A Knowledge Driven Dialogue Model With Reinforcement Learning

YONGNAN JIA\textsuperscript{1}, GAOCHEN MIN\textsuperscript{1}, CONG XU\textsuperscript{1,2}, XISHENG LI\textsuperscript{1}, AND DEZHENG ZHANG\textsuperscript{2,3}

\textsuperscript{1}School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
\textsuperscript{2}Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
\textsuperscript{3}School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Corresponding author: Cong Xu (b20140357@xs.ustb.edu.cn)

ABSTRACT In recent decades, many researchers pay a lot of attention on generating informative responses in end-to-end neural dialogue systems. In order to output the responses with knowledge and fact, many works leverage external knowledge to guide the process of response generation. However, human dialogue is not a simple sequence to sequence task but a process heavily relying on their background knowledge about the topic. Thus, the key of generating informative responses is leveraging the appropriate knowledge associated with current topic. This paper focus on addressing incorporating the appropriate knowledge in response generation. We adopt the reinforcement learning to select the most proper knowledge as the input information of the response generation part. Then we design an end-to-end dialogue model consisting of the knowledge decision part and the response generation part. The proposed model is able to effectively complete the knowledge driven dialogue task with specific topic. Our experiments clearly demonstrate the superior performance of our model over other baselines.

INDEX TERMS Dialogue model, policy gradient, knowledge graph, transformer network.

I. INTRODUCTION

Over the past few years, we are highly dependent on a variety of electronic computers in daily life. Previously these ice-cold machines were unable to understand human intentions and provide appropriate services. Today’s dialogue system allows the machine to communicate directly with humans and gives the machine a variety of anthropomorphic images that make the interaction more stereoscopic and vivid between human and machine. However, in order to achieve unimpeded communication with humans, the dialogue system must have rich knowledge and strong ability of decision-making. Thus, the dialogue system will understand the content of the dialogue with background knowledge and recall the relevant knowledge from memory or external resources. Finally the responses with appropriate information will be generated after reasoning these concepts.

In recent years, many researchers have put a lot of effort into the chatbot [1]–[3]. However, most models usually generate universal responses or inconsistent responses when talking about a fixed topic. The reason is that it is difficult for a model to learn semantic interaction patterns only from conversational data without the help of background knowledge. The task of this paper will give a sequence of topics as the purpose of the chat. The model needs to guide the conversation with humans along this given sequence of topics. For example, from “David Lynch” to “Yin and Yang” film.

The existence form of knowledge can be divided into structured and unstructured knowledge. The structured knowledge usually stores in database and knowledge graph. And the unstructured knowledge usually stores in documents, images, audio and so on. Since the knowledge graph has the capability of expressing the relationship and reasoning between the entities, it becomes an important carrier to organize and store knowledge. In the conventional knowledge graph, each piece of knowledge can be represented by a triple (entity-1, relation, entity-2). Due to the interrelation between the entities of knowledge graph, the relationship between topic entities can be represented as the edge in knowledge graph. Therefore, this paper also uses the knowledge graph as the knowledge storage form in this task.

II. RELATED WORK

Recently, researchers have begun to introduce external knowledge into open-domain chat conversations [4], [5]. Ghazvininejad et al. [6] proposed a knowledge-driven neural
network model that incorporates dialogue history and external facts to standardize responses based on the traditional Seq2Seq model. Liu et al. [7] proposed a model to integrate knowledge into the dialogue system. The model can not only match the fact entities in the user input but also diverge into other entities. Using fact matching and entity divergence, it is guaranteed that the responses generated by the dialogue system can diverge and converge in the knowledge graph. The model proposed by Vougiouklis et al. [8] used the alignment of two different data sources to generate context-sensitive and knowledge-containing responses given a series of historical conversations and a set of sentences representing background knowledge. Young et al. [9] studied how to integrate common sense into the retrieval model effectively. Through the Tri-LSTM model, the dialogue system can comprehensively consider input information and common sense for selecting the appropriate response. Zhou et al. [10] proposed a static and dynamic graph attention mechanism to fuse the knowledge graph into the end-to-end chatbot model. The method solves the semantic comprehension problem in the dialogue through common sense. Dinan et al. [11] considered that there is currently no large dataset in the open-domain dialogue, they collected the knowledge retrieved from Wikipedia and designed a model that can retrieve and understand the knowledge to output a natural fluent response.

