ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

Research on Adaptive Dialogue Strategies Combining Retrieval and Generative Chatbots

Siyi Wu[1,a], Weizhi Luo[2,b], Zihao Wang[2,c], Junxi Li[3,d,*]

[1]*Nanjing University of Science and Technology, Nanjing, China*
[2]*Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia*
[3]*Herzen University, Saint Petersburg, Russia*
[a]*chownate623@163.com,* [b]*lo2.v@edu.spbstu.ru,* [c]*van39.ts@edu.spbstu.ru,* [d]*lijunxi111@yandex.com*
[*]*Corresponding author*

Abstract: *During the development of chatbot technology, we are faced with the challenge of how to improve the interaction capability and user experience of bots. As technology continues to advance, we propose novel architectures that combine retrieval-based chatbots and generative chatbots, aiming to improve the interaction capability and user experience of chatbots. Retrieval-based bots rely on an existing knowledge base to answer questions, while generative bots can generate coherent dialogues. Combining the advantages of both, retrieval-based bots can quickly find information and then use generative bots to generate more natural dialogue. Experiments show that this architecture greatly improves the interaction capabilities and user satisfaction of chatbots.*

Keywords: *Retrieval Robot; Generative Chatbot; Generative Network; Attention Mechanism*

1. Introduction

The combination of retrieval-based and generative robots has become a popular research topic in today's artificial intelligence field. We aim to explore how to combine the powerful search capabilities of retrieval robots with the innovative generative capabilities of generative robots to improve the efficiency and quality of AI in a variety of application areas. Through an in-depth analysis of the strengths and limitations of these two types of robots, we will explore how to optimize their interaction for more efficient and accurate information processing and generation.

Retrieval robots rely on pre-stored knowledge bases, which limit the range of questions they can answer [1]. Retrieval robots are usually unable to understand context or perform complex reasoning, making it difficult to provide satisfactory answers when dealing with questions that require deeper understanding [2]. Since retrieval robots base their answers on predefined templates or knowledge bases, the answers they generate may be too fixed and lack personalization and innovation [3].

Generative robots may have inconsistent quality when generating text, where some of the output may be inaccurate, incoherent, or lacking in logic [4]. Generative robots typically require large amounts of data for training, especially when generating high-quality content, which can lead to high demands on computational resources and data [5]. Generative robots may be influenced by misleading information when generating content, especially when dealing with sensitive topics or involving false information, often giving answers that contradict common sense [6].

In this paper, we mainly discuss the combination of retrieval robots and generative robots, which is a very cutting-edge and challenging research field. In this field, retrieval bots rely mainly on pre-stored information libraries to answer questions, while generative bots are able to generate new text or answers based on input. Combining the advantages of these two robots can achieve a more efficient and flexible dialogue system.

Our retrieval-based robotic architecture relies heavily on Convolutional Neural Networks (CNNs) and Self-Attention Mechanisms (SAMs) to process and retrieve information. CNNs automatically extract features from an image or text through a combination of Convolutional, Pooling, and Fully-Connected Layers in order to perform accurate information retrieval. The self-attention mechanism, on the other hand, enables the model to process sequence data by assigning different attentional weights to different parts of the input sequence to better capture the important information in the sequence. In addition, neural structure search is used as an automated method for searching the optimal neural

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

network structure for efficient information retrieval.

Generative Class Robot Architecture We use Long Short-Term Memory Networks (LSTMs) as well as Transformer structures and Generative Adversarial Network (GANs) neural network architectures. LSTMs are able to process long sequential data and capture long term dependencies, Transformer structures are based on self-attentive mechanisms to process sequential data in parallel and increase the efficiency of generating text, and GANs generate high-quality new data through the adversarial process of generators and discriminators to generate new data with high quality.

We evaluated our architecture in three different ways, firstly assessing the consistency of the answers generated by the system with the actual answers. Secondly, to assess the performance of the system in dealing with different types of questions and its ability to generate new content according to the needs of the user. Finally, we evaluate whether the generated content is of high quality, including language fluency, content accuracy and innovation.

2. Architecture

2.1 Architecture Design Principles

We recognize that different user groups may have different question frameworks and habits of using certain terms or slang. Therefore, one of our design principles is to enable chatbots to adapt to these variations to improve their effectiveness and user satisfaction.

Combining the advantages of both retrieval and generation chatbots, our novel architecture aims to combine the advantages of both in the dialogue process to improve the interaction capability and user experience of the dialogue system. First, through the accuracy and speed of retrieval-based bots, we can ensure that chatbots can quickly find and provide relevant information. At the same time, by using the responses of the retrieval bots as the source data input for the generative bots, we can generate more varied and innovative responses while ensuring the coherence and fluidity of the dialogue.

