Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Smarter Response with Proactive Suggestion: A New Generative Neural Conversation Paradigm

Rui Yan[1,2], **Dongyan Zhao**[1,2]
[1] Institute of Computer Science and Technology (ICST), Peking University
[2] Beijing Institute of Big Data Research
{ruiyan, zhaody}@pku.edu.cn

Abstract

Conversational systems are becoming more and more promising by playing an important role in human-computer communications. A conversational system is supposed to be intelligent to enable human-like interactions. The long-term goal of smart human-computer conversations is challenging and heavily driven by data. Thanks to the prosperity of Web 2.0, a large volume of conversational data become available to establish human-computer conversational systems. Given a human issued message, namely a $query$, a traditional conversational system would provide a $response$ after proper training of how to respond like humans. In this paper, we propose a new paradigm for neural generative conversations: smarter $response$ with a $suggestion$ is provided given the $query$. We assume that the new conversation mode which proactively introduces contents as next utterances, keeping user actively engaged. To address the task, we propose a novel integrated model to handle both the response generation and the suggestion generation. From the experimental results, we verify the effectiveness of the new neural generative conversation paradigm.

1 Introduction

Generally speaking, talking to each other in human languages is one of the fundamental ways to communicate. Computers are powerful tools, actually as closely connected partners to us in the modern world. Researchers would expect that people can directly interact with computers using natural languages. The conversational user interface is simple, easy, straightforward, and ideally behaves like humans. During the past few years, human-computer conversational systems have been attracting more and more attention due to their functional, social, and entertainment roles in the AI era.

Building an intelligent human-computer conversational system is extremely challenging, and requires extensive analysis of natural languages. After proper understanding about human utterances, a conversational system needs to respond accordingly, synthesizing responses by organizing terms, known as a generation-based system; or the system can retrieve existing responses from a pre-collected conversational data repository, namely a retrieval-based systems. Huge efforts are devoted on how to maintain relevant, meaningful and user-engaging conversations as human-computer interactions in natural languages.

Thanks to the prosperity of Web 2.0! People are willing to have conversations on public websites, such as online forums or social media: such information repository provides a great opportunity to establish an immense data collection of human conversations [Wang *et al.*, 2013; Yan *et al.*, 2016a]. The big repository accelerates the fast development of data-driven technology for conversational research. With the help of human conversational data, an intelligent system would eventually learn how to converse. Here we concentrate on research in the generation-based human-computer conversational system in the open domain, known as non-task-oriented "*chatbots*" [Yan *et al.*, 2016a; 2016b].

For all these years, people have formulated a well-defined paradigm for mainstream chatbot systems. By taking a human utterance as the *query*, the computer generates a *response*. A traditional chatbot system presumes that humans will take the initiative role, and computers need only to "respond" [Li *et al.*, 2016d; Yoshino and Kawahara, 2015]. Such a process is typically regarded as "passive".[1]

Note that query suggestion is a successful task in Information Retrieval. What if we introduce the concept to generative conversations by providing a $response$ as well as a $suggestion$ given a $query$? The suggestion can be used as a next utterance. In general, "suggestions" bring information from an "external" scope and provide additional contents based on the entire corpus. Will the new generative neural conversation paradigm be effective for human-computer conversations? Probably yes. We will investigate the results in experiments.

Given a query q, the conversational system provides a pair of response and suggestion (r,s). r is to respond q, while s is to suggest how the conversation goes on given r and q. In contrast to the traditional q-r conversations, the system now provides additional engagements, which is a "proactive"

[1] Task-oriented systems are sometimes mixed-initiative since they aim to finish tasks, but for current chatbot systems/applications, they are almost all human-initiative.

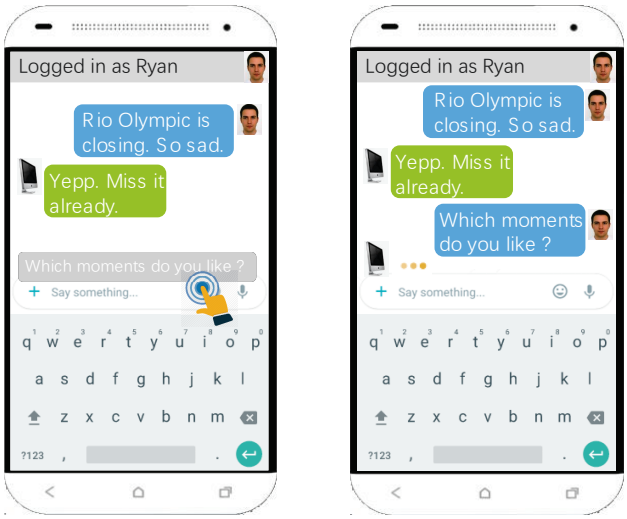

Figure 1: A proactive conversational system provides *responses* and *suggestions* in pairs given the *query*. Users click the suggestion (in gray) if agree to continue the conversation as suggested.

conversation style. We show how it works in Figure 1.