However, most of the research only integrates knowledge into the process of generating the response, that is, it only uses the similarity between the input sentence and the knowledge, and there is no effective mechanism to ensure that the appropriate knowledge is used in the generation of the response. We make the analogy to the process of using knowledge to conduct dialogues by humans. The human brain accepts external statements and finds relevant knowledge or facts in memory. Then it uses the received sentences to make decisions on the retrieved knowledge and uses appropriate knowledge to generate the last response. It can be seen that obtaining information-rich and appropriate response requires not only relevant knowledge, but also strong decision-making ability to select appropriate knowledge from the relevant knowledge set, as shown in Figure 1.

A feature of the embedded knowledge dialogue system is that it can be extended around a given topic. Namely, it can naturally switch from one chat topic to another related topic. Such ability of topic transformation can make people feel the dialogue system more intelligent and closer to the habit of human chat. This paper proposes an end-to-end neural network model that uses reinforcement learning algorithms to make decisions and generate responses. It mainly solves the problem of how to use the relevant knowledge to smoothly and naturally migrate chat topics during the dialogue process. In the process of chatting, the dialogue system is provided with a topic-related set of knowledge triplet by using entity linking or information retrieval technology. Since this chapter focuses on how to generate knowledge responses from relevant knowledge, the methods of entity linking or graph searching is not described in detail.

Dinan’s model is fit for solving problems in knowledge decision making. The model uses the Memory Network to retrieve relevant information from the knowledge base, and uses the Transformer model to generate a dialog response by receiving the retrieved knowledge set. There are generally two methods of using knowledge, one is a retrieval model that can produce output directly from candidate responses; the other is generating one word at a time by the generation technology to roughly select knowledge and get a much smaller candidate set. Then the model has a fine selection of knowledge from the candidate set.

This work proposes a knowledge retrieval mechanism, assuming that various artificially structured phrases and sentences are stored in a large knowledge graph. The current attention mechanism is unable to handle such a large-scale knowledge base, so you can use the standard information retrieval technology to roughly select knowledge and get a much smaller candidate set. Then the model has a fine selection of knowledge from the candidate set.

Furthermore, in order to make a more precise selection of knowledge, they adopt the attention mechanism to select the knowledge. After the process of rough selection and fine selection, the selected knowledge and the input sentence are concatenated and sent to the Transformer decoder network to finally generate an output response. The structure of the model is shown in Figure 2. This two-stage knowledge selection process can solve the problem that the knowledge graph is too large to be integrated in the end-to-end network effectively.

This paper also adopts this two-stage knowledge selection and proposes a knowledge embedded dialogue model. The rough selection stage also uses the information retrieval technology to generate candidate sets. Then we use the reinforcement learning method to select appropriate knowledge from the candidate set for text generation. And it can be found that Dinan’s model does not consider the response of the previous turn of output when using the attention mechanism for knowledge selection. However, the content of the last turn of response is actually directly related to the knowledge selection. Thus, this paper uses the input statement and response together to guide the knowledge decision process and
integrate them into the response generation process for the open domain task. To achieve this goal, we train a reinforcement learning model to learn how to choose the right knowledge under the conditions of known input sentences and historical responses. Due to the powerful ability of decision-making and the characteristics of interactive learning, the proposed model can select the appropriate knowledge for generating responses.

The main work of this paper are as follows: (1) This paper proposes a knowledge embedded dialogue model uses the reinforcement learning algorithm to make effective decision-making on selecting knowledge and integrates the process of knowledge decision-making into an end-to-end model. Due to the strong decision-making ability of reinforcement learning, the model proposed in this paper can select the appropriate knowledge and generate a response with related information. (2) Compared with other models, the performance of the proposed model is significantly improved in experiments. It also proves that embedding knowledge into the dialogue system can generate more appropriate and more informative responses.

III. COMBINATION MODEL
A. ARCHITECTURE OF MODEL
The system consists of two main parts:

(1) In the knowledge decision-making part, we choose knowledge through two-stage knowledge decision-making. The previous rough selection part is to use the existing information retrieval technology to select the top N knowledge triple as the candidate set. Different from the mentioned rough selection method, we roughly select the subgraph which contains the topic entities and relation from the whole graph. Since the target topic is supposed to be known in this paper, compared with the traditional chatting task, there are some constraints in the response. Therefore, in this paper the rough selection is to retrieve all relevant subgraphs as candidate set from the complete graph according to the target topic triple. Then the fine selection part uses the strategy gradient algorithm of reinforcement learning to choose the most appropriate knowledge in the subgraph as the optimal knowledge.

