History-Adaption Knowledge Incorporation Mechanism for Multi-Turn Dialogue System

Yajing Sun,[1,2] Yue Hu,[*,1,2] Luxi Xing,[1,2] Jing Yu,[1,2] Yuqiang Xie[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{sunyajing, huyue, xingluxi, yujing02, xieyuqiang}@iie.ac.cn

Abstract

Keeping the conversation consistent and avoiding its repetition are two key factors to construct an intelligent multi-turn knowledge-grounded dialogue system. Although some works tend to combine history with external knowledge such as personal background information to boost dialogue quality, they are prone to ignore the fact that incorporating the same knowledge multiple times into the conversation leads to repetition. The main reason is the lack of effective control over the use of knowledge on the conversation level. So we design a history-adaption knowledge incorporation mechanism to build an effective multi-turn dialogue model. Our proposed model addresses repetition by recurrently updating the knowledge from the conversation level and progressively incorporating it into the history step-by-step. And the knowledge-grounded history representation also enhances the conversation consistency. Experimental results show that our proposed model significantly outperforms several retrieval-based models on some benchmark datasets. The human evaluation demonstrates that our model can maintain conversation consistent and reduce conversation repetition.

Introduction

Building a human-machine conversational agent is one of the most important and challenging tasks in artificial intelligent(AI). Recently, building a chatbot for open-domain conversation (Serban et al. 2015; Ghazvininejad et al. 2017) has gained increasing interests due to both availabilities of a large amount of human conversation data and powerful models learned based on the neural networks. Existing methods are either retrieval-based or generation-based. The retrieval-based methods match the dialogue semantic representation with several response candidates, and then select an appropriate one as a reply to a human input (Shang, Lu, and Li 2015; Walker, Lin, and Sawyer 2012). On the other hand, the generated-based methods directly generate a response via natural language generation techniques (Mairesse and Walker 2007; 2008). We focus on the problem of response selection for retrieval-based chatbots since retrieval-based methods have the advantage of providing fluent and informative responses.

It's a crucial step to measure the relevance degree between the context of conversation and candidate responses. Although great progress has been made, common issues still exist in open-domain multi-turn dialogue. First, the models fail to capture the sequential and contextualized information of a long history, which leads to the loss of important information (Vinyals and Le 2015; Sordoni et al. 2015) and hinders the conversation understanding. Second, the models have poor ability to stay or change topics to keep dialogue consistent by itself (Li et al. 2016b; Zhang et al. 2018). Third, the repetitiveness of response, which incorporates and utilizes the same external knowledge multiple times to generate the same but meaningless content, remains without a de facto solution (Serban et al. 2016; Li et al. 2016a; Xu et al. 2017).

To address these problems, in recent years, several approaches have been developed to generate or select consistent and informative responses. Some works apply hierarchical recurrent network (Xing et al. 2018), convolution network (Wu et al. 2016) or transformer architecture (Mazaré et al. 2018) to encode the semantics of contexts and responses on words, n-grams, and sub-sequences of utterances. These methods capture both short-term and long-term dependencies among words. However, these works can't guarantee the consistency and interactivity of the dialogue only depending on the context of the conversation, since data-driven models are learned using the conversation data produced by the different speakers. There are two types of methods to model personalized neural conversation models. Li et al.; Kottur, Wang, and Carvalho; Zhang et al.(2016b; 2017; 2019) encode each speaker to a user vector, and the vector is fed to the sequence-to-sequence model to capture the speaking style of the speaker implicitly. Due to poor interpretability and data sparsity, the other methods generated or selected responses are conditioned either on a given personal profile (Zhang et al. 2018) or on a text-described external knowledge (Qian et al. 2017). It's a matter of fact that structured-form or unstructured-form information can greatly improve the consistency and interactivity of the dialogue. But most of the existing methods tend to pay attention to using per-

[*]Corresponding Author

sonal knowledge, while ignoring the fact that incorporating and utilizing the same knowledge for multiple times leads to repetition of dialogue (Lian et al. 2019). The main reason is the lack of effective control over the use of information. It's indispensable to keep track of the activation of each piece of information during the conversation flow.

