchical joint attention mechanism similar to the HRAN for multi-turn dialogue generation. [13] aimed to enhance the diversity of the replies generated by the RNN-based model and applied a nonnegative matrix factorization(NMF) [15] algorithm to replace the widely used LDA model for topic modeling, which is more efficient in calculation and easier to implement. Inspired by these work, our proposed model employs the self-attention mechanism and the NMF algorithm to exploit both the dialogue history and the topic information to increase the generation ability.

3. The Model
In this section, we introduce our proposed model. The architecture of the model is shown in Figure 1. We use the word-level encoder and the self-attention to get the representation of the dialogue context. In the process of response generation, the decoder employs two different attention to jointly affect the generation probability. Details will be described as follows.

3.1. Word-level encoder
Let \( c = \{u_1, \ldots, u_N\} \) denote the dialogue context, each utterance in \( c \) is defined as \( u_i = \{u_{i,1}, \ldots, u_{i,L_i}\} \). We employ a bidirectional recurrent neural network with gated recurrent units (Bi-GRU) [17] to encode each \( u_i \), \( i \in \{1, \ldots, N\} \) as fixed-dimension hidden vectors \( \{h_{i,1}, \ldots, h_{i,L_i}\} \). \( \forall k \in \{1, \ldots, L_i\} \), k-th hidden state of the forward GRU [18] and the backward GRU are defined as \( \overrightarrow{h}_{i,k} \) and \( \overleftarrow{h}_{i,k} \), respectively. \( h_{i,k} \) is the concatenation of \( \overrightarrow{h}_{i,k} \) and \( \overleftarrow{h}_{i,k} \). The \( \overrightarrow{h}_{i,k} \) is calculated as:

\[
\begin{align*}
    z_k &= \sigma(W_z w_{i,k} + U_z \overrightarrow{h}_{i,k-1}) \\
    r_k &= \sigma(W_r w_{i,k} + U_r \overrightarrow{h}_{i,k-1}) \\
    s_k &= \tanh(W_s w_{i,k} + U_s ( \overrightarrow{h}_{i,k-1} \circ r_k ) ) \\
    \overrightarrow{h}_{i,k} &= (1 - z_k) \circ s_k + z_k \circ \overrightarrow{h}_{i,k-1}
\end{align*}
\]

where \( z_k \) and \( r_k \) are the update gate and reset gate, respectively. \( w_{i,k} \) is the word embedding of \( u_{i,k} \) and \( \sigma \) is the sigmoid function. \( \circ \) represents the element wise multiplication. \( W_z, W_r, W_s, \)
$U_z, U_r, U_s$ are parameters. The computation of the backward hidden state $\overrightarrow{h}_{i,k}$ is conducted in reverse direction with a calculation similar to the $\overrightarrow{h}_{i,k}$. For each utterance $u_i$ in the context, we use $h_{i,L_i}$ as the representation of the utterance. Then, we acquire utterance representations \{ $h_{u_1}^{u_1}, \ldots, h_{u_N}^{u_N}$ \}.

It is well known that self-attention mechanism completely abandons the recurrence. So it is crucial to add some positional information to the input embeddings. In this paper, we use the same positional embeddings as the Transformer [10]. The utterance representations and the corresponding positional embeddings are summed to obtain the final input representations \{ $h_{u_1 Pos_1}, \ldots, h_{u_N Pos_N}$ \}.

![Figure 1. The overall architecture of the model.](image)

### 3.2. Context Self-Attention

In our work, we follow the same approach as ReCoSa [9] which adopt the multi-head attention mechanism [10] to compute a sequence’s representation. The matrices of queries, keys and values are $Q \in R^{n \times d}$, $K \in R^{n \times d}$, and $V \in R^{n \times d}$, respectively. $n$ is the number of query vectors. The attention scores are computed by the scaled dot-product attention [10] as:

$$Attention(Q, K, V) = softmax\left(\frac{Q K^T}{\sqrt{d}}\right)V$$

where $d$ denotes the dimension of the hidden vector computed by the word-level encoder.

We employ $H$ parallel heads. For the $i$-th head, we use parameter matrices $W_i^{Q} \in R^{d \times d/H}$, $W_i^{K} \in R^{d \times d/H}$, $W_i^{V} \in R^{d \times d/H}$ for projecting the corresponding queries, keys and values. Then the scale dot-product attention take these projected variables as parameters to output the mixed representations. The output of each parallel head are concatenated and projected to form the final representation:

$$head_i = Attention(W_i^{Q}, W_i^{K}, W_i^{V})$$

$$M = Contact(head_1, \ldots, head_H)$$

$$O = MW^O$$

where $M \in R^{n \times d}$ and $W^O \in R^{d \times d}$. 
The matrix of utterance vectors \( \{ h_{pos1}^{u1}, \ldots, h_{posN}^{uN} \} \) is fed to the multi-head attention mechanism as query, key, and value to get the output \( O \). Then, we use layer normalization [19] to further process the output and obtain the context representation \( O_c \).