(2) In response generation part, we adopt the Transformer network to encode and decode input text and knowledge. Transformer network is a deep network only consisting of attention mechanism. It has the advantage of being able to process all the words or symbols of the sequence in parallel. At the same time, owing to the self-attention mechanism, Transformer combines the context with the far words for improving the shortcomings of slow training of RNN network. By processing all words in parallel, each word will notice other words of the sentence in multiple processing steps. The depth of Transformer can also be very deep to fully exploit the specialty of deep neural network models and improve the accuracy of the model. The response generation part and the knowledge decision part are combined and trained at the same time. The training process of the model is similar to multi-task learning.

As shown in Figure 3, we can see that the whole structure of the dialog system is a closed-loop process. The sub-graph extraction method is to extract all sub-graphs containing entities and edges in the topic sequence from the knowledge graph by the information retrieval method. The reinforcement learning algorithm is used to further decision-making and inference on the extracted sub-graphs. By continuously selecting related entities or attributes and outputting the response text, the dialogue model will naturally transfer the content of conversation from the starting topic to the final target topic. The details of the two parts are described below.

B. KNOWLEDGE DECISION
This paper considers the whole process of knowledge decision making as a limited sequence decision problem on the improved knowledge graph. Because this task requires the smooth transition of the topic, the entity should not be skipped. The algorithm should only move one hop in the graph. As shown in Figure 3, this paper considers the process of knowledge decision-making as a Markov decision process. Given the input sentence and the initial topic entity, the reinforcement learning starts from the initial topic entity to find the optimal path and finally reaches the target topic entity through each turn of interactive dialogue. The reinforcement learning agent receives the known topic sequence containing the initial topic and the target topic and the dialogue dataset to find the optimal topic transfer path.

The differences between this task and the previous graph reasoning task are as follows. (1) In this paper the input
of the decision model is not only the triple information of the entity and the edge, but also the topic sequence and context. (2) The task of this paper not only needs to decide which knowledge of the topic entity is selected as background knowledge, but also needs to decide when to switch the topic. The above differences inquire some improvements to the previous model in order adapt to this task. In the following sections, we describe the state, reward, action, and specific definition of the network used in the reinforcement learning algorithm.

State: Defined at time step $t$, state $S_t \in S$ can be characterized by $S_t = (K_t, T_0, X, T_1, H_t)$. Where $K_t$ is the triplet knowledge selected by the current reinforcement learning algorithm, $T_0$ is the initial topic entity, $T_1$ is the target topic entity, $X$ is the sequence of text vectors representing the input, and $H_t$ represents the history of the conversation.

Action: The conventional knowledge graph inference task usually defines the action as the value of all edges, which means that the reinforcement learning algorithm needs to decide which side to go from the current entity node to the next entity node. Differing from the previous knowledge graph inference task, the transfer sequence of the entity node is known in our task and the decision is which triplet of the decision model is not only the triple information of the entity and the edge, but also the topic sequence and context. (2) The task of this paper not only needs to decide which knowledge of the topic entity is selected as background knowledge, but also needs to decide when to switch the topic. The above differences inquire some improvements to the previous model in order adapt to this task. In the following sections, we describe the state, reward, action, and specific definition of the network used in the reinforcement learning algorithm.

State: Defined at time step $t$, state $S_t \in S$ can be characterized by $S_t = (K_t, T_0, X, T_1, H_t)$. Where $K_t$ is the triplet knowledge selected by the current reinforcement learning algorithm, $T_0$ is the initial topic entity, $T_1$ is the target topic entity, $X$ is the sequence of text vectors representing the input, and $H_t$ represents the history of the conversation.

Action: The conventional knowledge graph inference task usually defines the action as the value of all edges, which means that the reinforcement learning algorithm needs to decide which side to go from the current entity node to the next entity node. Differing from the previous knowledge graph inference task, the transfer sequence of the entity node is known in our task and the decision is which triplet knowledge of the topic entity is selected as background knowledge, but also needs to decide when to switch the topic.