This combined method can not only improve the interactive capabilities of the chatbot, but also increase user satisfaction, allowing the chatbot to better meet the needs of users.

2.2 Datasets Construction

The dataset we use consists of three main parts, the first is the Ubuntu Dialogue dataset, the second is the Persona Dialogue dataset, and the third is the Dialogue NLI dataset, as shown in table 1.

Development of the dialogue repository Ubuntu Dialogue Corpus, built by Lowe et al. It is the largest public dialogue dataset currently available [7]. Persona dialogue dataset, a new dialogue dataset consisting of discourse between randomly paired crowdsourced workers and each asked to play a given role [8]. The Dialogue NLI dataset is a natural language inference dataset specifically designed for dialogue modelling, which contains a series of sentence pairs that are labelled as implicit, neutral, or contradictory [9].

Table 1: Chatbot dataset

Dataset name	Manual annotation	Dataset language	Training data size
Ubuntu Dialogue	N	English	1,000,000
Persona Dialogue	Y	English	10,907
Dialogue NLI	Y	English	310,110

The combined use of these three datasets to train chatbots has the following advantages:

1) The combined dataset provides a large amount of multi-round dialogue data, which is very helpful for training chatbots to understand and generate natural language dialogues.

2) The comprehensive dataset provides information on character descriptions and dialogue context, which is important for understanding the intent of the dialogue and generating more accurate responses.

3) The integrated dataset covers different types of dialogues, and this diversity can help chatbots better adapt to various dialogue scenarios.

4) The comprehensive dataset, models can be trained to better understand and predict implicit, neutral, or contradictory relationships in dialogues, thus improving the consistency and naturalness of

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

dialogues.

2.3 Component Modules and Functions

All our research is based on two large modules. The first major module is based on the architecture of retrieval robots and the second major module is based on the architecture of generative modules.

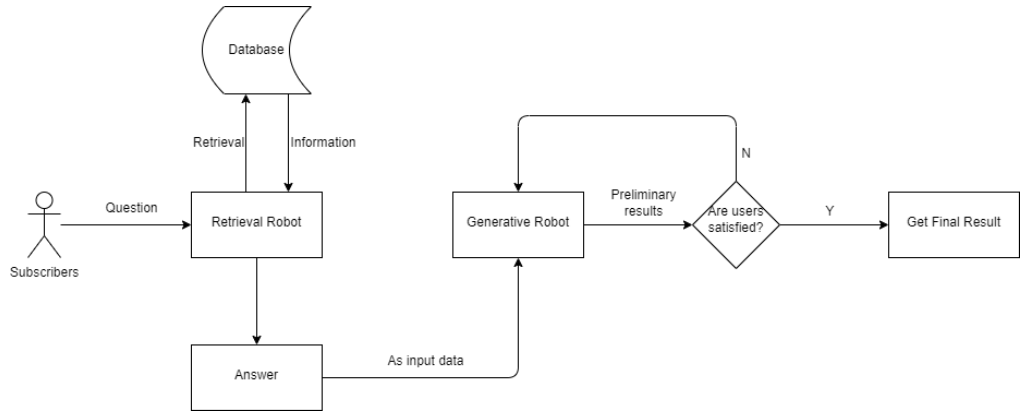

Figure 1: Chatbot Overall Architecture

As shown in figure 1, in the first stage, our users ask a question, our retrieval robot retrieves the relevant content of the question from a comprehensive database according to the question asked, and then passes the information to our retrieval robot, and finally outputs the results.

In the second stage, we take the output from the retrieval robot as input to our generative robot, which reorganizes and completes the relevant answers, and then asks if the user is satisfied with the question and answer, and if not, it returns to the generative robot to reorganize the language to generate a new answer, and vice versa to output the final answer.

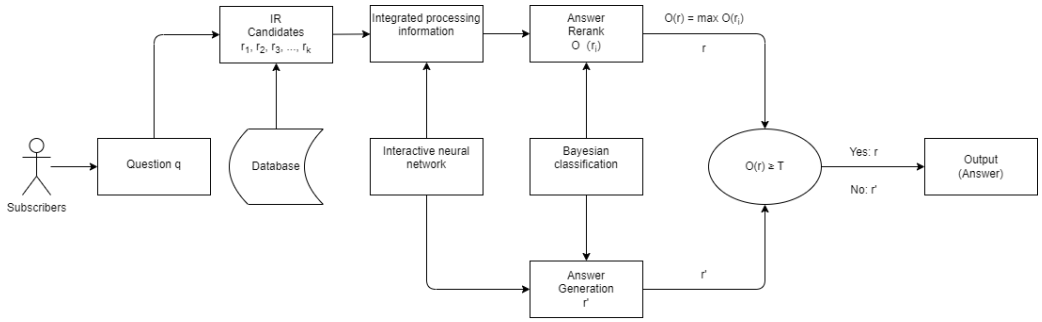

Figure 2: Retrieval Robot Module Architecture

As shown in figure 2, the core problem we solve in retrieval robots using traditional Bayesian classification methods is classification. An interactive neural network model using convolutional neural network and self-attention mechanism is used to process and retrieve the information, and then the match between the question and the response is calculated.