We propose a deep fusion recurrent neural network with gated recurrent units (GRU) for the proposed conversation paradigm. The model first generates the response given the query in a traditional way (sequence-to-sequence [Sutskever *et al.*, 2014]). Moreover, given the query and the generated response, the model integrates the information from both parts in a deep fusion way via dual GRU cells, namely Deep Dual Fusion Model. The model formulates two generation processes for query $\rightarrow$ response and response $\rightarrow$ suggestion conditioned on the query information. The model couples both parts in an end-to-end joint learnable manner.

To sum up, we have manifold contributions as follows:

• We are the 1st to investigate the generative conversational model featured with smarter responses and proactive suggestions. The proposed formulation for generative neural conversational systems is novel.

• We propose a Deep-Dual-Fusion Model based on the gated recurrent units with enhanced cells. The model framework includes a sequence-to-sequence model and a dual sequence-to-sequence model so as to generate the response and suggestion conditioned on the query information.

We conduct extensive experiments in a variety of human-computer conversation setups and evaluate the performance with automatic evaluation metrics and human judgments. In particular, we build a system upon a large conversation resource. We run experiments against several rival algorithms to verify if the new task useful and is the model effective? The experimental results are positive.

The rest of the paper is organized as follows. We introduce the proposed task in Section 2. The model details are elaborated in Section 3. We conduct experimental setups and investigate evaluations against a series of baselines and discuss results in Section 4. Related work is reviewed in Section 5 while the conclusion is drawn in Section 6.

2 Task Statement

2.1 Problem Formulation

The conversation task contains two hops of generations. Given a user utterance as a query q, the system would generate a response r to respond. For the generated r, the system will generate a suggestion s as the next utterance. Since the generated s should be related to the original query q, we formulate them as triples, each as (q,r,s).

Note that conversations have either 1) a single-turn or 2) multiple turns. For multi-turn conversations with preceding utterances before the query, known as the **context**, we concatenate the context sentences with the query to get a reformulated one, still denoted by q without loss of generality [Tian *et al.*, 2017]. Our proposed framework is compatible for both single-turn and multi-turn conversation scenarios.

Consecutive conversational utterances can be used for model training, with the last two utterances as r and s to train. During test time, the generator finds the most likely sequence via beam search through a softmax function. For the response and suggestion generation, we have the following objectives:

$$r^{\star} = \underset{r}{\operatorname{argmax}}\ \mathrm{p}(r|q) \tag{1}$$

$$s^{\star} = \underset{s}{\operatorname{argmax}}\ \mathrm{p}(s|q,r) \tag{2}$$

We adopt beam search algorithm similar as in machine translation systems. The size of beam is empirically tuned in our experiments for speedup consideration. Beam search is ended until the end-of-sentence symbol (*eos*) is generated.

2.2 Model Overview

We can decompose the whole task into two steps, and then analyze the major technical issues for each of them. The model overview is illustrated in Figure 2.

• RESPONSE GENERATION. Given an issued query **q**, we learn to generate a candidate response **r** from a collection of the conversational samples using Equation (1).

It is a key issue to generate a good candidate response given the query. If a generated response is not related to the query, the entire task might become meaningless. Response generation characterizes a standard sequence-to-sequence process from the query to a response. Here we apply a recurrent neural network structure with gated recurrent units (GRU).

• PROACTIVE SUGGESTION. Given a generated response **r** as well as the original (or reformulated) query (with contexts) **q**, we generate a suggestion **s** using Equation (2).

Note that the generation process here is different. Since we have two sequences available for information encoding, we literally need to have a dual sequence-to-sequence generation process for information fusing: the suggestion generation should not be isolated from the input query. We manage to record the hidden states of the query and hidden states of the generated response as "memories" of external information, and the dual sequence-to-sequence process will be influenced by the external memories where the memory is a soft gating mechanism for the information propagation.