Policy Network: in this paper the policy of network learning is how to take correct actions for knowledge selection. A policy network is a parameterized probability map between action space and degree of belief.

$$
\pi_\theta(a|s) = \pi(a|s; \theta) = \mathbb{P}(A_t = a|S_t = s, \theta_t = \theta)
$$

where $\theta$ is the parameter vertex, which is the weight of the neural network. To train the policy network, Williams and Sutton proposed a policy gradient algorithm. The policy gradient network algorithm is a model-free algorithm that directly fits the policy by adjusting the network parameters. The parameters are trained by a gradient optimization algorithm. We need to define the objective function $J$ that can guide the parameter search. This objective function measures the quality of the policy. One defined method is to use the average return value received by the agent, and the other is to calculate the weighted sum of all returns for each sample trajectory.

This paper designs a policy network consisting of two networks to decide which action should be taken currently based on the input $X$. The first part is the feed-forward network used to encode the conversation history. The LSTM is used to encode the conversation history $h_t = (H_t, A_{t-1}, O_t)$ as a continuous vector $h_t$, where $H_t$ and $A_{t-1}$ is an observation and action sequence, which is updated by LSTM.

$$
h_t = \text{LSTM}(h_{t-1}, [a_{t-1}; o_t])
$$

$$
e_X = \text{ReLU}(W_X e(X) + b_X)
$$

where $a_{t-1}$ and $o_t$ represent the vector of the action and the current state respectively, and the semicolon indicates that knowledge triplet and every word in response are computed respectively. We define the highest value of relatedness scores between the head entity and every word of response as $r_1$. Then we get $r_2$ and $r_3$ through the same operation of relation and tail entity. The reward is the sum of $r_1$, $r_2$ and $r_3$.

Policy Network: in this paper the policy of network learning is how to take correct actions for knowledge selection. A policy network is a parameterized probability map between action space and degree of belief.

$$
\pi_\theta(a|s) = \pi(a|s; \theta) = \mathbb{P}(A_t = a|S_t = s, \theta_t = \theta)
$$

where $\theta$ is the parameter vertex, which is the weight of the neural network. To train the policy network, Williams and Sutton proposed a policy gradient algorithm. The policy gradient network algorithm is a model-free algorithm that directly fits the policy by adjusting the network parameters. The parameters are trained by a gradient optimization algorithm. We need to define the objective function $J$ that can guide the parameter search. This objective function measures the quality of the policy. One defined method is to use the average return value received by the agent, and the other is to calculate the weighted sum of all returns for each sample trajectory.

This paper designs a policy network consisting of two networks to decide which action should be taken currently based on the input $X$. The first part is the feed-forward network used to encode the conversation history. The LSTM is used to encode the conversation history $h_t = (H_t, A_{t-1}, O_t)$ as a continuous vector $h_t$, where $H_t$ and $A_{t-1}$ is an observation and action sequence, which is updated by LSTM.

$$
h_t = \text{LSTM}(h_{t-1}, [a_{t-1}; o_t])
$$

$$
e_X = \text{ReLU}(W_X e(X) + b_X)
$$

where $a_{t-1}$ and $o_t$ represent the vector of the action and the current state respectively, and the semicolon indicates that

FIGURE 3. The architecture of our model.
the two vectors are spliced. The previous section mentions that the action represents the outgoing edge and the switching topic in the graph. And the state consists of the historical state \( h_t \), the current topic entity \( e_t \), and the sentence vector \( e_X \) of the input sentence \( X \). The sentence vector is calculated using a single-layer feed-forward network. Then the probability output of the decision network can be expressed as follows.

\[
P_{\text{decision}}(A_t|s_t) = \text{softmax}(W_2 \text{ReLU}(W_1 [h_t; a_t; e_X] + b_1) + b_2) \tag{4}
\]

By calculating the above-mentioned probability distribution covering all actions, we can select a discrete action \( a_t \) with the highest probability and select its corresponding triple knowledge \( k_t \) as the output.

\[
k_t \sim a_t = \arg \max_{i-L} P_{\text{decision}}(a_t|s_t) \tag{5}
\]

The decision network will be taken as input of the generating network at the same time. Then at the next turn, the decision network decides to generate new background knowledge according to the new state and continues to guide the generating process. The structure of the entire knowledge decision is as follows.

In order to train the policy network \( \pi_e \) mentioned above, we look for the parameter \( \theta \) that maximizes the expected reward.

\[
J(\theta) = \mathbb{E}_{a_1, \ldots, a_{t-1} \sim \pi_e} [R(S_T)|S_0 = (e_0, l_0, e_1)] \tag{6}
\]

where \( R \) represents the reward function. The parameters of network are adjusted by gradient back propagation until the network is convergent.

