2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Last Utterance-Context Attention Model for Multi-Turn Response Generation

1st Guodong Zhang
School of Internet of things engineering,
Jiangnan University
Wuxi, China
jndx_wlw518b@163.com

2nd Li Mao
School of Internet of things engineering,
Jiangnan University
Wuxi, China
wxmaoli@163.com

3rd Jun Sun
School of Internet of things engineering,
Jiangnan University
Wuxi, China
sunjun_wx@hotmail.com

***Abstract*—Recently, conversation response generation task is attracting the attention of more and more researchers. Different from single-turn response generation, multi-turn response generation not only focuses on fluency, but also needs to make use of contextual information. Therefore, we believe that an appropriate response should be coherent to the last utterance, and take conversation history into consideration at the same time. We propose a Last Utterance-Context Attention model. The last utterance attention calculates each word in last utterance and form them as a vector. Representation of each utterance is processed by the context attention and formed as a vector as well. Then the two vectors are concatenated as a context vector for decoding the response. In addition, we also apply the multi-head self-attention mechanism to focus more on the key words in each utterance. Both automatic and human evaluation results show that our model outperform baseline models for multi-turn response generation.**

Keywords—response generation, multi-turn, attention model, hierarchical model

I. INTRODUCTION

There are two main categories of conversational systems, called task-oriented and non-task-oriented. Task-oriented dialog systems are mainly used in some specific scenes，such as goods finding, hotel booking, restaurant booking etc. [1] Different from the Task-oriented dialog systems, non-task-oriented dialogue systems can be widely used in many scenes. We also call them "chatbot". They can also be divided into two categories: generative or retrieval-based. Retrieval-based chatbot select a proper response for the conversation from the available data. Generative chatbot can generate new response for the current conversation mainly using deep learning methods. Natural language processing has achieved great progress, especially in the area of deep learning. Therefore, data-driven generative models have become increasingly popular.

Seq2Seq models generate a reply **r** based on an input query **q** [3]. Based on encoder-decoder framework, researchers have achieved a great success in single-turn response generation [5]. Then, researchers have taken contextual information into consideration, hoping that response can be coherent and context-sensitive in multi-turn conversation.

[7]directly concatenated context utterances and the current query. In this case, the problem becomes a single-turn conversation generation. Hierarchical seq2seq models are widely used by many researchers, firstly they obtain the semantics at the sentence level, and then integrated them into the semantics of the whole context [8]. [9] committed to solving the following two issues: one is making the context information more useful, another is trying to figure out the effect of contextual information. We believe that a good response to a multi-turn conversation should be coherent to the last utterance at first, and then to the whole context. In addition, focusing more on the key words in each utterance can obviously improve the quality of response. Based on the concepts above, our contributions are as follows.

Firstly, to focus more on the key words when encoding the utterances, we apply the attention mechanisms called multi-head self-attention to get better utterance representations. Secondly, we propose a last utterance-context attention model which applies attention mechanism both on each word in the last utterance and on representation of each utterance. Several experiments are conducted on two datasets to compare our model with others. The evaluation is not only conducted in an automatic way, but also in a manual way. Both results show that our model outperforms the other models.

II. RELATED WORK

Deep learning is developing faster and faster in recent years, and many powerful deep learning models have emerged in the field of natural language processing, such as the seq2seq models. Research on single-turn conversational response generation have achieved a great success. Many researchers work on the coherence, diversity and personality of response [10] [11]. Recently, multi-turn response generation have attracted more and more attention from academia. [8] presented the HRED model to get hierarchy information. Based on the HRED model, [12] further proposed a hierarchical latent variable encoder-decoder model, which introduced a Gaussian random variable to improve the diversity of the response.

In the field of machine translation, attention mechanism was first applied [13]. Then, they are widely used in dialogue systems soon. [14] proposed a hierarchical recurrent attention network. In this model, attention mechanisms are used at both utterance level and word level, which was the first application of the hierarchical attention used in dialogue systems. [9] explicitly weights context vectors. [6] proposed two different types of attention mechanisms, one is dynamic attention and the other is static attention.

[4] proposed a particular attention mechanism called multi-head attention mechanism, in which self-attention is widely used in encoders and decoders. In this work, we use the particular attention mechanism in utterance encoding, making the utterance representation more effective.

III. LAST UTTERANCE-CONTEXT ATTENTION MODEL

In our work, we proposed a last utterance-context attention model, whose framework is shown in Figure 1. In the proposed model, we firstly apply multi-head self-attention mechanism at the word level in each utterance, and then encode each utterance as hidden vectors using a recurrent neural network (GRU). When decoding, we pay attention to both each word in the last utterance and all the contextual utterances. For the last utterance, a hidden vector

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00057

is calculated depending on each word and the decoder hidden states. For the contextual utterances, a hidden vector is calculated depending on representation of each utterance and the decoder hidden states. Finally, the two hidden vectors are concatenated as a context vector for decoding the response.