To be more specific, for the dual sequence-to-sequence model, we propose three cells for each generation unit. One

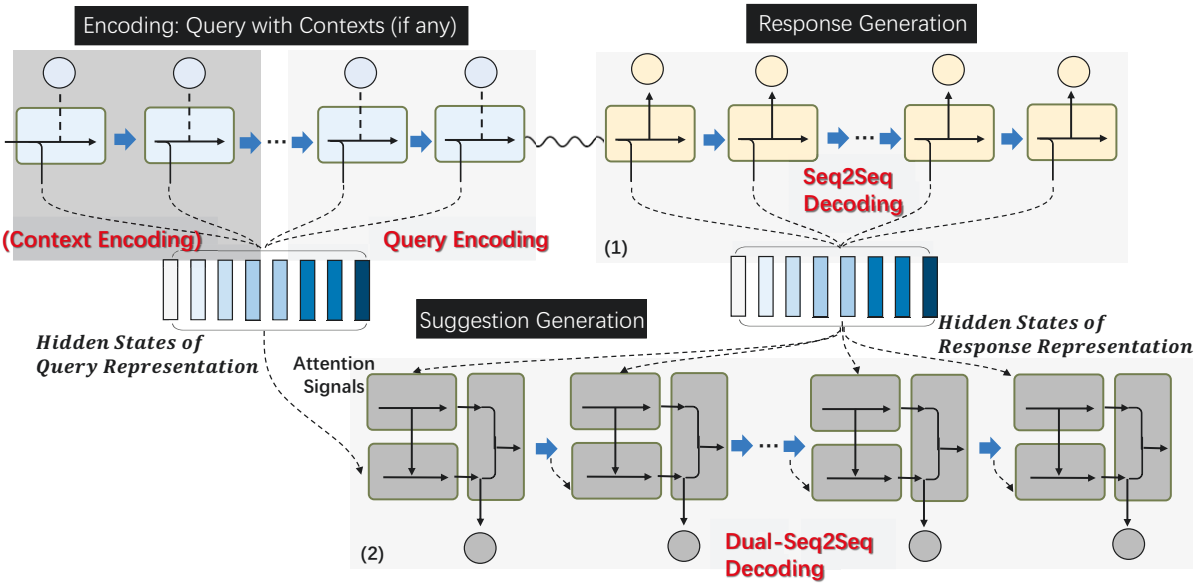

Figure 2: Model overview. The whole framework includes 1) query encoding, 2) response decoding, and 3) suggestion decoding.

cell captures the generation process from the decoded words within the generated suggestion sentence by attention to the memories from the response sequence, namely Sequential GRU Cell. The second cell aligns the hidden state from the Sequential GRU Cell with the memory block from the query which stores the history information, namely Alignment GRU Cell. The last cell is the fusion cell, which combines the output hidden states from the previous two cells together in a deep fusing manner. At each time step, a word of the suggestion will be decoded by the last cell of the deep fusion units. We will introduce the details in the next section.

The model includes two parts. If we omit the dual sequence-to-sequence part, the model degenerates to a traditional generation method for the human-computer conversation without proactive suggestion. In other words, a neural generative model for the standard human-computer conversation is part of our model.

3 Deep Dual Fusion Model

Response generation is a standard sequence-to-sequence process to generate responses given the input query. The key component of the proposed framework is the suggestion component which aims at proactive suggestion to generate next utterances. Due to strict page limits, we skip the description of standard sequence-to-sequence model [Sutskever *et al.*, 2014] for response generation.

The key challenge for proactive suggestion is to generate next utterances given the generated responses while the information from the query part should be taken into account. In this way, the information from three components, i.e., input, response, and suggestion, can be fused together sufficiently so that the conversation keeps in line. To this end, we propose the deep dual fusion units. The proposed units are illustrated in Figure 3. Each unit contains three cells: 1) Sequential GRU Cell, 2) Alignment GRU Cell, and 3) Fusion Cell.