In this paper, we propose a novel history-adaption knowledge incorporation mechanism to build an effective multi-turn dialogue model. We introduce personalized knowledge to keep consistent and pay attention to the fact that external information has different contributions to different dialogue histories. Distinct from existing approaches, we recurrently update the external knowledge from utterances level and progressively incorporate it into the history step-by-step, and finally we consider the sequential information between different history turns, and use hierarchical recurrent mechanism to synthesize them to a vector for scoring the candidate answers.There are two reasons why modeling history-adaption personalized information. The control and update of external information are closely related to the context of the conversation. On the one hand, the knowledge utilized in t-th turn tends to be semantically related to the partner's last utterance. On the other hand, the knowledge utilized in the t-th turn tends to be dissimilar with the former dialogue history. For sake of coherence, we need to select semantically relevant information for context understanding from utterance level. To avoid repetition, we consider updating external knowledge, which encourages extensive coverage and avoids unnecessary repetition.

Comprehensive experiments show that our model achieves a higher scores on hits@1 and F1 metrics than several benchmarks, which demonstrates that our proposed model has a better capability of capturing the semantic information of dialogue and selecting the more relevant answer from the provided answers. And the human evaluation in repetition, consistency, and repetition also outperforms higher than baselines.

Related Work

Dialogue Understanding Understanding the context and scene of the conversation is the heart of any dialogue system. With the development of deep learning, sequence-to-sequence (SEQ2SEQ) neural network provides the possibility to improve the quality of multi-turn chit chat dialogue. Various studies have been proposed for tackling the issue of understanding conversation. Serban et al. (2015) proposed HRED which considers the history at two levels: a sequence of the words and utterances. The HRED models the hierarchy of sequences with two RNNs to avoid losing important information. Xing et al. (2018) paid attention to the fact that the words and utterances of the dialogue are differently important and proposed HRAN to encode context using two levels attention mechanism, which aimed at modeling both the hierarchy and the importance variance in a unified framework. Besides, Luo et al. (2018) proposed to use memory network (Sukhbaatar et al. 2015) to address the problem that long history is easy to be forgotten. The model encodes and stores context that comes from itself and other similar users with the memory module. Wu et al. (2016) proposed a sequential matching network which uses the CNN network to extract important information from dialogue and uses RNN to capture and the sequential information of dialogue. In our work, we also focus on dialogue context understanding. Because memory network costs high calculating consumption and makes training slowly, we adopt hierarchical attention and utterance encoder mechanism for history encoder to pay attention to the important information in the history and integrate the latent semantics of conversation history more deeply.

Personalized Dialogue In recent years, some researchers have shown a growing interest in modeling of personality. Recent studies on personalized neural conversational models can be broadly classified into two types: one is implicit personalization and the other is explicit personalization. Li et al. (2016b) first introduced personality information into dialogue generation. The persona-based neural model uses the user embedding to capture the users' background information and speaking style into the model to keep consistency. Instead of learning latent personality embedding from dialogue data, explicit personalization approaches attempt to endow dialogue models with persona which is described by natural language sentences. Qian et al. (2017) constructed the structural personality knowledge and assigned a desired identity to a chatbot. In order to keep consistency, the model proposed the profile detector to select personality in the encoder. Recent works tried to construct high-quality data. Zhang et al. (2018) contributed the Persona-Chat dataset which gave a text-described persona, and they further proposed both ranking and generative models. Lian et al. (2019) pay attention to select appropriate profile knowledge with the method that employs the posterior knowledge distribution. But these methods just attempt to incorporate personality information to keep consistency. None of the existing models pay attention to control and coordinate the information utilization on the conversation level. The same personalized information contributes differently to different histories in a conversation, and its contribution to the current conversation is related to the previous conversation history. So our work focuses on building a multi-turn dialogue system by modeling the updating of profile information and incorporating the updated profile information into the history step-by-step.

Response Selection Some works focus on the response selection of the retrieval-based multi-turn dialogue. The modern neural networks can be roughly divided into two types, shallow and deep interaction networks. For shallow interaction networks, many efforts have been devoted to learning good representation for input and candidate independently. Wu et al. (2016) proposed a sequential matching network that utilized convolution and pooling operations to extract the difference levels information and encoded the representation with RNN to capture sequential information of dialogue. Zhou et al. (2016) proposed a multi-view response selection model for a multi-turn human-computer conversation. The model integrates the existing word sequence view with deep neural network and utterance sequence view with the recurrent neural network into a uni-

