DOI: 10.31534/engmod.2021.2.ri.02d
Original scientific paper
Received: 5.04.2021.

# Multi-turn Dialogue Model Based on the Improved Hierarchical Recurrent Attention Network

**Jiawei Miao**(1), **Jiansheng Wu**(1*)

(1) *School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114044, CHINA*
*e-mail: ssewu@163.com*

## SUMMARY

*When considering the multi-turn dialogue systems, the model needs to generate a natural and contextual response. At present, HRAN, one of the most advanced models for multi-turn dialogue problems, uses a hierarchical recurrent encoder-decoder combined with a hierarchical attention mechanism. However, for complex conversations, the traditional attention-based RNN does not fully understand the context, which results in attention to the wrong context that generates irrelevant responses. To solve this problem, we proposed an improved hierarchical recurrent attention network, a self-attention network (HSAN), instead of RNN, to learn word representations and utterances representations. Empirical studies on both Chinese and English datasets show that the proposed model has achieved significant improvement.*

**KEYWORDS:** *Multi-turn dialogue; hierarchical neural network; attention mechanism; self-attention.*

## 1. INTRODUCTION

Dialogue systems exist in two main forms: (1) task-oriented dialogue systems, which is very robust at complete specific tasks in the vertical domain [1-5], and (2) non-task-oriented dialogue systems that generate human-like responses in the open domain [6-9], namely chat-bots. It not only imitates human dialog in all the facets like answer questions about movies, music, travel, and other fields, but it also completes complex tasks like booking restaurant reservations [10], movie tickets, etc. Because of its high practical value, the research of dialogue systems has been widely paid attention to by academics and industry. With the explosive growth of data, a large conversational corpus is available [11], the focus of academic circles has gradually shifted to data-driven open domain dialogue systems.

Over the past years, the neural-network-based approach became popular [12]. Researchers have extensively used neural networks based on the Seq2Seq framework [13] to explore fully data-driven open-domain dialogue systems. The Seq2Seq framework, which uses end-to-end training, has become an increasingly popular technique for open-domain dialogue generation.

However, these seq2seq-based models work well only for the single-turn dialogue system but are still quite difficult for the multi-turn dialogue systems.

Recently, many research efforts have been devoted to generating more contextual and appropriate responses in the multi-turn dialogue generation. Serban et al. [14] extended a hierarchical recurrent encoder-decoder framework and proposed one of the most popular models (HRED), which modeled the hierarchy of context. Referring to the idea of applying the attention mechanism in single-turn dialogues [15], Xing et al. [16] extended HRED by introducing the attention mechanism and named as the hierarchical recurrent attention network (HRAN) model. HRAN not only models the hierarchy of the context but also focuses on the important parts used selectively to generate a proper response. An empirical study on Chinese dialogue datasets shows that the HRAN model has improved in performance compared with HRED. But it has the following two shortcomings:

- Low computational efficiency: When the attention mechanism based on RNN carries on the sequence alignment, the execution computation is serial and cannot be parallel. It cannot effectively utilize the accelerated computation hardware resources. However, on large data sets, computational efficiency is crucial and meaningful.
- Generate irrelevant response: The traditional attention mechanism is realized based on RNN [17-19]. Due to the characteristics of the RNN structure that the later sequence contains more information when the attention calculates the weight, the later sequence is given a larger weight, resulting in more attention to the situation of close distance. However, in the multi-turn dialogue generation task, some relevant contexts are far away from the response. Therefore, multi-turn dialogue needs to select and use the important part of the context correctly and generate the context-related response.

We suggest a hierarchical self-attention network to solve the above shortcomings. The improvement of the HSAN model is based on the HRAN and built in a hierarchical structure. HSAN model includes a word-level encoder, word-level attention, utterance-level encoder, utterance-level attention, and decoder. Rather than using RNN to encode utterance and context in the word-level encoder and utterance-level encoder, a multi-head self-attention [20] is employed to learn utterance representations and context representations.

Firstly, the word-level encoder using a multi-head self-attention network is employed to acquire more informative word representations. Secondly, the important words are attended to, and the utterances vector is generated, using the word-level attention. The utterance vector is taken as the utterance level encoder input. Thirdly, the utterance-level encoder which uses a multi-head self-attention network encodes utterances and obtains utterances representations. Then the context vector is obtained through the utterance level attention. Finally, the decoder uses the previously predicted words and the context vector as a guide to decoding, generating the predicted words one at a time.