3.1 Sequential GRU Cell

For this component, we apply the GRU-based sequential generation with the attention mechanism at the decoder part. Let $\mathbf{h}^{(1)}_{\mathbf{t-1}}$ be the last hidden state of the Sequential GRU Cell, $\mathbf{y_{t-1}}$ be the embedding of the last generated word for the next utterance generation process, and $\mathbf{c}^{(1)}_{\mathbf{t}}$ be the current attention-based context for this cell. The current hidden state of the sequential GRU decoding, $\mathbf{h}^{(1)}_{\mathbf{y}}$, is defined as follows:

$$\begin{aligned}
\mathbf{h}^{(1)}_{\mathbf{t}} &= (\mathbf{I} - \mathbf{z}^{(1)}_{\mathbf{t}}) \odot \mathbf{h}^{(1)}_{\mathbf{t-1}} + \mathbf{z}^{(1)}_{\mathbf{t}} \odot \hat{\mathbf{h}}^{(1)}_{\mathbf{t}} \\
\mathbf{z}^{(1)}_{\mathbf{t}} &= \sigma(\mathbf{W_z y_{t-1}} + \mathbf{U_z h}^{(1)}_{\mathbf{t-1}} + \mathbf{M_z c}^{(1)}_{\mathbf{t}}) \\
\hat{\mathbf{h}}^{(1)}_{\mathbf{t}} &= \tanh(\mathbf{W y_{t-1}} + \mathbf{U}(\mathbf{r}^{(1)}_{\mathbf{t}} \odot \mathbf{h}^{(1)}_{\mathbf{t-1}}) + \mathbf{M c}^{(1)}_{\mathbf{t}}) \\
\mathbf{r}^{(1)}_{\mathbf{t}} &= \sigma(\mathbf{W_r y_{t-1}} + \mathbf{U_r h}^{(1)}_{\mathbf{t-1}} + \mathbf{M_r c}^{(1)}_{\mathbf{t}})
\end{aligned} \tag{3}$$

where all $\mathbf{W}$'s $\in \mathcal{R}^{\text{dim}\times\text{E}}$ and all $\mathbf{U}$'s $\in \mathcal{R}^{\text{dim}\times\text{dim}}$ are weighted matrices. Bias items are omitted in Equation (3). E indicates the dimensionality of word embeddings and dim indicates the dimensionality of hidden states.

The attention-based contexts are calculated as:

$$\mathbf{c}^{(1)}_{\mathbf{t}} = \sum_{1}^{\text{T}} \alpha_{ij} \mathbf{h}^{(1)}_{\mathbf{j}} \tag{4}$$

The attention signal is a scalar computed by:

$$\begin{aligned}
\alpha_{ij} &= \text{softmax}(\phi(\mathbf{h}^{(1)}_{\mathbf{j}}, \mathbf{y_{i-1}})) \\
\phi(\mathbf{h}^{(1)}_{\mathbf{j}}, \mathbf{y_{i-1}}) &= \mathbf{v}^T \tanh(\mathbf{W} \cdot [\mathbf{h}^{(1)}_{\mathbf{j}}, \mathbf{y_{i-1}}] + \mathbf{b})
\end{aligned} \tag{5}$$

α_{ij} is a normalized score which is a soft alignment model measuring how well the context status and the output hidden states are matched, which represents the attention distribution over the external memory, i.e., hidden states. $\phi(.)$ is a perceptron-like function where $\mathbf{W}$ are the weight matrices, $\mathbf{v}$ is the weight vector and $\mathbf{v}^T$ denotes its transpose.

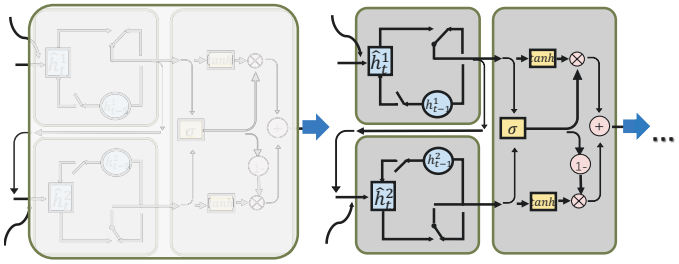

Figure 3: Deep Dual Fusion model with three cells: 1) Sequential GRU Cell with hidden states $\mathbf{h}_t^{(1)}$, 2) Alignment GRU Cell with hidden states $\mathbf{h}_t^{(2)}$, and 3) Fusion Cell with hidden states $\mathbf{h}_t$.

3.2 Alignment GRU Cell

To generate more meaningful suggestions for responses, we need to keep the conversation in line, which means the query, the generated response and suggestion will be aligned. Naturally, the key issue lies in how to incorporate query information into the suggestion generation process. One of the feasible methods is to apply the neural cell with additional inputs via various gating mechanisms. However, such revisions are generally designed for a particular scenario with less transferability to other tasks. Since we aim at tackling the multiple sequence-to-sequence framework, which would be compatible for many scenarios, we propose the Alignment GRU Cell, which is another independent neural cell to deal with the alignment issue with auxiliary information. The advantage of this neural cell is that it can be replaced or reused with better flexibility under the proposed framework.