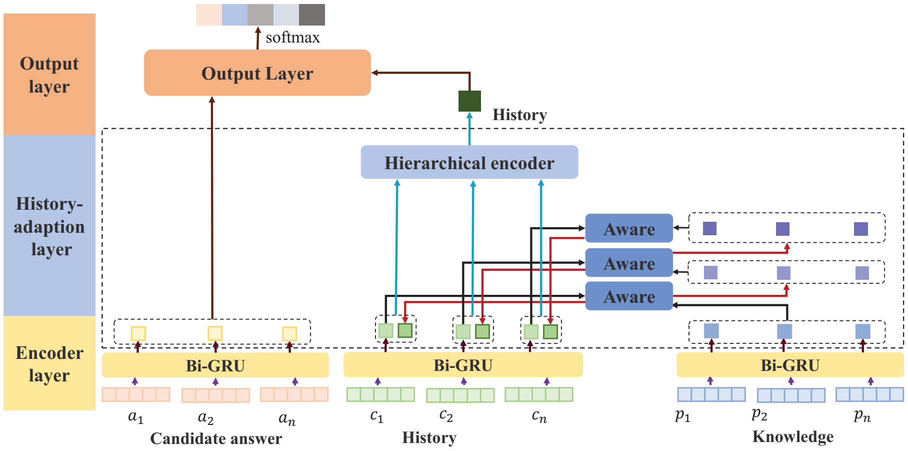


Figure 1: Model Overview. There are three parts including the encoder layer, the history-adaption layer, and the output layer. The network takes context C, knowledge sentences P and candidate answers A as inputs and selects appropriate answer. The parameters of the encoder are not shared, and the history-adaption layer is shared.

fied multi-view model. For deep interaction networks, various approaches are proposed to interact query and response to generate a single feature vector that preserves all query-response interaction information at different levels of abstraction. Hu et al. (2014) propose deep convolutional architectures for matching natural language sentences, which can nicely combine the hierarchical modeling of individual sentences and the patterns of their matching.

Model

Task Definition

Suppose that we have a data set $\mathcal{D} = (P, C, A)$. Let $P = \{p_1, p_2, \ldots, p_{l_p}\}$, where $p_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,k}\}$ represents the knowledge. l_p is the number of knowledge, k is the length of a sentence. And $C = \{c_1, c_2, \ldots, c_{l_c}\}$, and $c_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,k}\}$ represents history information that the current question is located in the last turn. l_c is the turn of history. And the candidate answers are $A = \{a_i, a_2, \ldots, a_{l_a}\}$, where l_a is the number of candidate answers. Our goal is to learn a score model $g(\cdot)$ with $\mathcal{D}$ to select right and proper answer from candidate answers set A.

Model Overview

As shown in Figure 1, our model is composed of three parts: the encoder layer, the history-adaption layer, and the output layer. First, we use Bi-GRU to capture the context information of the external knowledge, history information and candidate answers respectively. And then we use the self-attention mechanism to identify the important information and ignore irrelevant information from them. In the history-adaption layer, we recurrently update the external knowledge and progressively incorporate it into history, which not only helps history to capture the knowledge but also dynamically updates the knowledge based on history. Then we use the hierarchical recurrent network to extract sequential and contextualized information. Finally, in the output layer, we compute the similarity of history and candidate answers. The detail of the three parts is described in the following sections.

Encoder Layer

This layer is responsible for extracting context information. Specifically, the knowledge information P, history information C and candidate information A are respectively processed by an encoding module to encode long-term dependency among words into representation.

The encoding module consists of a Bi-GRU component and a self-attention component. Without loss of generality, let $X = \{x_1, x_2, \cdots, x_t, \cdots, x_l\}, x_t \in \mathbb{R}^{k \times d}$ denotes the embedding matrices of sentence, where l is the number of sentences and k represents the length of a sentence and d stands for hidden size.