In short, our main contributions are as follows:

- Our proposed model considers how to select and use the dialogue context correctly. The attention performance can be improved by using the multi-head self-attention network instead of the recurrent neural networks. It is the first attempt to replace RNN coding with the self-attention network in a hierarchical attention structure which concentrates on a multi-turn dialogue system.

- Empirical studies of automatic evaluation and human judgment in both Chinese and English datasets show that our proposed HSAN model outperforms the baselines and produces an appropriate, informative response.

## 2. RELATED WORK

The original idea of implement an open domain dialogue system was to think of response generation as machine translation. Inspired by the successful application in machine translation, researchers gradually adopted Neural Machine Translation (NMT) model [19] for dialogue generation. Shang et al. [15] proposed a Neural Responding Machine (NRM) for short-text conversation and trained on about 400,000 responses from Weibo. As you can see from the experiment results, the sequential model is effective in the dialogue task, which sets off a boom in the research of constructing an open domain generative dialog system using deep learning technology. Vinyals et al. [21] applied the encoder-decoder structure of the Seq2Seq to the open domain response generation task in 2015. All the above work is about the study of single-turn dialogues, but they ignored that the hierarchy of the context is vital to response generation. Therefore, Sordoni et al. [22], considering the influence of conversational history, proposed a DCGM model to encode context information. Serban et al. [14] proposed the HRED model to simulate the hierarchical structure of the context by referring to the hierarchical neural network, which encodes the semantics inside sentences and the context semantics between sentences, respectively. The HRED model focuses on the hierarchy of the context but does not concern the impact of important parts of the context on the response, which introduces noise into the model and loss of important information in the context, resulting in irrelevant responses. Xing et al. [16] extended the structure of HRED by introduced the attention mechanism applied to single-turn dialogue to the multi-turn domain and proposed a Hierarchical Recurrent Attention Network (HRAN) model.

Inspired by those who focused their attention on the target area when studying, researchers proposed an attention mechanism. Then Bahdanau et al. [19] apply it to the NLP field, and then researchers rapidly employed it to single-turn dialogue system. Xing et al. [16] adopted a hierarchical attention network in multi-turn dialogue, which simulates the role of words and utterances in dialogue generation. Recently, Vaswani et al. [20] abandoned the traditional Encoder-Decoder model that combines the inherent patterns of CNN or RNN and used only the attention mechanism to construct a sequence coding layer for machine translation tasks. Experiments show that the model can reduce computation and improve parallel efficiency, and self-attention is superior in capturing long-distance dependence.

## 3. METHOD

### 3.1 PROBLEM DEFINITION AND OVERVIEW

Consider a data set $D = \{(C_i\,, Y_i)\}_{i=1}^{T_d}$, in which $C_i$ represents dialogue context and $Y_i$represents the response. Each $C_i = \{U_1, U_2, \ldots, U_M\}$contains a sequence of $M$ utterances. Each utterance $U_m = \{w_{m,1}, w_{m,2}, \ldots, w_{m,N}\}$ consists of $N$ tokens, where *m∈{1,..., M]*. Our goal is to establish a dialogue model that produces informative and coherent responses $Y = \{y_1, y_2, \ldots, y_T\}$ based on dialogue history $C$.

Figure 1 shows the overview of our proposed HSAN model composed of the word-level encoder, word-level attention, utterance-level encoder, utterance-level attention, and decoder. Firstly, a word-level encoder encodes the words of each utterance in the context as hidden states. Secondly, word-level attention is used to attend to the important words in utterances and generate utterances vectors. Then the utterance vectors are taken as the utterance level encoder input. The utterance level encoder using a multi-head self-attention network encodes utterances and gets utterances representations. Finally, the utterance level attention emphasizes the important utterances of the context and encodes them into a context vector. The context vector is used as decoder input, guiding the generation of responses. Next, HSAN details are given.