In the retrieval robot phase, after the user asks a question, an alternating number of answers are retrieved from the database and then processed by our interactive neural network to obtain the more comprehensive information, and finally our Bayesian approach classifies the answers, compares the scores, and outputs the answer with the highest score.

The interaction-based neural network module captures both the important information in the context and replies and understands the relationships between individual sentences [10].

The interaction-based neural network consists of three main modules: dialogue branch-response match computation, match accumulation, and match prediction.

These three modules are implemented by different layers of the neural network. Given a dialogue history $c=\{u_1, u_2, \ldots, U_n\}$ and a candidate response r, the method first computes a vector of matches between each sentence in the dialogue history and the candidate response.

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

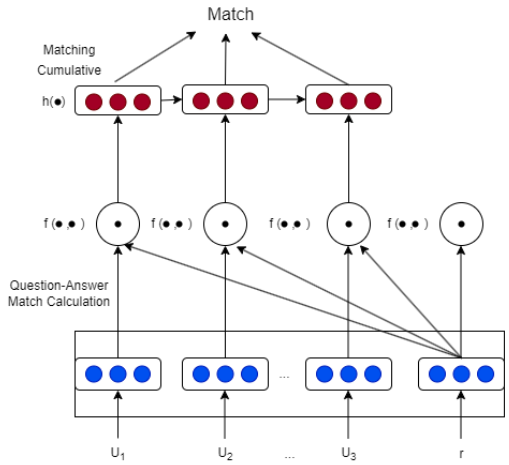

Figure 3: Interaction-based neural network module

As shown in figure 3, after obtaining these matching vectors, the vectors are passed to the second layer for the next computation, which models the dependencies of the different sentences in the dialogue history and then computes the final matching equivalence scores.

A Bayesian classifier is a classification method based on Bayes' theorem that makes classification decisions by calculating the posterior probability of each category [11]. The core idea of a Bayesian classifier is to use the information in the training dataset to estimate the prior probability of each category and the conditional probability of each feature given the category, and then classify by maximizing the posterior probability.

When we train through multiple rounds of the interactive neural network module, we get multiple matches for the question and then classify them using a Bayesian classifier to finally get the answer that best matches the question and output it to our generative robot.

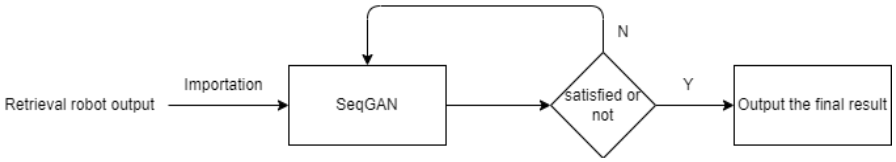

Figure 4: Generative Robot Module

As shown in figure 4, the full name of SeqGAN is Sequence Generative Adversarial Nets with Policy Gradient. SeqGAN is a new approach to training generative models that solves the problem of generating sequential data by using decision-making strategies from reinforcement learning [12]. SeqGAN makes the process of generating sequential data more efficient by passing the generator differentiation problem by performing gradient policy updates directly.

The generative robot takes the initial answer given by our retrieval robot as input and reorganizes the language through SeqGAN for output; if the user is satisfied with the generated answer, the final result is output directly; if the user is not satisfied with the style or content of the output result, it must go back to SeqGAN to generate it again and repeat the process until the user is satisfied.

3. Evaluation

We tested the combinatorial bots on both business logic and dialogue comprehension. For business logic, we asked the combined robot complex logic questions and tested whether the combined robot could fully understand the meaning of the questions and give correct answers. For dialogue understanding, we acted as a user and asked the combined robot questions and observed whether the combined robot's answers matched the scenarios in which the questions were asked.

The first test is on complex logical thinking topics, mainly on logical reasoning topics, puzzles, cognitive skills, and paradoxical topics.

It was first tested using logical reasoning questions, where we provided a series of sentences with

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

conditions and conclusions and asked the robot to answer relevant questions based on this information.