**C. RESPONSE GENERATION**

The response generation part adopts two different generation methods. One is direct generation, that is, directly using the selected knowledge and the input sentence \( X \) as the input of the Transformer decoder. The other is indirect generation. The input of the Transformer decoder is the context vector which is the output of a Transformer encoder through the knowledge \( k \) of the input sentence \( X \).

The whole model consists of the above knowledge decision part and response generation part. The two type of complete network structures are shown in Figure 5.

In Figure 5 the input sentence \( X \) will be directly sent to the Transformer decoder after passing through the LSTM encoder. Input sentence \( X \) isn’t encoded and is fed directly into the process of response generation. Another way is to transform the selected knowledge and input sentences into a context vector through the Transformer encoder and send the context vector to the Transformer decoder, as shown in Figure 6. In the experiment, we evaluate the performance of these two models separately.

**D. MODEL TRAINING**

The proposed model has two training objectives, one is to reduce the gap between the generated response and the target response, and the other is to correctly select the appropriate background knowledge. Therefore, the objective function of our model consists of two parts. The target function of the response generation part is Log-Likelihood.

\[
L_1 = \frac{1}{m} \sum_{i=1}^{m} P_{\theta}(Y_i|Y_{1:t-1}, X, K) \tag{7}
\]

where \( Y_i \) represents the word in target sentence at time step \( t \), \( Y_{1:t-1} \) represents the partial sentence before time step \( t \). The target function of knowledge decision network is expressed by:

\[
L_2 = \mathbb{E}_{a_1, \ldots, a_{t-1} \sim \pi_0} [R(s_t, a_t)] \tag{8}
\]

Finally, we combine the above functions as the final target function \( L \).

\[
L = L_1 + L_2 = \frac{1}{m} \sum_{i=1}^{m} P_{\theta}(Y_i|Y_{1:t-1}, X, K) + \mathbb{E}_{a_1, \ldots, a_{t-1} \sim \pi_0} [R(s_t, a_t)] \tag{9}
\]

**IV. EXPERIMENT**

**A. DATASET**

This paper adopts Baidu’s collection of dialogue dataset containing domain knowledge. The knowledge in the dataset comes from the dialogue with the useful information in the field of movies and entertainment characters, such as box office, director, evaluation, etc. The knowledge is in the form of triplet SPO organization. The topic in the dialogue goal is a movie or entertainment character entity. There are 30,000 sessions in the data sets, about 120,000 turns of dialogue, including 100,000 examples for the training set, 10,000 examples for the development set and 10,000 examples for the test set. The given topic in the data set is a specific entity, such as a movie name or a star name. Each dialogue is marked by two tags, one emulating the role of the conversation agent and the other simulating the role of the user. The dialogue agent needs to use the background knowledge to guide the conversation in the order of the topic. The user does not need to have any background knowledge and only needs to communicate with the dialogue agent naturally. In our experiment, we collect all retrieved triplet given in datasets as a whole knowledge graph and the retrieved entities involved in every model are obtained by the same IR system from the whole knowledge graph.

Each dialogue in the training data contains the goal of the dialogue (topic sequence), background knowledge and dialogue data pairs. Each item of the data is as below: Dialogue goal: Contains two lines, the first line is the path of the
conversation topics, such as [Start, TOPIC\textsubscript{A}, TOPIC\textsubscript{B}], and the second line is the relation between topic A and topic B. Knowledge: The original data directly provides knowledge related to all current topic entities, that is, the task directly gives the knowledge subset obtained after the rough selection process. Dialogue: Dialogue rounds are generally 4 to 8 turns.

**B. SETTING**

This paper conducts experiments to compare the Transformer RL model which is proposed in this paper with the Seq2Seq model, the HRED model (hierarchical recurrent encoder-decoder model) [12], the MemoryNet model [6], the GTTP model [13] and the Transformer MemNet [11] model.
**Seq2Seq**: Using a sequence-to-sequence model which is widely used in the open domain.

**HRED**: Hierarchical recursive coding-decoding model.

**MemoryNet**: A memory network conversation model that combines with end-to-end knowledge.

**GTTP**: This is an end-to-end text summary model proposed in 2017. It will be used in the process of generating dialogue responses for fusion knowledge in the experiment.

**Transformer MemNet**: As mentioned above, this model is quite suitable for this task. It is used in this experiment after a slight change in the input in this paper.