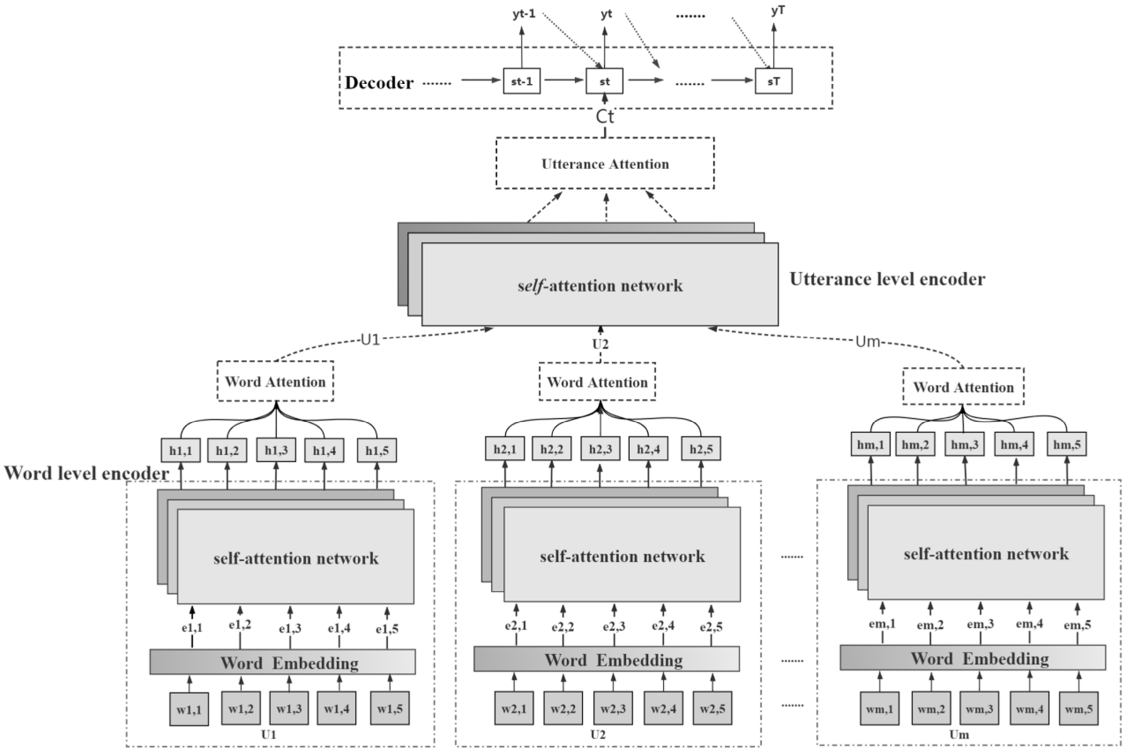

**Fig. 1** *The architecture of Hierarchical Self-Attention Network*

### 3.2 WORD LEVEL ENCODER

The word-level encoder module includes two layers. One is word embedding used to convert utterance from a series of words into a series of low-dimensional embedding vectors. Suppose utterance $U_m$ with $N$-words as $\{w_{m,i}\}_{i=1}^{N}$ and each $U_m$ in $C_i$, *m*∈*{1,..., M]* as input of the word-level encoder. Through the first layer, it is converted into a vector sequence $\{e_{m,i}\}_{i=1}^{N}$.

The other layer used word level multi-head self-attention network [20]. Utterance representations must focus on the important words among the utterance. The experiment shows that the multi-head self-attention network represents better the word representation by capturing word interaction [20]. Thus, we employ the multi-head self-attention network to encode words $\{e_{m,i}\}_{i=1}^{N}$ in $U_m$ as word-level hidden states $\{h_{m,i}\}_{i=1}^{N}$. The *k-th* self-attention head representation of the *i-th* word in $U_m$, $h_{m,i}^{k}$ is given by:

$$\alpha_{i,j}^{k}=\frac{exp\left(e_{m,i}^{T}Q_{k}^{w}e_{m,j}\right)}{\sum_{n=1}^{N}exp\left(e_{m,i}^{T}Q_{k}^{w}e_{m,n}\right)} \quad (1)$$

$$h_{m,i}^{k}=V_{k}^{w}\left(\sum_{j=1}^{N}\alpha_{i,j}^{k}\,e_{j}\right) \quad (2)$$

where $Q_k^w$and $V_k^w$are the trainable parameters in the *k-th* head, and $\alpha_{i,j}^k$ indicates the relative importance of the interaction between the *i-th* and *j-th* words, *j∈{1,..., N]*. The multi-head representation $h_{m,i}$ of the *i-th* word by connecting the representation of *h=8* separate self-attention heads reads:

$$h_{m,i}=Concat\left(h_{m,i}^{1},h_{m,i}^{2},\ldots,h_{m,i}^{h}\right) \quad (3)$$