Mathematically, x_t is first to encoded to $h_t \in \mathbb{R}^{k \times d}$. Then we apply a self-attention mechanism to calculate the sentence-level representation. The sentence-level representation focuses more on the key information in an sentence. Practically, for each sentence $h_t = \{h_{t,1}, h_{t,2}, \ldots, h_{t,k}\}$,

$$e_{t,j} = v^{\mathrm{T}} \tanh(W_1 h_{t,j}), \tag{1}$$

$$h'_t = \sum_{j=1}^{k} \frac{\exp e_{t,j}}{\sum_{n=1}^{k} \exp e_{t,n}} h_{t,j} \tag{2}$$

$h'_t \in \mathbb{R}^{1 \times d}$. We denote the whole encode module as $f_{enc}(\cdot)$. P, C and A are represented through $f_{enc}(\cdot)$:

$$h^p = f_{enc}(P) \tag{3}$$

$$h^c = f_{enc}(C) \tag{4}$$

$$h^a = f_{enc}(A). \tag{5}$$

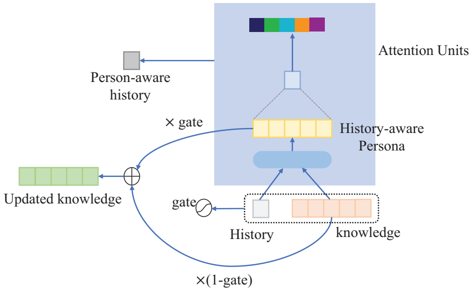


Figure 2: Illustration of history attending to the external knowledge module. The key module takes current turn history representation and knowledge representation as input, and the output includes two parts: updated knowledge representation and knowledge-aware history representation.

History-adaption Layer

It is indispensable to keep track of the activation of each piece of external knowledge during the conversation flow to address repetition and consistency. And the control and update of external knowledge are closely related to the context of the conversation.

Based on this principle, we design a novel gate mechanism to control the flowing of the external knowledge information to the current and next turn dialogue. On the one hand, the module integrates the knowledge information into the current dialogue and get knowledge-aware representation, which helps keep dialogue consistency. On the other hand, the knowledge-aware representation also helps to control to update knowledge updating which are used in the next turn to address repetition.

Distinct from existing approaches, we recurrently update the external knowledge from utterance level and progressively incorporate it into the history step-by-step. Finally we consider the sequential information between different history turns, and use hierarchical recurrent mechanism to calculate the final history representation. The representation is served as the encoder feature output for scoring the candidate answers.

History-aware knowledge updating Suppose the initial state of knowledge information is $h^{p(1)} = h^p \in \mathbb{R}^{l_p \times d}$, then we recurrently update the knowledge based on the history information $h^c = \{h_1^c, h_2^c, \cdots, h_{l_c}^c\}$. The dynamic updating path is showed as follows:

$$(h^{p(1)}, h_1^c) \rightarrow (h^{p(2)}, h_1^{c'}) \tag{6}$$

$$(h^{p(t)}, h_t^c) \rightarrow (h^{p(t+1)}, h_t^{c'}) \tag{7}$$

As shown in the Figure 2, we update the external knowledge $h^{p(t)}$ in step t as follows,

$$U = V_1^T Tanh\{W_2[(h^{p(t)} \oplus h_t^c); (h^{p(t)} \odot h_t^c)] + b\} \tag{8}$$

where $U \in \mathbb{R}^{l_p \times d}$. U is used for two purposes: 1) Controlling the information of knowledge that flows to the next turn. This step guarantees effective and thorough control over the use of knowledge. 2) Selecting semantically relevant knowledge information for context understanding from utterance level.

For the first purpose, we perform an nonlinear operation between $h^{p(t)}$ and U to get the new representation.

$$h^{p(t')} = \sigma\{W_3[(U \oplus h^{p(t)}); (U - h^{p(t)}); (U \odot h^{p(t)})] + b\} \tag{9}$$

Afterwards, we design the forget gate to select the old and new knowledge representation as follows,

$$gate = Tanh(W_4 h_t^c) \tag{10}$$

$$h^{p(t+1)} = gate \cdot h^{p(t')} + (1 - gate) \cdot h^{p(t)} \tag{11}$$

The purpose of designing such architecture is to take consistency and repetition of the dialogue into consideration. We apply gate mechanism to encourage to select the knowledge that is semantically relevant to the current turn dialogue and avoid to select repetitive but unnecessary knowledge with dialogue. In this way, we can avoid unnecessary repetition and encourage knowledge coverage.

For the second purpose, we normalize U to score each piece of knowledge representation and get knowledge-aware history representation, which incorporates the external knowledge to keep consistency. Specifically,

$$\alpha = softmax(V_2^T U) \tag{12}$$

$$h_t^{c'} = \sum_{i=1}^{l_p} \alpha h_i^{p(t)} \tag{13}$$

Then we update the external knowledge and progressively incorporate it into the history step-by-step. Note that the update process adaptively depends on the length of the history.