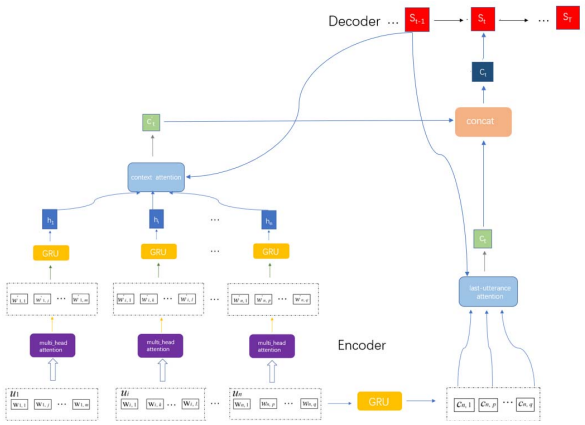

Fig. 1. The proposed last utterance-context attention model

A. Utterance Encoding using Multi-Head Self-Attention

Attention mechanisms can be employed so that the weights of the different words can be calculated. We note that multi-head self-attention mechanism [4] has achieved a great success in machine translation task, so that this method is employed in the proposed model to help encode utterances.

1) Multi-Head Attention Mechanism

The particular attention mechanism in multi-head is called scaled dot-product attention. The mechanism takes three vectors as input, called queries, keys and values. The dimension of them are d_k , d_k and d_v. To obtain the weights on the values, firstly, take the dot product on query and keys, then divide the result by $\sqrt{d_k}$. Finally, a softmax function is used . The matrix of outputs is computed as:

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where $\sqrt{d_k}$ is a scaling factor.

2) Utterance Encoding

Given a conversation that contains several contextual utterances $U = (u_1,...u_i,...u_n)$, and a corresponding response $Y = (y_1,...y_j,...y_m)$, where n is the number of utterances and m is the number of words in the response. For any utterance u_i , $W_i = (w_{i,1},...w_{i,k},...w_{i,p})$ are the words in the utterance, where p is the number of words. Follow the work in [4], we can get:

$$\begin{aligned} w'_{i,k} &= MultiHead(w_{i,k}, w_{i,k}, w_{i,k}) \\ &= Concat(head_1,...,head_h)W^O \end{aligned} \tag{2}$$

where query=key=value, namely self-attention. d_{model} is word vector dimension, $d_k = d_v$ are the hyper-parameters.

Since we have got $W'_i = (w'_{i,1},...w'_{i,k},...w'_{i,p})$, the new word vector can be considered to have the ability to measure the importance of each word. Then, we take each word in order as the GRU input and take the network output in the last step as the representation vector:

$$h_i = GRU(w'_{i,1},...w'_{i,k},...w'_{i,p}) \tag{3}$$

The representations of all the utterances are obtained as the above equation, from which we can get $H = (h_1,...h_i,...h_n)$.

B. Last Utterance-Context Attention

1) Last Utterance Attention

With $W_n = (w_{n,1},...w_{n,k},...w_{n,p})$ denoting the words in the last utterance u_n , the last utterance attention can be calculated by:

$$\begin{aligned} e_{k,t} &= V^T \tanh(W w_{n,k} + U s_{t-1}) \\ \alpha_{k,t} &= \frac{\exp(e_{k,t})}{\sum_k \exp(e_{k,t})} \\ c_t &= \sum_k \alpha_{k,t} w_{n,k} \end{aligned} \tag{4}$$

where t represents each step of decoding, V , W , U are parameter matrices, and S_{t-1} is the decoder hidden state in the time step of t-1.

2) Context Attention

For all the contextual utterances $U = (u_1,...u_i,...u_n)$, we can get their representations by equation (3), and the context attention is calculated by

$$\begin{aligned} e_{i,t} &= V'^T \tanh(W' h_i + U' s_{t-1}) \\ \alpha_{i,t} &= \frac{\exp(e_{i,t})}{\sum_i \exp(e_{i,t})} \\ c'_t &= \sum_k \alpha_{i,t} h_i \end{aligned} \tag{5}$$

where t represents each step of decoding, V' , W' , U' are parameter matrices, and S_{t-1} is the decoder hidden state in the time step of t-1.