### 3.3 WORD LEVEL ATTENTION

The words in a sentence may have different effects on expression. Thus, we introduce an attention mechanism to attend to important words in utterance for learning more informative utterance representations and aggregate these important words into an utterance vector. Suppose the decoder has generated *t-1* words at step *t*, according to the step *t-1* decoder hidden state $s_{t-1}$, the word level attention represents utterance $U_m$ as utterance vector $r_m$ by using word-level hidden states $\{h_{m,i}\}_{i=1}^{N}$, formulated as:

$$\gamma_{i,t}^{w}=softmax\left(h_{m,i}^{T}W_{w}s_{t-1}\right) \quad (4)$$

$$r_{m}=\sum_{i=1}^{N}\gamma_{i,t}^{w}h_{m,i} \quad (5)$$

where $\{\gamma_{i,t}^{w}\}_{i=1}^{N}$ expresses the importance of words in utterance $U_m$ , and $W_w$ is a trainable weight matrix.

### 3.4 UTTERANCE LEVEL ENCODER

Utterance vectors $\{r_m\}_{m=1}^{M}$ as the utterance level encoder input also uses the multi-head self-attention network to encode utterances $\{r_m\}_{m=1}^{M}$to $\{u_m\}_{m=1}^{M}$ as hidden vectors of the context. The representation of the *i-th* utterance learned by the *k-th* head is as follows:

$$\beta_{i,j}^{k}=\frac{exp\left(r_{i}^{T}Q_{k}^{u}r_{j}\right)}{\sum_{m=1}^{M}exp\left(r_{i}^{T}Q_{k}^{u}r_{m}\right)} \quad (6)$$

$$u_{i}^{k}=V_{k}^{u}\left(\sum_{j=1}^{M}\beta_{i,j}^{k}\,r_{j}\right) \quad (7)$$

where $Q_u^k$ and $V_u^k$ are the trainable parameters in the *k-th* head, and $\beta_{i,j}^k$ indicates the relative importance of the interaction between the *i-th* and *j-th* utterance, *j∈{1,...N}*. The multi-head representation $u_i$of the *i-th* word by connecting the representation of *h=8* separate self-attention heads reads:

$$u_{i}=Concat\left(u_{i}^{1},u_{i}^{2},\ldots,u_{i}^{h}\right) \quad (8)$$

### 3.5 UTTERANCE LEVEL ENCODER

Different utterances may have different informativeness when expressing the context of a dialogue. Therefore, attention mechanisms are also used to attain utterances that are important for learning contextual expressions. Context vector $c_t$, calculated by utterance level attention, is defined as:

$$\gamma_{i}^{u}=softmax\left(u_{i}^{T}W_{u}s_{t-1}\right) \quad (9)$$

$$c_t = \sum_{i=1}^{M} \gamma_i^u u_i \tag{10}$$

where $\{\gamma_i^u\}_{i=1}^{M}$ measures the importance of utterances in $C_i = \{U_1, U_2, \ldots, U_M\}$ and $W_u$ is a trainable weight matrix.

### 3.5 DECODER

Language models (LM) are trained to calculate that the probability of a sequence of tokens being a linguistic sentence. When generating a sequence based on an LM, we can generate one word at a time based on the previously predicted words. The decoder of HSAN is an RNN language model [23] conditioned on the context vectors $\{c_t\}_{t=1}^{T}$ is given by Eq. (10). Frequently, the probability of predicting the word at step $t$ is as follows:

$$P(y_t|c_t, y_1, \ldots, y_{t-1}) = softmax(\beta_t) \tag{11}$$

where $\beta_t$ is given by:

$$\beta_t \ = linear\{concat(c_t, s_t, e_{y_{t-1}})\} \tag{12}$$

$$s_t \ = f(s_{t-1}, e_{y_{t-1}}) \tag{13}$$

where $f$ function is a GRU [24], the embedding of the step *t-1*'s output $e_{y_{t-1}}$ and the last hidden state $s_{t-1}$ as the input of GRU, give the new decoder state $s_t$. Then, concat $c_t$, $s_t$, $e_{y_{t-1}}$ and map a V-dimensional vector by linear layer, where V represents vocabulary size. Finally, $\beta_t$ through a softmax layer obtains the word probability. The likelihood of every response sequence $Y = \{y_1, \ldots y_T\}$ is computed by:

$$P(Y|C;\theta) = \prod_{t=1}^{T} P(y_t|C, y_1, \ldots, y_{t-1}) \tag{14}$$

We denote θ as the parameter set of HSAN and trained objective to search parameters θ from $D$ by maximize the likelihood of every sentence:

$$\theta = argmax\{L(\theta, D)\} \tag{15}$$

where:

$$L(\theta, D) = \sum_{i=1}^{T_d} P(Y_i|C_i;\theta).$$

## 4. EXPERIMENTS

In this section, we analyze the Chinese and English datasets used. Experiments were performed to evaluate our model and compared with several baseline models.

### 4.1 DATASETS

We used two publicly available multi-turn dialogue datasets, one is DailyDialog [25], an English dialogue dataset between people in daily life, and the other is KdConv [26], Chinese multi-domain knowledge-driven dialogue datasets. DailyDialog contains 11,318 hand-written dialogues which cover a variety of topics in our daily life. KdConv, recently proposed by Zhou et al. [26], contains an in-depth discussion of related topics in three fields: film, music, tourism, and a natural transition between multiple topics. We mainly study the multi-turn dialogue generation in the field of film, so only dialogues of such a topic are selected. In reality, multi-turn dialogues are not limited to one or two topics, but the above datasets are adjusted for multi-turn dialogues generation tasks.

In the experiment, we eliminated the knowledge of KdConv first. As preprocessing, we split each dialogue into conversation pairs, then deleted less than 3 turns and more than 15 turns of dialogue. For KdConv we employed the Jieba Chinese word segmenter for tokenization and removed more than 50 words from sentences. In the end, we randomly divided the data into training, validation, and test sets and obtained 15,483, 1,924, 1,885 pairs in KdConv, and 11,118, 1,000, 1,000 pairs in DailyDialog, respectively.

## 4.2 BASELINES AND IMPLEMENTATION DETAILS

In this experiment, we compared the proposed HSAN model with the following set of models:

- **Attn-Seq2Seq,** a standard Seq2Seq model with attention [19], that is extensively applied in open-domain dialogue generation.
- **HRED**, a hierarchical encoder-decoder model proposed by [14], that is the most basic multi-turn dialogue model.
- **HRAN**, a hierarchical recurrent attention network proposed by [16], that is one of the best models in the multi-turn dialogues systems at this stage.

The parameter setting of the model has an important effect on the experiment. To be specific, both Chinese and English vocabulary have a size of *25,000* and the dimension of word embedding is set to *200*. During the decoding process, the special UNK token indicates that the generated words are not in the vocabulary. The batch size is set to *32*, and all multi-head self-attention networks with *512* hidden states and heads=*8*. The dropout method [27] is used in the experimental training of this paper to prevent the overfitting of parameters with the dropout rate set to *0.3*. And gradient clipping is used to prevent gradient explosion. Adam optimizer is used in the optimization process. During the training process, the dynamic learning rate method is adopted, and the initial learning rate is set to *0.001*. When the perplexity (PPL) of the valid set did not improve in *3* consecutive times, the learning rate was reduced by *10%*. We run all models on a GTX 1070 Ti machine.

## 4.3 EVALUATION MEASURES

Evaluating non-task-driven dialogue systems is still an open question [28]. Now, there is no mature automatic evaluation method for generative dialogues systems, and there is no unified standard for automatic evaluation of generated. According to the problems to be solved, different automatic evaluation metrics are selected [29]. Xing et al. [16] mentioned that due to the diversity of responses, BLEU [30] is not suitable for an evaluation metric, so it is not used in this paper.

### 4.3.1 AUTOMATIC EVALUATION

Due to the strong generalization ability, perplexity [31] has been proposed previously to evaluate whether the generation results of the generative dialogue models are grammatical and fluent. For automatic evaluation, perplexity was used as the evaluation metric. The lower perplexity of the language model, the higher the probability of expected statement appearance, which indicates better generation performance. During the training process, we referred to the perplexity on the validation set to determine when to stop the training. It is considered that the training has reached convergence and terminated if the perplexity of *10* consecutive turns does not decrease in the valid set.