3.3. Topic Representation
In this paper, we use a nonnegative matrix factorization (NMF) [15] algorithm to obtain topic vectors. Given the original matrix \( V \), the idea of NMF is to find such two matrices \( W \) and \( H \), where the product of \( W \) and \( H \) approximates to \( V \) as:

\[
V \approx WH \tag{4}
\]

The object of NMF is to minimize the frobenius norm [15] as the following:

\[
\text{minimize } \parallel V - WH \parallel_2^2 \tag{5}
\]

In our work, \( V \in \mathbb{R}^{m \times n} \) is the representation of the corpus for topic modeling, where \( m \) and \( n \) are the number of words and documents in the text corpus, respectively. Each value of the matrix \( V \) is the term frequency–inverse document frequency (TF-IDF) [20] of the corresponding word in the corpus, \( W \in \mathbb{R}^{m \times r} \) and \( H \in \mathbb{R}^{r \times n} \) are learned from \( V \) through NMF, where \( r \) is the number of topics. ∀ \( i \in \{1, \ldots, r\} \), \( x_i \) is the \( i \)-th column vector of the topic matrix \( W \). For each context \( c \) in the dialogue corpus, the matrix \( H \) of \( c \) is defined as \( H(c) \) and it can be calculated by the NMF as shown in Equation (6):

\[
c \approx W \times H(c) \tag{6}
\]

where \( W \) is obtained from a large text corpus which is used for training the NMF. In order to avoid the sparse problem, we use the alternative method in RNN-NMF chatbot [13] to compute \( H(c) \) according to:

\[
H(c) = W^T c \tag{7}
\]

The \( H(c) = (k_1, \ldots, k_r)^T \) will indicate the model how much attention should be given to each topic during the process of response generation.

3.4. Decoder
Given the context representation and topic vectors, the decoder utilizes two attention mechanisms, namely message self-attention and topic attention to generate responses step by step. Message self-attention is a multi-head attention. It enables the decoder to consider which parts of the context information are relevant to focus on. For this attention, the previous hidden state of the decoder \( s_{t-1} \) is the query, and the context representation \( O_c \) obtained by the encoder is the key and value. We use the hidden representation of the last utterance in the context to initialize the \( s_0 \). At the current time step \( t \), the output of the message self-attention \( O_m^c \) is computed via the equation (3).

Suppose that the context representation \( O_c \) and the hidden state at the previous time step \( s_{t-1} \) are known, the current topic attention can be calculated as the following:

\[
\alpha_{tj} = \frac{\exp(\eta(s_{t-1}, x_j, O_c))}{\sum_{i=1}^{r} \exp(\eta(s_{t-1}, x_i, O_c))} \tag{8}
\]

\[
c_t = \sum_{j=1}^{r} \alpha_{tj} x_j
\]

where \( x_j \) denotes the \( j \)-th column vector of \( W \) obtained by NMF according to the equation (4), \( \eta \) is a multi-layer perception (MLP). We use context representation \( O_c \) as a parameter in the
process of computing topic attention, because $O_c$ contains the context features extracted by the self-attention mechanism. Then, the decoder updates the current hidden state as:

$$s_t = f(y_{t-1}, s_{t-1}, O_t^m, c_t)$$

where $y_{t-1}$ is the predicted distribution at the previous time step, $f$ is a GRU unit.

With the dialogue context $c$ and the topic matrix $W$ calculated by the NMF, we can obtain the $H(c) = (k_1, \ldots, k_r)^T$ according to the equation (4). $\forall i \in \{1, \ldots, r\}$, $k_i$ indicates the importance of the topic vector $x_i$ in $W$, when a word $w$ belongs to the topic vector $x_i$, the generation probability of $w$ is additionally biased to the corresponding $k_i$ and the topic attention. We define the generation probability as follows:

$$\begin{align*}
\Psi_t^m(w) &= \delta(w^T(W^m_y y_{t-1} + W^m_s s_{t-1} + W^m_o O_t^m + b_m)) \\
\Psi_t^c(w) &= \delta(w^T(W^c_y y_{t-1} + W^c_s s_{t-1} + W^c_c c_t + b_c)) \\
p(y_t = w) &\propto \exp(\Psi_t^m(w)) + (\sum_{j=1}^r k_j(w \in x_j)) \exp(\Psi_t^c(w))
\end{align*}$$

where $\delta$ is the tanh function, $w$ is the one-hot indicator vector for word $w$, $\propto$ denotes that generation probability is proportional to the right side of the formula. $W^m_y, W^m_s, W^m_o, W^c_y, W^c_s, W^c_c, b_m, b_c$ are parameters.