3) Concatenating and Decoding

c_t and c'_t are calculated by equation (4) and equation (5) to get the information of the last utterance and the whole context respectively, and then we concatenate can them as a context vector

$$C_t = [c_t; c'_t] \tag{6}$$

and the decoder hidden state in step t can be obtained by

$$s_t = f(y_{t-1}, s_{t-1}, C_t) \tag{7}$$

where y_{t-1} is the decoder output at t-1 step. The output of the decoder in t step can be expressed in terms of conditional probabilities

$$y_t = \arg\max P(y_t) = \prod_{t=1}^{T} p(y_t \mid \{y_1,...y_{t-1}\}, C_t) \tag{8}$$

which can be reduced to:

$$P(y_t | y_{t-1}, y_{t-2}, \ldots y_1, C_t) = g(s_t, y_{t-1}, C_t) \quad (9)$$

IV. EXPERIMENT

A. Experiment Settings

In this paper, two data sets are selected to verify the model. Firstly, the English dataset called DailyDialog is used. Then, to verify the performance in the Chinese context, we also select a Chinese dataset called Douban Conversation Corpus. Many of the existing conversation datasets do not originate from real conversations but from social networks and movie lines. Compared with the previous corpus, the English dataset DailyDialog has less noise and covers several major themes of life. There are more than 13,000 dialogues, averaging 8 rounds per dialog. We split the dataset into three parts, the validation set and testing set are both 1,000, and the rest are training set. The Chinese dataset Douban Conversation Corpus, crawled form Douban group. This dataset is large in size, with more informal abbreviations and Internet slang, which is harmful to the model. Limited to computing resources, we only randomly selected a small part of the dialogue, dataset size and partition is basically equivalent to the English dataset.

In the experiments, the maximum number of dialogues was set to 15, and the dialogues less than 3 rounds were removed. We set the hidden units number to 512, and set the word vector dimension to 300.

B. Baselines

- S2SA: We directly concatenate all the utterances as the input, then the problem is transformed into single-turn. Using [3] as the baseline model.
- HRED: The first to use the hierarchical model to get context information, which is proposed by [8].
- Dynamic Attention: The dynamic attention model proposed by [6].

C. Evaluation and Results

1) Automatic Evaluation

In this paper, BLEU [2] was used as an automatic evaluation index, which is an indicator used to evaluate the difference between the sentences generated by the model (Candidate) and the actual sentences (reference). From Table 1, obviously that our model gets the highest in nearly all BLEU scores.

TABLE I. BLEU SCORES FOR EACH MODEL

Models	DailyDialog			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
S2SA	15.08	6.80	4.89	3.87
HRED	**16.55**	7.92	5.92	4.64
Dynamic Attention	15.89	7.90	6.04	4.73
Last-Context	16.53	**8.35**	**6.25**	**4.78**

Models	Douban Conversation Corpus			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
S2SA	3.00	0.981	0.665	0.560
HRED	4.12	1.16	0.837	0.694
Dynamic Attention	3.42	1.17	0.842	0.713
Last-Context	**4.31**	**1.38**	**0.908**	**0.750**

2) Human Evaluation

How to evaluate the quality of dialog system automatically has always been a difficult problem. Current mainstream evaluation indicators have a variety of defects, and BLEU scores cannot measure the quality of the model generated by the response very accurately. Because of the complexity and diversity of human language, some sentences may not have overlapping vocabulary. They are irrelevant sentences if the context is ignored, but can be used as a response to the same conversational situation. In this case, human evaluation is introduced to further measure the quality of the response generated by the proposed model and each baseline model. Human evaluation indicators used in this paper include coherence and fluency. Coherence measures whether the generated responses are consistent with the context and whether they can connect the conversation. The coherence score ranged from 0 to 2. 0 represents no coherence in the response, which means that it cannot be used as a response to the current dialogue; 1 represents general coherence, meaning that it can be used as a response to the current dialogue; 2 represents high consistence with the current dialogue situation, implying that the answer is very appropriate. Fluency measures whether the generated response is grammatically fluent and error-free. The fluency score ranged from 0 to 1. 0 means that the sentence is not smooth, and there are grammatical errors; the score of 1 means that the sentence is smooth without grammatical errors. For each model in our experiments, 200 test sentences were randomly selected for human evaluation. The evaluator was completely unaware of the content of the experiment. Table 2 shows the evaluation results, from which we can see that the model we proposed achieved the best results in both correlation and fluency.

TABLE II. HUMAN EVALUATION RESULTS

Models	DailyDialog	
	Coherence	Fluency
S2SA	0.695	0.630
HRED	0.735	0.595
Dynamic Attention	0.750	0.655
Last-Context	**0.845**	**0.670**

Models	Douban Conversation Corpus	
	Coherence	Fluency
S2SA	0.335	0.460
HRED	0.360	0.390
Dynamic Attention	0.355	0.530
Last-Context	**0.490**	**0.625**

3) Case Study and discussion

Table 3 shows a few typical cases of our model and the best performing model in the other models. From the first case we can see our model accurately obtains the key information of Barack Obama in the context, while the response generated by the baseline model can also be used as a reply, but obviously lacks the contextual information. In the second case, our model captures the context of appointment with a doctor, and the response is not only fluent but also interesting. Although the baseline model also captures the contextual information of doctor, the reply is not smooth. In the third case, the replies which are generated by our model conform to the roles of the characters and are rich in information. In contrast, the baseline model also falls into the problem of generating general replies.