#### 4.3.2 HUMAN JUDGEMENT

For human judgment, we randomly sampled 300 contexts from the test datasets and generated responses for each model. Three annotators (all students majoring in NLP) were asked to compare the HSAN model and baselines, grading with a win, loss, tie. The win indicates that the HSAN model is more relevant, logically consistent with the context, and fluent than the baseline model; the loss that the baseline model works better; and the tie indicates that it is impossible to judge which one is better.

### 4.4 EXPERIMENTAL RESULTS

The perplexity results of each model shown in Table 1, and all models achieved the lowest perplexity in validation and test. The smaller the PPL, the closer the generated response is to the standard response. And you can see from the results, the HSAN model outperforms all baselines on perplexity. The Attn-Seq2Seq model always generates short and meaningless responses because it does not consider the structure of the conversational context. The HRED model produces a more flexible response. By analyzing the result of the HRED model, we can see that it is important to model the hierarchy of context for response generation. HRAN models the hierarchy of the context focusing on the important parts of the context, which generates more relevant responses than HRED. Compared to the above-mentioned models, the HSAN model not only understands the meaning of the conversation context but also selects correct contextual information and generates more proper and informative responses. A significance test was carried out for the perplexity and the results showed that the improvement of HSAN was statistically significant ($p$-value < *0.01*).

**Table 1** *Perplexity results*

| *Model* | *DailyDialog* | | *KdConv* | |
|---|---|---|---|---|
| | *Validation perplexity* | *Test perplexity* | *Validation perplexity* | *Test perplexity* |
| *Attn-Seq2Seq* | *49.17* | *51.94* | *40.58* | *37.42* |
| *HRED* | *47.40* | *49.86* | *37.48* | *34.26* |
| *HRAN* | *45.32* | *47.94* | *36.23* | *32.08* |
| *HSAN* | ***42.23*** | ***44.09*** | ***34.39*** | ***30.14*** |

The human evaluation results are shown in Tables 2 and 3. We can see from the results that the winning score is always greater than loss, indicating that the performance of the HSAN is superior to all baselines. KdConv, for example, compared with Attn-Seq2Seq, HRED, HRAN, the HSAN model achieves a preference of *12.68%*, *11.8%*, and *7.61%*, respectively. In this paper, Kappa values [32] are used to verify the consistency among the annotators, and the results show that they have reached a relatively high agreement in judgment. We also performed significance tests, and the results show that the improvement of our model is meaningful on both the Chinese and English datasets ($p$-value <*0.01*).

**Table 2** *Human evaluation results of KdConv datasets*

| *Model* | *Win(%)* | *Loss(%)* | *Tie(%)* | *kappa* |
|---|---|---|---|---|
| *HSAN vs. Attn-Seq2Seq* | *35.70* | *23.02* | *41.28* | *0.43* |
| *HSAN vs. HRED* | *34.98* | *23.18* | *41.84* | *0.41* |
| *HSAN vs. HRAN* | *32.40* | *24.79* | *42.81* | *0.42* |

**Table 3** *Human evaluation results of DailyDialog datasets*

| *Model* | *Win(%)* | *Loss(%)* | *Tie(%)* | *kappa* |
|---|---|---|---|---|
| *HSAN vs. Attn-Seq2Seq* | *34.16* | *18.56* | *47.28* | *0.44* |
| *HSAN vs. HRED* | *30.48* | *19.89* | *49.63* | *0.41* |
| *HSAN vs. HRAN* | *31.43* | *22.77* | *45.80* | *0.42* |

## 4.5 CASE STUDY

To further analyze the model performance, we list some cases with large turns on the test set as shown in Tables 4, 5 to compare the effect of the HSAN model with the baseline model on multi-turn dialogue. In Example 1, The Attn-Seq2Seq generates illogical responses, ex. “*I know it's a Japanese movie*”. For the HRED model, it is easy to generate some common responses, such as “*I don't know about that*’”. The HRAN model can bring up a new topic, ex. “*But I do know where he is from?*”, but the new topic will ignore the history of the conversation, ex. ‘*Do you know where he's from?*’. HSAN can take the initiative to change the topic when the long conversation is coming to an end, ex. “*I don't know, but his movies are very good. Do you like movies?*’”, and the topic discussed in the context will not be brought up again, preventing the repetition of dialogue.