Hierarchical History Encoder Similar to the (Xing et al. 2018), we identify the different importance between the words and utterance level in the dialogue. We use different GRU from the encoder layer to encode the contextual history information and use the self-attention mechanism to pick up the important information to a vector. Before encoding, history $H = [h_1, h_2, \cdots, h_{l_c}]$, where $h_t = [h_t^c; h_t^{c'}]$ is fed to the GRU and self-attention architecture. We represent the final meaningful history representation as O. Specifically,

$$H^s = GRU(H) \tag{14}$$

$$e^s = V_3^T Tanh(W_5 H^s) \tag{15}$$

$$O = \sum_{i=1}^{l_c} \frac{\exp e_i^s}{\sum_{j=1}^{l_c} \exp e_j^s} h_i^s \tag{16}$$

Output Layer

The output layer is responsible to calculate the similarity of history and candidate answers to select consistent in knowledge and contextual coherence response from candidate answers. We define the score function as following:

$$g(O, A) = softmax(O^T h^a) \tag{17}$$

method	PERSONA-CHAT						CMUDoG		
	original persona			revised persona					
	r@1	r@2	r@5	r@1	r@2	r@5	r@1	r@2	r@5
KV profile Mmeory (Zhang et al. 2018)	51.1	61.8	77.4	35.1	45.7	66.3	56.1	69.9	80.3
Transformer (Mazaré et al. 2018)	54.2	68.3	83.8	42.1	56.5	75.0	60.3	74.4	80.3
our model	**57.6**	**72.9**	**89.9**	**52.4**	**68.5**	**87.0**	**82.7**	**93.8**	**99.5**

Table 1: Experimental results of automatic metrics with the different models on the persona-chat data and CMUDoG data. And in the Persona-Chat data, there are two different settings: conditioned on the speakers given persona(”original persona”), or a revised persona that does not have word overlap.

We train the whole model by minimizing the standard cross entropy of $g(\cdot,\cdot)$. Let Θ denote the parameters of the model and the objective function can be formulated as:

$$\begin{aligned}\mathcal{L}(\mathcal{D},\Theta) = -\sum_{i=1}^{l}[y_i \log(g(O, A_i^s)) \\ + (1-y_i)\log(1-g(O, A_i^s))].\end{aligned} \tag{18}$$

Experiments

Dataset

We performed our experiments on the two public available datasets – Persona-Chat (Zhang et al. 2018) and CMUDoG dataset published recently in Zhou, Prabhumoye, and Black(2018).

The Persona-Chat dataset is obtained from crowdworkers on Amazon Mechanical Turk who are required to chat with each other according to their assigned profiles. Each dialogue contains average 4.49 sentences served as profile, average 7.35 turns utterances as conversation history and an utterance treated as a positive response which is associated with 19 negative response candidates. Persona-Chat has separated training and test set. In total, there are 8939 dialogues (65719 turns) in the training set and 968 dialogues (7512 turns) in the test set. We can get dataset on ParlAI [1].

In addition to Persona-Chat dataset, we also experiment with CMUDoG dataset published in Zhou, Prabhumoye, and Black(2018). The dataset addresses the concern of the grounding in conversation responses, context, and coherence in responses.Each movie document consists of four sections corresponding to basic information and three key scenes of the movies. The 4 sections are shown to one or both workers one by one every 3 turns. The dataset consists of total 4112 conversations with an average of 21.43 turns and has been divided into a training set, validation set and test set by publishers. Besides, we randomly sample 19 negative response candidates for each utterance from the same set. The candidate answers which are selected randomly may be easily distinguished from the correct answer. we will try data enhancement methods to select or generate better candidate answers in the future work.

To supervise knowledge selection in the history-adaption layer in every turn, we label each utterance with its corresponding knowledge by calculating TF-IDF similarity.

Comprehensive comparisons have been made to the following methods:

- **KV Profile Memory:** the model in Zhang et al.(2018) performs best. The model considers the keys as dialog histories and the values as the next dialogue utterances. It performs attention over the profile which is then used to attend over the keys and outputs a weighted sum of values.
- **Transformer:** the model exhibits state-of-art performance on the Persona-Chat data which is reported in Mazaré et al.(2018).