During each time step, we have a hidden state from the Sequential GRU Cell, namely $\mathbf{h}_t^{(1)}$. Since we keep track of the hidden states from the input query, we can align the hidden state from the Sequential GRU Cell by paying attention to some hidden state(s) from the input query. To this end, information from the query can be better utilized for fusion so as to decode an aligned proactive suggestion. Given the last hidden state $\mathbf{h}_{t-1}^{(2)}$ and the current attention-based context, the new hidden state of the auxiliary decoding $\mathbf{h}_t^{(2)}$ is computed by following equations:

$$\begin{aligned}
\mathbf{h}_t^{(2)} &= (\mathbf{I} - \mathbf{z}_t^{(2)}) \odot \mathbf{h}_{t-1}^{(2)} + \mathbf{z}_t^{(2)} \odot \hat{\mathbf{h}}_t^{(2)} \\
\mathbf{z}_t^{(2)} &= \sigma(\mathbf{U}_z \mathbf{h}_{t-1}^{(2)} + \mathbf{M}_z \mathbf{c}_t^{(2)}) \\
\hat{\mathbf{h}}_t^{(2)} &= \tanh(\mathbf{U}(\mathbf{r}_t^{(2)} \odot \mathbf{h}_{t-1}^{(2)}) + \mathbf{M}\mathbf{c}_t^{(2)}) \\
\mathbf{r}_t^{(2)} &= \sigma(\mathbf{U}_r \mathbf{h}_{t-1}^{(2)} + \mathbf{M}_r \mathbf{c}_t^{(2)})
\end{aligned} \tag{6}$$

In the Alignment GRU Cell, we do not include the $\mathbf{y}$ information so that the generated words will not be double-counted, while we focus on the alignment function in the sequence. All dimensionality calculation is analogous to Equation (3), and the attention context is:

$$\mathbf{c}_t^{(2)} = \sum_1^{\mathrm{T}} \beta_{ij} \mathbf{h}_j^{(2)} \tag{7}$$

where

$$\begin{aligned}
\beta_{ij} &= \mathrm{softmax}(\pi(\mathbf{h}_j^{(1)}, \mathbf{y}_j, \mathbf{h}_i^{(2)})) \\
\pi(\mathbf{h}_j^{(1)}, \mathbf{y}_j, \mathbf{h}_i^{(2)}) &= \mathbf{v}^T \tanh(\mathbf{W} \cdot [\mathbf{h}_j^{(1)}, \mathbf{y}_j, \mathbf{h}_i^{(2)}] + \mathbf{b})
\end{aligned} \tag{8}$$

Note that both GRU cells do not share parameter matrices.

3.3 Fusion Cell

We have obtained two hidden states from the Sequential GRU Cell and the Alignment GRU Cell. There are several ways to combine them together such as concatenation or pooling [Zoph and Knight, 2016]. These methods are simple with shallow interactions between two outputs. Here we apply the fusion cell [Arevalo *et al.*, 2017] by deeply integrating the hidden states of both cells to compute the current hidden state $\mathbf{h}_t$. The equations are as follows:

$$\begin{aligned}
\mathbf{h}_t &= k \odot \mathbf{h}_t^{(1)} + (1 - k) \odot \mathbf{h}_t^{(2)} \\
k &= \sigma(\mathbf{W}_k[\hat{\mathbf{h}}_t^{(1)}, \hat{\mathbf{h}}_t^{(2)}]) \\
\hat{\mathbf{h}}_t^{(1)} &= \tanh(\mathbf{W}_1 \mathbf{h}_t^{(1)}) \\
\hat{\mathbf{h}}_t^{(2)} &= \tanh(\mathbf{W}_2 \mathbf{h}_t^{(2)})
\end{aligned} \tag{9}$$

The deep fusion unit provides an alternative solution for information integration. The gate neuron k controls the contribution of the information calculated from $\mathbf{h}_t^{(1)}$ and $\mathbf{h}_t^{(2)}$ to the overall output of the unit. The weighted matrices $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_k$ are all parameters to be learned.

Given the final output of $\mathbf{h}$ after the Fusion Cell, we decode the most likely generated word one by one chosen from the vocabulary. The output words are fed into the Sequential GRU Cell at the next time step until the variable-length suggestion sentence is fully generated.

4 Experiments and Evaluation

The objectives of our experiments are to verify the effectiveness the proposed model for the new generative neural conversation paradigm.