In our model the hyper-parameters of Transformer network are the same as described in the original paper. The policy network consists of a 128 units bidirectional LSTM and a softmax layer. Word embeddings are set as 512 dimensions and randomly initialized. When training our model, we adopt mini-batch training with batch size 128. The optimization algorithm is Adam [6] and the threshold of gradient clipping is set to 5. Our code is implemented in Pytorch and runs on two GPU device GTX 2080Ti. We adopted the BLEU-2 [14] and ROUGE-L [15] which are commonly used in previous studies to evaluate the similarity between the output response of the model and the target sentence.
TABLE 1. The experimental results on dialogue dataset.

| Model                   | BLEU-2 | ROUGE-L | DISTINCT2 | F1 score (%) |
|-------------------------|--------|---------|-----------|--------------|
| Seq2Seq                 | 0.098  | 0.155   | 0.081     | 7.98         |
| IRDSD                   | 0.063  | 0.210   | 0.094     | 11.32        |
| MemoryNet               | 0.103  | 0.187   | 0.127     | 24.15        |
| GTPP                    | 0.201  | 0.339   | 0.072     | 18.50        |
| Transformer MemNet      | 0.265  | 0.288   | 0.312     | 42.47        |
| Transformer RL (directly)| 0.253  | 0.316   | 0.274     | 47.75        |
| Transformer RL (indirectly)| 0.281  | 0.371   | 0.297     | 46.83        |

TABLE 2. Sample responses generated by Transformer RL and Transformer MemNet.

| Goal | Transformer RL | Transformer MemNet |
|------|----------------|--------------------|
| Agree| Do you usually like to watch action movies? | Have recently recommended a good movie to you. |
| User | Occasionally. Do you have any recommendation? | What movie is it? |
| Agree| This Sweet Alibis have a good reputation in Minsk. | You can watch Sweet Alibis by Lin Baihong. |
| User | Yes. Did you watch it? | Can you watch Sweet Alibis by Lin Baihong? |
| User | Yes, I’d watch it right away. | I’d watch it right away. |
| User | I haven’t seen that film. | I also haven’t seen that film. |
| User | We should go to the theatre. | We should go to the theatre. |
| Agree| Lin Baihong did a good job in the movie. | The story of the movie is very good. |

DISTINCT2 is used to evaluate the diversity of generated responses. The F1 score is used to assess the accuracy-recall of the output response relative to the standard response at the word level. All experimental results are listed in Table 1. As shown in Table 1, the F1 score is used to evaluate the knowledge decision accuracy of the model. The proposed Transformer RL (directly) model obtains the highest accuracy of knowledge selection, and the Transformer RL (indirectly) obtains the second highest accuracy rate. This shows that the reinforcement learning algorithm in the proposed models can better select the background knowledge, which is 5.28% and 4.36% higher than the third model. In terms of the degree of conformity between the generated response and the target text, the model in this paper achieves better results than other benchmark models in both BLEU-2 and ROUGE-L, which indicates that the responses part of the model can learn the human dialogue pattern from the dialogue data sets very well. From the diversity of text generation, Transformer MemNet can generate more diverse responses, and the model proposed in this paper has obtained only behind the scores of Transformer MemNet.

C. CASE STUDY

The results of the representative responses generated in the experiment are shown in Table 2. As shown in Table 2, the proposed model can complete well the task of changing topic. At the beginning, the topic ‘Sweet Alibis’ is introduced by the knowledge of [sweet killer, type, action film], and then through [Sweet Killer, Time Network Rating, 6.2] the model chats with the user about the related topics and starts switching topic according to the topic sequence. Finally, the model uses [Sweet Killer, Starring, Lin Baihong] to switch to the topic entity “Lin Baihong” and continues to communicate with the user under this topic ‘Lin Baihong’. In contrast, we can see that Transformer MemNet is also very good at communicating with users, but it has not learnt how to use the appropriate knowledge and when to switch to the topic smoothly.

V. CONCLUSION

Based on the purpose of enhancing the intelligence of the dialogue system, this paper proposes a dialogue model that uses the reinforcement learning algorithm to make knowledge decision. The model includes a knowledge decision part by using the reinforcement learning algorithm and a response generation part based on the Transformer model. The accuracy and efficiency of the proposed model in knowledge decision-making and response generation tasks are proved in experiments. This paper has done some exploration work in the open domain dialogue task of fusion knowledge, but the current knowledge decision-making ability of the model is limited to the choice of given knowledge, not all the knowledge on the Internet. And the knowledge used in this paper is a structured factual knowledge graph, the usage of non-factual knowledge or unstructured knowledge remains to be further explored.