Note that we don’t include models which are pre-trained on the large-scale dataset since the comparison is unfair. And we leave the study as future work.

Experiment Settings

We train our model on the two datasets using the following setting:

- The number of history length is set as 6, 7, 8 to compare the performance in Persona-chat dataset. And the history length in CMUDoG dataset is set as 7.
- We use Adam optimizer with a batch size of 128 and an initial rate of 0.001. Then we use the smaller learning rate 1e-5 to fine-tune the model. To avoid model overfitting, we set dropout as 0.5. Specifically, we clip the gradient when its norm exceeds 5. For CMUDoG dataset, we apply early stop.
- For word embedding representation, we use GloVe (Pennington, Socher, and Manning 2014) with an embedding size of 300. Note that we lock the embedding weights and set their gradients to zero.
- In the encoder layer, we use a layer BiGRU model with a hidden size of 300. The hierarchical encoder layer is one layer GRU with hidden size of 300.

Results

Automatic Evaluation

Following Zhang et al.(2018), We apply r@k and F1 as automatic metrics. Table 1 reports evaluation on the two datasets. We can see our model improves notably compared with the baselines in both datasets. The improvements on r@k and F1 mean that the proposed model has a better ability to capture deep semantic information of the dialog and select the more

[1] http://convai.io/data/

Method	original persona			revised persona		
	r@1	r@2	r@5	r@1	r@2	r@5
our model	57.6	72.9	89.9	52.4	68.5	87.0
w/o knowledge	47.1	62.8	82.4	46.3	62.3	82.8
w/o knowledge update	55.7	70.8	87.8	49.4	64.7	84.5

Table 2: The ablation tests. Results in r@k are significantly different from the ablated models.

relevant answer from the provided answers. Besides, the improvements in our model on the CMUDoG data is larger than that on the persona-chat. The reason might be that CMUDoG dataset contains more knowledge and richer semantics. The history-adaption knowledge updating mechanism can be trained better to help the conversation to control the use of knowledge effectively on the CMUDoG dataset.

Ablation Study

To investigate the influence of the history-adaption knowledge incorporation mechanism, we conducted an ablation test on the persona-chat. Table 2 displays the performance of ablating the history-adaption module and the knowledge-aware module compared to our model respectively. The performance only using knowledge is significantly higher than the model without knowledge, but it's lower than our model with knowledge updated. The sharp reduction of r@k without the knowledge or updated knowledge demonstrates that incorporating knowledge into the conversation is necessary and the updated knowledge plays an important role to improve the quality of dialogue. The main reasons are the two following aspects. On the one hand, the module incorporates knowledge into the each turn history, which is useful to keep conversation consistency. On the other hand, updated knowledge recurrently can memorize the used information previously and help each turn dialogue to capture appropriate knowledge information, which improves the relevance between the dialogue and proper response and avoids dialogue repetition.

Length Analysis

It is statistical that the dialogue contains an average 7.35 turns utterances as conversation history in the persona-chat. So we compare the model with a history length of 6, 7, and 8. As Table 3 shown, the model with length 7 achieves the best results. It demonstrates that our model can capture the sequential and contextualized information of a long history. The main reason is the combination of the hierarchical recurrent network and history-adaption mechanism. Specifically, the history-adaption mechanism promotes mutual understanding between knowledge and history information. And then hierarchical recurrent network captures sequential information for the knowledge-aware history. Besides experiments also show that increasing the length doesn't influence the cost of training time much.

Human Evaluation

As automated metrics are poor for evaluating the quality of our model, we adopt human evaluation which is suggested by Zhang et al. (2018). We also evaluate human performance, which effectively gives upper bound scores. In both settings, we present external texts, input messages, as well as the selected response by the baseline and our model. Considering the candidate responses are pre-defined and it's difficult to select randomly examples to verify repetition and consistency. So we manually select 100 turns original persona dialogues in which a part of candidate responses contain repetitive and inconsistent answers with the history and external knowledge. Then we adopt crowd-sourced to score. Note that each conversation example is required to evaluate twice by two participants. And the participants are required to score the answers with the following standards.

Length	original persona		revised persona	
	r@1	F1	r@1	F1
6	55.5	62.1	50.0	57.5
7	57.6	64.0	52.7	59.8
8	54.8	61.3	50.4	57.9

Table 3: Auto metrics performance for different length of history.