4.1 Experimental Setups

Dataset. We use the data which contain a large number of human conversations from [Yao *et al.*, 2017; Tao *et al.*, 2018]. The data are crawled from open Web, where the users publish messages visible to the public, and then receive a bunch of subsequent replies to their utterances. We conducted the same data filtering and cleaning used in [Yan *et al.*, 2016a]. We establish training samples by extracting the last three consecutive utterances (as a triple) from all conversations, and other preceding utterances as contexts.

For the test process, we have manually judged appropriateness for all utterances using crowdsourcing. Each sample was judged by at least 7 annotators via majority voting based on the *appropriateness*: "1" denotes an appropriate utterance (*response* or *suggestion*) and "0" indicates an inappropriate one. We examine the appropriateness of responses given the query, as well as the appropriateness of the proactive suggestion given the query and a response.

Evaluation Metrics. We evaluate the appropriateness of the response and the suggestion given a particular query. We also have test cases for automatic evaluations.

Given the annotated results for test queries, we evaluated the performance in terms of BLEU, ROUGE and human judgments. BLEU and ROUGE are widely used as evaluation metrics for machine translation and summarization systems, and recently for many conversational studies as well. Both metrics measure the word-overlapping information. The system generates the candidates. With the appropriateness judgments from humans, we are able to indicate the fraction of suitable utterances among the top results generated, which is quite similar to the metric of precision@1.

4.2 Hyperparameters

We use 512-dimensional word embeddings, and they were initialized randomly. All parameters are learned during training. As our dataset is in Chinese, we performed standard Chinese word segmentation. We maintained a vocabulary by choosing those with more than 2 occurrences, and we omitted the others.

The cell units have 300 hidden units for each dimension. We used stochastic gradient descent (with a mini-batch size of 100) for optimization, gradient computed by standard back propagation. Initial learning rate was set to 0.8, and a multiplicative learning rate decay was applied. We applied the validation set. All of the parameters were chosen and tuned empirically.

4.3 Competing Algorithms

We compared the proposed model against several baselines. Since our proposed approach is technically a generative method, and the evaluation for generation-based conversational systems is different from retrieval-based systems, we mainly focus on other generative baselines. For fairness we use the same pre-processing procedure for all algorithms.

Plain Seq2Seq. We apply the standard GRU Seq2Seq model for the proposed conversation paradigm: in fact the process is a 2-hop Seq2Seq. The first Seq2Seq generates the response and the second Seq2Seq generates the suggestion.

Multi-Seq2Seq. Multi-Seq2Seq [Zoph and Knight, 2016] was proposed for multi-source translation originally. In our scenario, the generation of suggestion comes from two parts. We apply the method into our scenario.

Attentive Multi-Seq2Seq. The Attentive Multi-Seq2Seq method was proposed with different ways to combine attentions with Multi-Seq2Seq learning. We implement the best reported configuration in [Libovický and Helcl, 2017].

Copying Seq2Seq. The copying mechanism [Gu *et al.*, 2016] is an effective way to incorporate terms from one source of sequence to another. We apply the mechanism into our task with copying score considered to fuse information.

Deep Dual Fusion. Given the original query and the generated response, the new model integrates the information from two parts in a deep fusion way via dual GRU cells, namely Deep Dual Fusion Model.

Note that for all methods listed above, we have the standard Seq2Seq to generate the response part, implementing different ways to combine the original query and the output response together to generate a suggestion.

Model	BLEU	ROUGE	Human Score
Plain	2.125	1.623	0.116
Multi	4.033	3.364	0.304
Attentive	4.625	3.877	0.327
Copy	3.229	2.942	0.315
Dual Fusion	**6.631**	**5.796**	**0.392**

Table 1: Appropriateness results by automatic and human evaluations for all methods. All methods use standard Seq2Seq to generate the response. Hence, we do not compare the performance for response generation, and the main focus is on suggestion generation.

4.4 Overall Performance

We conduct the appropriateness evaluation to see the performance of all methods in Table 1. Since response generation given the query is the same for all methods, we focus on the suggestion evaluation part. We report the BLEU, ROUGE score and human judgment scores. The human scores measure the appropriateness of the suggestion given both the query and the generated response.