**Table 4** *Example 1*

u1：知道新海诚吗？

Do you know *Makoto Shinkai*?

u2：是的，他是一名日本人,是个很出名的导演。

Yes, He is a Japanese and a famous director.

u1：他也是一个作家，你知道他是哪里人吗？

He's also a writer. Do you know where he's from?

u2：他是日本长野县南佐久郡小海町出生的。

He was born in Koachi, Minamisaka, Nagano Prefecture, Japan.

u1：看过他的电影《你的名字》吗？

Have you seen his movie *Your Name*?

u2：看过，这个电影很火的，不过我更喜欢《追逐繁星的孩子》。

Yes, it was a hit, but I liked *Children Who Chase Lost Voices from Deep Below*.

u1：这个动画我也很喜欢看，它是哪年上映的啊？

I also like this cartoon very much. When was it released?

u2：它是2016年08月26日上映的.

It was released on August 26, 2016.

u1：谁配音的啊？

That's great. It's so popular. Who voices it?

u2：上白石萌音、神木隆之介等人。这个电影有没有得奖啊？

Mone Kamishiraishi and Kamiki Ryunosuke et al. Did the movie win any awards?

Attn-Seq2Seq：这个就不知道了， 不过我知道是在日本电影 。

I don't know about this, but I know it's a Japanese movie.

HRAN：这个我就不清楚了。

I don't know about that.

HRED： 这个我不知道，不过我知道他是哪里人啊？

I don't know that, but I do know where he is from?

HSAN：这个我就不清楚了，不过他的电影很好看，你喜欢看电影吗？

I don't know, but his movies are very good. Do you like movies?

In Example 2, the HSAN model is able to find relevant context and understand what it really means, ex. ‘This animation doesn't last long, does it?”, and give appropriate answers, such as “141 minutes. What do you think?”. However, baselines do not understand conversational questions and give an irrelevant or meaningless response.

**Table 5** *Example 2. The bold part represents the relevant context of the response*

u1：你看过《龙猫》吗？

Have you seen *My Neighbor Totoro*?

u2：看过，这是 2018 年 12 月 14 日在中国公映的。

Yes, it was released in China on December 14,2018.

u1：那在日本是哪年上映的呢？

What year was it released in Japan?

u2：1988 年 04 月 16 日。

April 16, 1988.

u1：时隔这么久呢，这个电影又叫做《邻家的豆豆龙》。

After all this time, the movie is called The *Beanbeandragon Next Door*.

u2：这个动画的片长不长吧？

This animation doesn't last long, does it?

Attn-Seq2Seq: 是的，我喜欢看。Yes, I like it.

HRAN： 是的，它的导演是谁啊？Yes, who is its director?

HRED： 是的，是在日本的动画。Yes, it's animation in Japan.

HSAN：片长是 141 分钟，你觉得怎么样？141 minutes. What do you think?

## 5. CONCLUSIONS

In this paper, we aimed to solve the defect of traditional attention for long-distance situations capture. We propose a novel multi-turn dialogue generation model to find relevant context, named HSAN. Our model enhances the capture ability of long-distance situations by introducing advanced technology self-attention network. Empirical results on Chinese and English datasets show that the HSAN model outperformed the most popular models. One epoch training time of the HSAN model is reduced by half compared with the HRAN model.

However, it is found from the test results of the KdConv datasets that most of the responses do not match the facts when the conversation involves specific facts. One of the reasons for this phenomenon is that the model lacks knowledge. Future research should be introducing knowledge to avoid irrelevant responses and ensure more accurate ones without sacrificing efficiency and performance.

## 6. ACKNOWLEDGMENT

The research work was supported by the National Natural Science Foundation of China (No.51774179).

## 7. REFERENCES

[1] S. Young, M. Gasic, B. Thomson and J. Williams, POMDP-Based Statistical Spoken Dialog Systems: A Review, Proceedings of the IEEE, Vol. 101, No. 5, pp. 1160-1179, 2013. https://doi.org/10.1109/JPROC.2012.2225812

[2] M. Henderson, B. Thomson, S. Young, Word based dialog state tracking with recurrent neural networks, Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 292–299, 2014. https://doi.org/10.3115/v1/W14-4340

[3] M. Henderson, Machine learning for dialog state tracking: A review, Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing, 2015.