We also analyze some error cases and compare the performance of the same model on two datasets of the same size. The most common feature of the wrong cases is that they tend to have more dialogue turns, more words or unfamiliar words. In this case, the model is often unable to obtain effective contextual information, or even can't guarantee the fluency of the reply. In addition, based on the comparison of the same model on two datasets, we find that on the Chinese dataset, the performance has great decline compared with on the English dataset. Probably because the Chinese dataset is crawled form the social network with big noise. Moreover, due to the need of word segmentation, the vocabulary is much larger than that in English. However, we find that our model shows great advantages over the baseline models. Fluency, in particular, is similar to that on the English dataset.

In general, this article is a combinatorial attempt to make use of existing techniques, mainly to distinguish the importance of context and the last utterance. Limited by computing resources and model training skills, the experiments in this paper may be a little rough. Given the analysis of the error cases, we plan to improve the performance of the model in two directions in our future work: (1) For too many dialogue turns, selectively forget parts of the content. (2) Consider the treatment of unfamiliar words.

TABLE III. SEVERAL CASES（_ eou_ is used to separate two interlocutors）

Context	Response
I can ' t believe Obama is our President . ___eou__ Did you vote for him , because I know that I did . __eou__ I did , too . __eou__ I have nothing but faith in Obama . __eou__ Really ? __eou__ I ' m sure he will help better America . __eou__ You ' re absolutely right . __eou__	Last-Context：I 'm excited to Barack Obama. Dynamic Attention: I agree with you.
Hi , I need to make an appointment to see the doctor . __eou__ I have openings on Tuesday or Wednesday . Which would be best for you ? __eou__ I need an appointment on Tuesday . __eou__ Fine , I am putting you down for 9 __eou__ I would like to see Dr . Jones . __eou__	Last-Context: It will be long until your schedule. Dynamic Attention: I don't know that will be the doctor. Thank you, Goodbye.
Welcome to Flash Foot , sir. How can I help you ? __eou__ I ' d like extra prints made of these photos . __eou__ What size do you want the prints ? __eou__ Four by six , except this one . I want a ten by thirteen print of this one . __eou__	Last-Context：Okay, they'll be ready for you soon. Dynamic Attention: All right.

V. CONCLUSION

In this paper, a last utterance-context attention model for multi-turn response generation is proposed. In this model, the multi-head self-attention mechanism is used during the utterance encoding stage, in order to highlight the importance of key words in each utterance. In addition, our model takes the last utterance and the contextual information into consideration at the same time. To verify our model, we also conduct several experiments on two datasets. In order to verify the applicability of our model in different language environments, the datasets include both Chinese and English. The evaluation is not only conducted in an automatic way, but also in a manual way. Both results show that our model outperforms the other models.

ACKNOWLEDGMENT

The work described in this paper was fully supported by a grant from the National Key R&D Program of China (No. 2018YFC1603303).

REFERENCES

[1] Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. CoRR, abs/1605.07683.

[2] Kishore Papineni, Salim Roukos, Todd Ward, et al. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA., pages 311–318.

[3] Shang, Lu, and Hang. 2015. Neural responding machine for short-text conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pages 1577–1586.

[4] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. 2017.

[5] Iulian Vlad Serban, Tim Klinger, Gerald Tesauro,et al. 2016. Multiresolution recurrent neural networks: An application to dialogue response generation. arXiv preprint arXiv:1606.00776.

[6] Zhang, W., Cui, Y., Wang, Y., et al. (2018, August). Context-Sensitive Generation of Open-Domain Conversational Responses. In Proceedings of the 27th International Conference on Computational Linguistics(pp. 2437-2447).

[7] Rui Yan, Yiping Song, and Hua Wu. 2016. 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. pages 55–64.

[8] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, et al 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16).

[9] Tian, Z., Yan, R., Mou, L., et al. (2017, July). How to make context more useful? an empirical study on context-aware neural conversational models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 231-236).

[10] Chen Xing, Wei Wu, Yu Wu, et al . 2016. Topic aware neural response generation. arXiv preprint arXiv:1606.08340.

[11] Jiwei Li, Michel Galley, Chris Brockett, et al. 2016. A persona-based neural conversation model. arXiv preprint arXiv:1603.06155.

[12] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, et al. 2016b. A hierarchical latent variable encoder-decoder model for generating dialogues. arXiv preprint arXiv:1605.06069.

[13] Dzmitry Bahdanau,Kyunghyun Cho,and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[14] Chen Xing, Wei Wu, Yu Wu, et al. 2017. Hierarchical recurrent attention network for response generation. arXiv preprint arXiv:1701.07149.