4. Experiments

In this section, we conduct experiments on a public multi-turn dialogue dataset to evaluate our model.

4.1. Datasets

Two public datasets 20NewsGroups [21] and DailyDialog [22] are used in our work. The 20NewsGroups dataset is used as the text corpus for training the NMF algorithm to obtain the topic vectors. For each document in this dataset, we extract TF-IDF features and convert the document into a bag-of-words representation. These representations are combined into the original matrix for NMF to learn the matrix which contains 10 topic vectors, each topic vector corresponds to 1,000 topic words. The DailyDialog dataset is a popular multi-turn dialogue dataset extracted from websites for people to practice conversational English. It contains 13,000 conversions which are relevant to our daily life. We convert the dataset into the form of context-response pair, each context contains at least one utterance and utterances will be separated from each other by a special token if there are more than one in the context. After that, we obtain 76,052, 7,069, 6,740 pairs as training, validation, and test set.

4.2. Baselines and Experiment Setting

In our work, we use three baseline models including the Seq2Seq model with an attention mechanism (S2SA) [2], RNN-NMF [13], and ReCoSa [9].

The batch size is set to 128 and the size of hidden state in all models is set to 512. The number of parallel heads in our model and the ReCoSa is 8. For DailyDialog, the vocabulary size is set to 16,000. The Adam optimizer [23] is used in the process of training our model and we set the learning rate to 0.0001. Both our proposed model and the baselines are run on a NVIDIA Quadro M4000 GPU with Pytorch.
4.3. Evaluation metrics
We follow the existing works [9,11] and use two automatic metrics as well as the human judgment to evaluate our model as follows:

**Perplexity:** we use the perplexity (PPL) as the metric to measure the fluency of responses generated by the model. Lower perplexity generally means the model has a better performance in generating responses. In our experiments, The perplexity on the entire test will be calculated to evaluate the model’s generation ability.

**Distinct-1, Distinct-2:** distinct-1 and distinct-2 are employed to evaluate the diversity of the generated responses which calculate the radios of distinct unigrams and bigrams in the replies generated by the model, respectively. Higher ratios means that the responses generated by the model tend to be longer and more diverse.

**Human judgment:** In our work, we conduct human judgment to evaluate the performance of different models. 150 samples are randomly selected from the test set, each sample contains the dialogue context and two responses which are generated by our model and a baseline, respectively. Two annotators are required to compare our response with the response generated by the baseline in terms of coherence, diversity and relevance. The result of each comparison is marked as one of "win", "loss", and "tie". For instance, "win" indicates that the quality of the response given by our model is better than the baseline. We calculate and record the percentage of "win", "lose", and "tie" in each comparison with another baseline, and use these data as an important reference for giving the evaluation of our model’s performance.

4.4. Evaluation Result
The results of automatic evaluation metrics are shown in Table 1. We can see that ReCoSa and our proposed model outperform the other baseline models, because both of them utilize relevant information of different utterances in the dialogue history by the self-attention. RNN-NMF performs better than the traditional S2SA, which indicates that topic vectors learned by the NMF algorithm also affect the generation ability of the model, especially the diversity of the responses. Notice that our proposed model performs best on both the perplexity and the distinct scores. This means responses generated by our model are more coherent and diverse than all baselines.

| Model  | PPL   | Distinct-1 | Distinct-2 |
|--------|-------|------------|------------|
| S2SA   | 33.667| 0.489      | 1.686      |
| RNN-NMF| 33.095| 0.662      | 2.767      |
| ReCoSa | 32.822| 1.162      | 5.372      |
| ours   | 32.786| 1.461      | 7.736      |

Human judgment results are shown in Table 2. Notice that the percentage of "win" in each comparison is higher than the percentage of "lose", which indicates that our model achieves better performance compared to other baseline models. The Kappa scores [24] are calculated to measure the consistency between the annotators in our experiment.

4.5. Case Study
Table 3 provides some examples of the responses generated by our proposed model and other baseline models. From the table we can see that responses generated by our model are inclined...
to be more diverse and informative than other baseline models, which indicates that our model
can not only effectively leverage the relevant and valuable information in the dialogue history
to generate context-sensitive responses, but also take the advantage of the topic information
obtained by the NMF to enrich the content of the generated responses. For example, in the
case 1, our model considers the relevant context and utilizes topic words "look" and "north" to
generate the appropriate response.