[4] Y.N. Chen, D. Hakkani-Tur, G. Tur, J. Gao, and L. Deng, End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding, Proceedings of The 17th Annual Meeting of the International Speech Communication Association, pp. 3245–3249, 2016. https://doi.org/10.21437/Interspeech.2016-312

[5] F. Strub, H.D. Vries, J. Mary, B. Piot, A.C. Courville and O. Pietquin, End-to-end optimization of goal-driven and visually grounded dialogue systems, Proceedings of the 26th International Joint Conference on Artifificial Intelligence (IJCAI), pp. 2765–2771, 2017. https://doi.org/10.24963/ijcai.2017/385

[6] A. Ritter, C. Cherry, W.B. Dolan, Data-driven response generation in social media, Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp. 583-593, 2011.

[7] R.E. Banchs, H. Li, IRIS: a chat-oriented dialogue system based on the vector space model, Proceedings of the Association for Computational Linguistics, System Demonstrations, pp. 37–42, 2012.

[8] J.Gu, Z. Lu, H. Li and V.O. Li, Incorporating copying mechanism in sequence-to sequence learning, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1631‑1640, 2016. https://doi.org/10.18653/v1/P16-1154

[9] S. He, C. Liu, K. Liu and J. Zhao, Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning, ACL, Vol. 1, pp. 199‑208, 2017. https://doi.org/10.18653/v1/P17-1019

[10] Ł. Kaiser, S. Bengio, Can active memory replace attention?, Advances in Neural Information Processing Systems, pp. 3781‑3789, 2016.

[11] R. Lowe, N. Pow, I. Serban and J. Pineau, The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 285‑294, 2015.

https://doi.org/10.18653/v1/W15-4640

[12] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, A neural probabilistic language model, *Journal of machine learning research*, pp. 1137-1155, 2003.

[13] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Advances in neural information processing systems, 2014.

[14] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, Proceedings 30th AAAI Conference Artificial Intelligence, 2016.

[15] L. Shang, Z. Lu and H. Li, Neural responding machine for short-text conversation, arXiv preprint arXiv:1503.02364, 2015. https://doi.org/10.3115/v1/P15-1152

[16] C. Xing, Y. Wu, W. Wu, Y. Huang and M. Zhou, Hierarchical recurrent attention network for response generation, 23rd AAAI Conference Artificial Intelligence, 2018.

[17] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, pp. 107‑116, 1998. https://doi.org/10.1142/S0218488598000094

[18] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural computation, pp. 1735‑1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[19] D. Bahdanau, K. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, Computer Science, 2014.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems, pp. 6000‑6010, 2017.

[21] O. Vinyals and Q. Le, A neural conversational model, arXiv preprint arXiv:1506.05869, 2015.

[22] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.Y. Nie, J. Gao and B. Dolan, A neural network approach to context-sensitive generation of conversational responses, arXiv preprint arXiv:1506.06714, 2015. https://doi.org/10.3115/v1/N15-1020

[23] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky and S. Khudanpur, Recurrent neural network based language model, 11th Annuals Conference International Speech Communication Association, pp. 1045-1048, 2010.

https://doi.org/10.1109/ICASSP.2011.5947611

[24] K. Cho et al., Learning phrase representations using rnn encoder-decoder for statistical machine translation Computer Science, 2014.

https://doi.org/10.3115/v1/D14-1179

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

[26] H. Zhou, C. Zheng, K. Huang et al., KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation, Proceedings 58th Annuals Meeting Association for Computational Linguistics, 2020.

https://doi.org/10.18653/v1/2020.acl-main.635

[27] N. Srivastava, G. Hinton, A. Krizhevsky et al., Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929-1958, 2014.

[28] W.N. Zhang, Y.Z. Zhang, T. Liu, A review of dialogue system evaluation methods, Science in China: Information Science, Vol. 047, No. 008, pp. 953-966, 2017.

https://doi.org/10.1360/N112017-00125

[29] C.W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin and J. Pineau, How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, arXiv 2016. arXiv:1603.08023.

https://doi.org/10.18653/v1/D16-1230

[30] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, Bleu: a method for automatic evaluation of machine translation, Proceedings 40th Annuals meeting on association for computational linguistics, pp. 311-318, 2002.

https://doi.org/10.3115/1073083.1073135

[31] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville and J. Pineau, Hierarchical neural network generative models for movie dialogues, arXiv preprint arXiv:1507.04808, Vol. 7, No. 8, 2015.

[32] J.L. Fleiss, Measuring nominal scale agreement among many raters, American Psychological Association, 1971. https://doi.org/10.1037/h0031619