REFERENCES

[1] C. Xu, Q. Li, D. Zhang, J. Cui, Z. Sun, and H. Zhou, “A model with length-variable attention for spoken language understanding,” *Neurocomputing*, vol. 379, pp. 197–202, Feb. 2020. Accessed: Jul. 24, 2019, doi: 10.1016/j.neucom.2019.07.112.

[2] B. Wu, B. Wang, and H. Xue, “Ranking responses oriented to conversational relevance in chat-bots,” in *Proc. COLING*, Osaka, Japan, 2016, pp. 652–662.

[3] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li, “Building task-oriented dialogue systems for online shopping,” in *Proc. AAAI*, San Francisco, CA, USA, 2017, pp. 652–662.

[4] Z. Xu, B. Liu, C. Sun, and X. Wang, “Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AK, USA, May 2017, pp. 3506–3513.

[5] L. Yao, Y. Zhang, Y. Feng, D. Zhao, and R. Yan, “Towards implicit content-introducing for generative short-text conversation systems,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 2190–2199.

[6] M. Ghazvininejad, C. Brockett, and M. Chang, “A knowledge-grounded neural conversation model,” in *Proc. AAAI*, New Orleans, LA, USA, 2018, pp. 5110–5117.

[7] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, and D. Yin, “Knowledge diffusion for neural dialogue generation,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 1489–1498.

[8] P. Vougiklous, J. Hare, and E. Simperl, “A neural network approach for knowledge-driven response generation,” in *Proc. COLING*, Osaka, Japan, 2016, pp. 3370–3380.

[9] T. Young, E. Cambria, I. Chaturvedi, and H. Zhou, “Augmenting end-to-end dialogue systems with commonsense knowledge,” in *Proc. AAAI*, New Orleans, LA, USA, 2018, pp. 4970–4977.

[10] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, “Commonsense knowledge aware conversation generation with graph attention,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 4623–4629.

[11] E. Dinan, S. Roller, K. Shuster, A. Fan, and M. Auli, “Wizard of wikipedia: Knowledge-powered conversational agents,” in *Proc. ICLR*, New Orleans, LA, USA, 2019, pp. 1–18.
[12] I. Serban, A. Sordoni, R. Lowe, and L. Charlin, “A hierarchical latent variable encoder-decoder model for generating dialogues,” in Proc. AAAI, San Francisco, CA, USA, 2017, pp. 3295–3301.

[13] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra, “Towards exploiting background knowledge for building conversation systems,” in Proc. Conf. Empirical Methods Natural Lang. Process., Brussels, Belgium, 2018, pp. 2322–2332.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), Philadelphia, PA, USA, 2001, pp. 311–318.

[15] C.-Y. Lin and E. Hovy, “Manual and automatic evaluation of summaries,” in Proc. Workshop Autom. Summarization (ACL), Barcelona, Spain, 2002, pp. 74–81.

YONGNAN JIA received the Ph.D. degree from the Intelligent Control Laboratory, College of Engineering, Peking University, in 2014. She is currently a Lecturer with the School of Automation and Electrical Engineering, University of Sciences and Technology Beijing. Her research interests include intelligent robots, biomimetic robots, and collective behaviors of multirobot systems and machine learning.

GAOCHEN MIN is currently pursuing the bachelor’s degree with the University of Science and Technology Beijing, China. He attended the Student Research Training Program on path planning for manipulator, in 2019. His current research interests include intelligent optimization and artificial intelligence.

CONG XU is currently pursuing the Ph.D. student with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, China. His research interests include deep reinforcement learning, dialogue systems, and machine translation.

XISHENG LI received the Ph.D. degree from the University of Science and Technology Beijing (USTB), China, in 2000. He is a Professor with the School of Automation and Electrical Engineering, USTB. His research interests include advanced sensing technology, information fusion, and geomagnetic navigation.

DEZHENG ZHANG received the Ph.D. degree from the University of Science and Technology Beijing (USTB), China. He is a Professor with the School of Computer and Communication Engineering, USTB and the Director of the Beijing Key Laboratory of Knowledge Engineering for Materials Science. His research interests include knowledge acquisition, big data mining, artificial intelligent, and deep learning with satellite image.