Plain Seq2Seq shows the basic performance, which is not surprising. For *Multi-Seq2Seq* model, the performance is improved, and we assume that the information from both sequences of the original query and the generated response bring the advantages. The *Attentive Multi-Seq2Seq* model is an enhanced version with different attention strategies applied on the *Multi-Seq2Seq* model. Therefore, the performance gets boosted due to effective attention mechanisms. For the *Copying Seq2Seq* model, the automatic evaluation scores of BLEU and ROUGE are not quite promising and we understand that copying part of the query sequence may not be "favored" by BLEU or ROUGE metrics given the ground truth. For human evaluations, partially repeating the original query seems not to be a bad idea: the human score is better than that of *Multi-Seq2Seq*. Our model combines information in a deep dual fusion way and shows prominent improvement for both automatic and human evaluations.

4.5 Case Study

We show typical cases in Tables 2-3 to demonstrate the advantage of our proposed model. The 2-hop Plain Seq2Seq generates a good candidate suggestion to respond the response, but the suggestion diverges from the query. The Deep Dual Fusion model still keeps the suggestion in line with the query and the response. The information of "go hiking" passes along from the query to the response and then to the suggestion utterance.

5 Related Work

People have continuously devoted efforts for studies on conversational systems. At the very beginning, researchers usually focused on conversation systems based on rules or templates [Walker *et al.*, 2001; Williams *et al.*, 2013]. The idea is rather straightforward and such methods require few data for training. However, such systems require great human

Query
Is it a good day to go hiking?
Response
Sure let me know when you return.
Suggestion
Sorry but I don't know.

Table 2: An example of generated results for 2-hop Plain Seq2Seq. Utterances in the dataset are originally in Chinese while here we display the translated English version.

Query
Is it a good day to go hiking?
Response
Sure let me know when you return.
Suggestion
There are always good stories to know for hiking.

Table 3: An example of generated results for our proposed Deep Dual Fusion Model. Utterances are also translated into English.

efforts to create a large number of rules or templates to make the system operational. To this end, systems powered by humans are quite costly.

The trend for conversational systems gradually shifts from human-driven systems to data-driven systems. In this way, the need for a bigger amount of data for training is largely increasing. Thanks to the prosperity of online forums, social media (such as microblogs), and other Web resources, people now get used to having conversations on the Web. It is therefore practical to collect abundant human-to-human conversation data [Wang *et al.*, 2013].

Recently, deep neural network techniques are developing fast. With the help of deep learning, retrieval-based conversational systems are improved significantly. A series of neural retrieval-based methods are applied to short-text conversations, either for single-turn conversations [Ji *et al.*, 2014; Li and Xu, 2014; Lu and Li, 2013] or multi-turn conversations [Yan *et al.*, 2016b; Zhou *et al.*, 2016; Yan *et al.*, 2016a]. Basically, sentence representation using convolutional [Lu and Li, 2013; Hu *et al.*, 2014] or recurrent [Palangi *et al.*, 2015; Wan *et al.*, 2016] units is demonstrated to be effective so as to construct conversational systems.

Not surprisingly, generation-based conversational systems are also developing fast due to deep learning. In general, the sequence-to-sequence model is the dominant generation manner for generative conversational systems [Sutskever *et al.*, 2014]. A neural responding machine is proposed for single-turn conversations [Shang *et al.*, 2015]. Since conversations contain multi-turns, researchers extend the conversation scenario into multi-turns: plain contexts [Sordoni *et al.*, 2015] and hierarchical contexts [Serban *et al.*, 2016; Tian *et al.*, 2017] are investigated. Additional elements can also be incorporated into the generation process, such as diversity [Song *et al.*, 2018; Li *et al.*, 2016a], persona [Li *et al.*, 2016b], topic [Xing *et al.*, 2016], and contents [Yao *et al.*, 2017; Mou *et al.*, 2016]. A conversational system can be learned incrementally using reinforcement learning [Li *et al.*, 2016c] and/or adversarial learning [Li *et al.*, 2017].

In this paper, we propose a new conversation paradigm featured with smarter response and proactive suggestion. The difference compared with related work is quite clear. Most of the mentioned studies basically generate responses given the queries. In the new paradigm, the system generates responses and suggestions as a *pair* for proactive content introducing. The new task and the model have been preliminarily explored in retrieval-based system [Yan *et al.*, 2017], while we investigate the new *generative* conversational paradigm and extend the single-turn conversations in [Yan *et al.*, 2017] to context-aware conversations in this paper.

6 Conclusion

We propose a new generative conversation paradigm between humans and computers, featured by smarter response with proactive suggestions. We investigate the utilization in experiments: the system attracts users by providing more information proactively and hence makes users more interactive in the chit-chat conversations. The results are not surprising since users are assumed to maintain an open style in open-domain conversations.