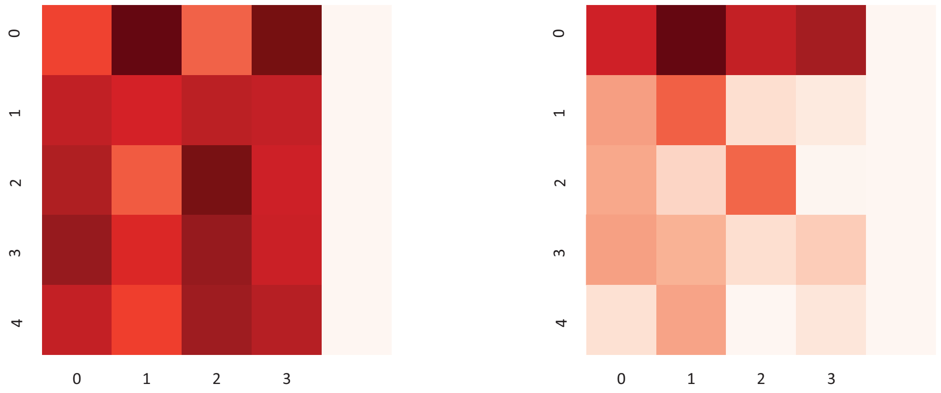
- **Repetition** The repetition refers to the problem that the model tends to utilize the same knowledge, whereas it is irrelevant to the current-turn conversation. So we ask them to judge the repetition as a score from 1 to 3. 1 means the response uses the same information, 2 means the response doesn't use information, and 3 means response uses the new information.
- **Consistency** We ask them to judge the consistency of personalized information in the dialogue compared to the given persona by us. The score is from 1-3. Specifically, 1 is not consistent, 2 is neutral and 3 is consistent. Score 2 is aimed at the instances that belong to daily conversations such as *I'm fine, how about you?*.
- **Relevance** We ask them to judge the relevance by giving a 3-scale rating. 1: the response is total not relevant; 2: the response is a little relevant; 3: the response is relevant.

We calculated the Fleiss' kappa to measure inter-rater consistency. The Fleiss' kappa for repetition, consistency and relevance is 0.6530, 0.6728 and 0.5379, indicating "substantial agreement", "substantial agreement", and "moderate agreement" respectively.

From the results reported in Table 4, we can conclude: First, our model is better than the baseline in terms of human rating, which demonstrates the effectiveness of the proposed history-adaption mechanism. Second, our model alleviates repetition. Due to the small proportion of examples that contains the repetitive and consistent answers, the overall scores in repetition and consistency are relatively high. However, we can still see that our model scores higher than the baseline model.

Case Study

We use the updated knowledge representation and knowledge-grounded history information to obtain the



(a)

(b)

Figure 3: The figure (a) depicts Attention weights over persona for different turns history. The x-axis represents the four profile information. And y-axis is the turns of history. Besides, the darker the color, the greater the value. The figure (b) shows the mean scalar of each updated persona representation for different turns history

Method	repetition	consistency	relevance
KV-profile network	2.07	2.10	1.85
our model	2.36	2.25	1.98
human performance	2.56	2.50	2.25

Table 4: Human evaluation for benchmarks, along with a comparison to human performance.

history	
0	i keep all my cars in great condition. that is good . i wish i had someone to play with , boring being
1	try going to the gym . the secret is never working out. no one to take me to the gym or anywhere else .
2	how old are you then ? 8 , and i live in a cloud by myself
3	sounds lonely . i'm lonely too . my girlfriends always dump me. well you could stop time like me before you get dumped .
4	they just do not understand my love for my cars .
persona	
0	I like to make time stop.
1	I'm very lonely.
2	I live in the cloud.
3	I am a little girl.

Table 5: Example dialog corresponding to the Figure 3b and Figure 3a. The conversation consists of 5 turn history, which corresponds to the y-axis in Figure 3b. And 4 profile information corresponds to the x-axis.

weights of candidate answers. Thus, we expect to observe the change of knowledge representation and the weights over representation as the conversation going on. For each piece of knowledge, we calculate the mean of the hidden vector. Then we visualize the knowledge representation and the attention of them over different history information. The corresponding example is showed in Table 5. For Figure 3a and 3b, the x-axis represents the different knowledge which showed in the Table 5 and the y-axis represents the history turns.