Table 2. Human judgment results.

| win(%) | lose(%) | tie(%) | Kappa |
|--------|---------|-------|-------|
| ours vs S2SA | 39.3 | 4.0 | 56.7 | 0.521 |
| ours vs RNN-NMF | 36.0 | 5.7 | 58.3 | 0.546 |
| ours vs ReCoSa | 31.3 | 10.3 | 58.3 | 0.559 |

Table 3. Case study.

| Context | A: We are approaching the top of the mountain. B: What a beautiful view we have here. A: Yes, you can have a bird’s eye view of the whole city. | A: what’s wrong with you? B: I have a sore throat and headache. |
|---------|--------------------------------------------------------------------------------------------------------|---------------------------------------------------------------|
| S2SA    | Response: I see.                                                                                     | Response: I’m sorry.                                           |
| RNN-NMF | Response: Yes, we have a few minutes.                                                                 | Response: What’s the problem?                                  |
| ReCoSa  | Response: That’s a good idea.                                                                          | Response: How long have you been working?                      |
| ours    | Response: Well, I’ll take a look at the famous scenery in the north.                                    | Response: Oh, that’s too bad, have you got a fever?             |

5. Conclusion
In this paper, we propose a new model for multi-turn dialogue generation. The model leverages
the self-attention for extracting relevant context and the NMF algorithm for topic modeling.
The generation process of our model is jointly affected by the topic information and the context.
Empirical results on automatic evaluation metrics and human judgement show that our proposed
model outperforms all the baseline models in the ability of generating more diverse and context-
sensitive responses.

References
[1] Shang, L., Lu, Z. and Li, H., 2015. Neural responding machine for short-text conversation. preprint
arXiv:1503.02364
[2] Vinyals, O. and Le, Q., 2015. A neural conversational model. preprint arXiv:1506.05869
[3] Serban, I., Sordoni, A., Bengio, Y., Courville, A. and Pineau, J., 2016, March. Building end-to-end dialogue
systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on
Artificial Intelligence vol 30, no 1
[4] Serban, I., Klingier, T., Tesauro, G., Talamadupula, K., Zhou, B., Bengio, Y. and Courville, A., 2017, February. Multiresolution recurrent neural networks: An application to dialogue response generation. In Proceedings of the AAAI Conference on Artificial Intelligence vol 31, no 1

[5] Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A. and Bengio, Y., 2017, February. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the AAAI Conference on Artificial Intelligence vol 31, no 1

[6] Elman, J.L., 1990. Finding structure in time. Cognitive science. 14 179-211

[7] Xing, C., Wu, Y., Wu, W., Huang, Y. and Zhou, M., 2018, April. Hierarchical recurrent attention network for response generation. In Proceedings of the AAAI Conference on Artificial Intelligence vol 32, no 1

[8] Kolen, J.F. and Kremer, S.C., 2001. Gradient flow in recurrent nets: The difficulty of learning longterm dependencies

[9] Zhang, H., Lan, Y., Pang, L., Guo, J. and Cheng, X., 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. preprint arXiv:1907.05339

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. In Advances in neural information processing systems 5998-6008

[11] Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M. and Ma, W.Y., 2017, February. Topic aware neural response generation. In Proceedings of the AAAI Conference on Artificial Intelligence vol 31, no 1

[12] Dziiri, N., Kamalloo, E., Mathewson, K.W. and Zaiane, O., 2018. Augmenting neural response generation with context-aware topical attention. preprint arXiv:1811.01063

[13] Guo, Y., Haonian, N., Lin, Z., Liiskij, N., Lyu, H., Needell, D., Qu, J., Sojico, H., Wang, Y., Xiong, Z. and Zou, Z., 2019. Topic-aware chatbot using Recurrent Neural Networks and Nonnegative Matrix Factorization. preprint arXiv:1912.00315

[14] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research. 3 993-1022

[15] Lee, D.D. and Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature. 401 788-791

[16] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H. and Li, X., 2011, April. Comparing twitter and traditional media using topic models. In European conference on information retrieval (Berlin: Springer) pp 338-349

[17] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. Preprint arXiv:1409.0473

[18] Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. Preprint arXiv:1409.1259

[19] Ba, J.L., Kiros, J.R. and Hinton, G.E., 2016. Layer normalization. Preprint arXiv:1607.06450

[20] Salton, G. and Yu, C.T., 1973. On the construction of effective vocabularies for information retrieval. Acm Sigplan Notices, 10. 48-60

[21] Lang, K., 2008. The 20 news groups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/

[22] Li, Y., Su, H., Shen, X., Li, W., Cao, Z. and Niu, S., 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. preprint arXiv:1710.03957

[23] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. preprint arXiv:1412.6980

[24] Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychological bulletin. 76 378