We propose a Deep Dual Fusion Model for the generative task. The model fuses information from dual sequences via deep interactions through cell units and gatings. It shows promising appropriateness results in automatic evaluations as well as human judgments compared with baselines. In the future, we will investigate how to incorporate additional elements for more options of proactive suggestions.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61672058). Rui Yan was sponsored by CCF-Tencent Open Research Fund and Microsoft Research Asia (MSRA) Collaborative Research Program.

References

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, and et al. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[Gu *et al.*, 2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. pages 1631–1640, 2016.

[Hu *et al.*, 2014] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *NIPS*, pages 2042–2050, 2014.

[Ji *et al.*, 2014] Zongcheng Ji, Zhengdong Lu, and Hang Li. An information retrieval approach to short text conversation. *CoRR*, abs/1408.6988, 2014.

[Li and Xu, 2014] Hang Li and Jun Xu. Semantic matching in search. *Foundations and Trends in Information Retrieval*, 8:89, 2014.

[Li *et al.*, 2016a] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL'16*, pages 110–119, 2016.

[Li *et al.*, 2016b] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *ACL'16*, pages 994–1003, 2016.

[Li *et al.*, 2016c] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. In *EMNLP'16*, pages 1192–1202, 2016.

[Li *et al.*, 2016d] Xiang Li, Lili Mou, Rui Yan, and Ming Zhang. Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. In *IJCAI'16*, pages 2845–2851, 2016.

[Li *et al.*, 2017] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP'17*, pages 2157–2169, 2017.

[Libovickỳ and Helcl, 2017] Jindřich Libovickỳ and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *ACL'17*, pages 196–202, 2017.

[Lu and Li, 2013] Zhengdong Lu and Hang Li. A deep architecture for matching short texts. In *NIPS*, pages 1367–1375, 2013.

[Mou *et al.*, 2016] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING'16*, pages 3349–3358, 2016.

[Palangi *et al.*, 2015] Hamid Palangi, Li Deng, and et al. Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval. *arXiv:1502.06922*, 2015.

[Serban *et al.*, 2016] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI'16*, pages 3776–3783, 2016.

[Shang *et al.*, 2015] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *ACL-IJCNLP'15*, pages 1577–1586, 2015.

[Song *et al.*, 2018] Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. Towards a neural conversation model with diversity net using determinantal point processes. In *AAAI'18*, 2018.

[Sordoni *et al.*, 2015] Alessandro Sordoni, Michel Galley, and et al. A neural network approach to context-sensitive generation of conversational responses. In *NAACL'15*, pages 196–205, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Tao *et al.*, 2018] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI'18*, 2018.

[Tian *et al.*, 2017] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. How to make contexts more useful? an empirical study to context-aware neural conversation models. In *ACL'17*, pages 231–236, 2017.

[Walker *et al.*, 2001] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *ACL*, pages 515–522, 2001.

[Wan *et al.*, 2016] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *AAAI'16*, pages 2835–2841, 2016.

[Wang *et al.*, 2013] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. A dataset for research on short-text conversations. In *EMNLP'13*, pages 935–945, 2013.

[Williams *et al.*, 2013] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In *SIGDIAL*, pages 404–413, 2013.

[Xing *et al.*, 2016] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic augmented neural response generation with a joint attention mechanism. *arXiv preprint arXiv:1606.08340*, 2016.

[Yan *et al.*, 2016a] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR'16*, pages 55–64, 2016.

[Yan *et al.*, 2016b] Rui Yan, Yiping Song, Xiangyang Zhou, and Hua Wu. Shall i be your chat companion?: Towards an online human-computer conversation system. In *CIKM'16*, pages 649–658, 2016.

[Yan *et al.*, 2017] Rui Yan, Dongyan Zhao, and Weinan E. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *SIGIR'17*, pages 685–694, 2017.

[Yao *et al.*, 2017] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP'17*, pages 2190–2199, 2017.

[Yoshino and Kawahara, 2015] Koichiro Yoshino and Tatsuya Kawahara. News navigation system based on proactive dialogue strategy. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 15–25. 2015.

[Zhou *et al.*, 2016] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view response selection for human-computer conversation. In *EMNLP'16*, pages 372–381, 2016.

[Zoph and Knight, 2016] Barret Zoph and Kevin Knight. Multi-source neural translation. In *NAACL'16*, pages 30–34, 2016.