As shown in Figure 3a, we observe that different piece of knowledge has different weights in the same context of dialogue, which shows that our model can incorporate appropriate information into history turns. For instance, in the second history turn *8, and I live in a cloud by myself*, the weights for third knowledge *I live in the cloud* are significantly larger than others, which corresponds to the dialogue example in the Table 5. Besides, we also observe that the attention of the same information is sharp in the process of conversation, which demonstrates our model can avoid using the same knowledge multiple times.

As shown in Figure 3b, we can see that the representation of the same knowledge is changing with the same tendency as the conversation goes on. It means that the proposed mechanism of history-adaption knowledge updating can prevent the same knowledge from being utilized multiple times. Meanwhile, there is a distinctive representation of different knowledge in the same history. The result illustrates the fact that history can distinguish the importance among the given knowledge, which won't be influenced by the length of history.

Conclusion

In this paper, we propose a history-adaption knowledge incorporation mechanism to model an effective multi-turn dialogue. The experimental evaluation shows that our model can improve the conversation quality on several benchmark datasets.

In the future, we will try to apply our history-adaption mechanism to the generation-based method. We also explore to capture hidden information in the conversation and enhance knowledge using and understanding.

Acknowledgements

This work is supported by the National Key Research and Development Programs (Grant No.2017YFB0803301).

References

Ghazvininejad, M.; Brockett, C.; Chang, M.; Dolan, B.; Gao, J.; Yih, W.; and Galley, M. 2017. A knowledge-grounded neural conversation model. *CoRR* abs/1702.01932.

Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2042–2050.

Kottur, S.; Wang, X.; and Carvalho, V. 2017. Exploring personalized neural conversational models. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 3728–3734.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 110–119.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016b. A persona-based neural conversation model. *CoRR* abs/1603.06155.

Lian, R.; Xie, M.; Wang, F.; Peng, J.; and Wu, H. 2019. Learning to select knowledge for response generation in dialog systems. *CoRR* abs/1902.04911.

Luo, L.; Huang, W.; Zeng, Q.; Nie, Z.; and Sun, X. 2018. Learning personalized end-to-end goal-oriented dialog. *CoRR* abs/1811.04604.

Mairesse, F., and Walker, M. A. 2007. PERSONAGE: personality generation for dialogue. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Mairesse, F., and Walker, M. A. 2008. A personality-based framework for utterance generation in dialogue applications. In *Emotion, Personality, and Social Behavior, Papers from the 2008 AAAI Spring Symposium, Technical Report SS-08-04, Stanford, California, USA, March 26-28, 2008*, 80–87.

Mazaré, P.; Humeau, S.; Raison, M.; and Bordes, A. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2775–2779.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Qian, Q.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *CoRR* abs/1706.02861.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR* abs/1507.04808.

Serban, I. V.; Lowe, R.; Charlin, L.; and Pineau, J. 2016. Generative deep neural networks for dialogue: A short review. *CoRR* abs/1611.06216.

Shang, L.; Lu, Z.; and Li, H. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 1577–1586.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. *CoRR* abs/1506.06714.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. Weakly supervised memory networks. *CoRR* abs/1503.08895.

Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.

Walker, M. A.; Lin, G. I.; and Sawyer, J. 2012. An annotated corpus of film dialogue for learning and characterizing character style. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, 1373–1378.

Wu, Y.; Wu, W.; Zhou, M.; and Li, Z. 2016. Sequential match network: A new architecture for multi-turn response selection in retrieval-based chatbots. *CoRR* abs/1612.01627.

Xing, C.; Wu, Y.; Wu, W.; Huang, Y.; and Zhou, M. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5610–5617.

Xu, Z.; Liu, B.; Wang, B.; Sun, C.; Wang, X.; Wang, Z.; and Qi, C. 2017. Neural response generation via GAN with an approximate embedding layer. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 617–626.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *CoRR* abs/1801.07243.

Zhang, W.; Zhu, Q.; Wang, Y.; Zhao, Y.; and Liu, T. 2019. Neural personalized response generation as domain adaptation. *World Wide Web* 22(4):1427–1446.

Zhou, X.; Dong, D.; Wu, H.; Zhao, S.; Yu, D.; Tian, H.; Liu, X.; and Yan, R. 2016. Multi-view response selection for human-computer conversation. In *EMNLP*.

Zhou, K.; Prabhumoye, S.; and Black, A. W. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 708–713.