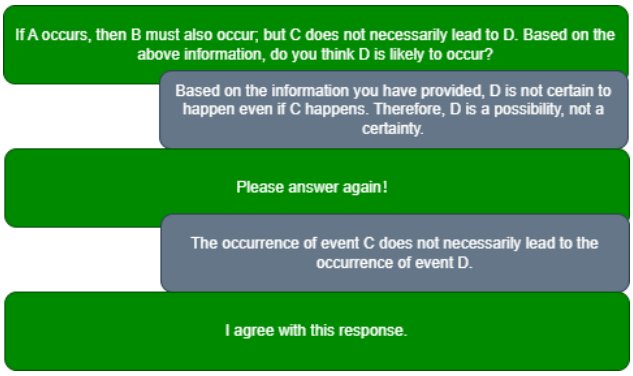

Figure 5: Logical Thinking Topic Test

As shown in figure 5, if we ask a more complex logical question, our robot will answer according to the question, and if we are not satisfied, we can ask the robot to answer again until we are satisfied.

Next, we will test our robots with puzzles that require them to reason and analyze.

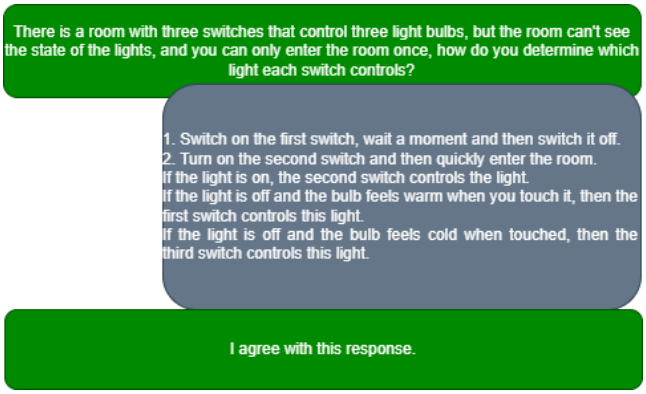

Figure 6: Riddle Test

As shown in figure 6, when we are given riddle type questions, we usually wait a little longer because the reasoning is more difficult. Here we can see that the robot gives reasonable answers to the puzzles.

Next, we will use cognitive questions to test whether the robot can extract some scenes or stories with implicit information, and ultimately require the robot to be able to understand the meaning and draw relevant conclusions.

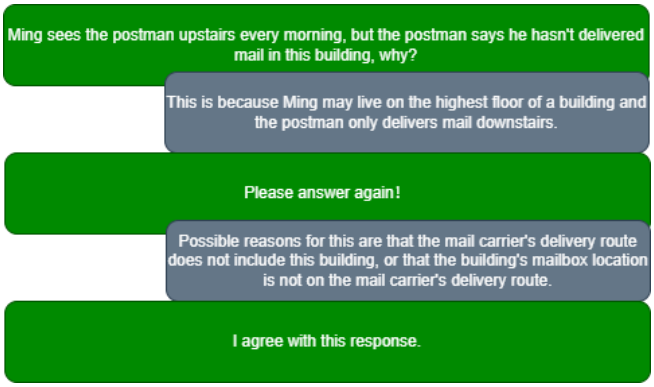

Figure 7: Cognitive Ability Test

As shown in figure 7, when we tested the robot's cognitive abilities, the first answer was not very reasonable; we asked the robot to answer the question again, and the second answer was more reasonable.

Finally, we tested with paradox questions, whose main purpose is to provide scenarios that may be

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

confusing or contradictory for the robot to reason and explain.

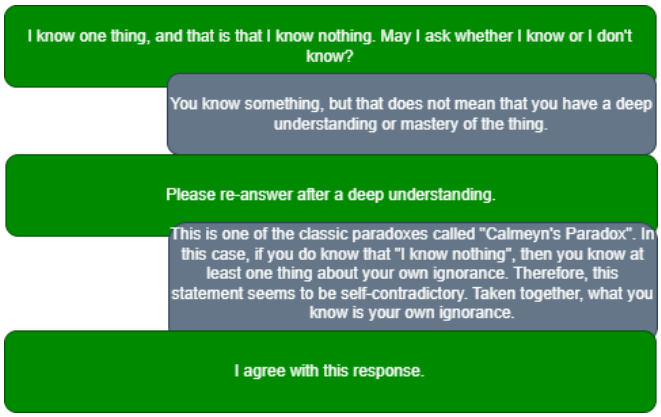

Figure 8: Paradox Test

As shown in figure 8, when we did the paradox test, we were given a Socratic maxim, but the robot didn't understand it well; when we asked again for a deeper understanding of the question, the second time the answer was close to what we humans understand, but of course there was still a gap, and we think that the lack of a philosophical corpus dataset when the robot was trained caused these problems.

The second phase of the test was dialogue comprehension, where we set up a dialogue about travel plans. We played the role of a traveler and the robot was responsible for answering the questions we asked. The main topics are location, climate, culture, and cuisine, as well as transport and accommodation. Through such dialogue scenarios, we can investigate the robot's overall performance in dialogue comprehension, information extraction, recommendation suggestions and human-robot interaction.

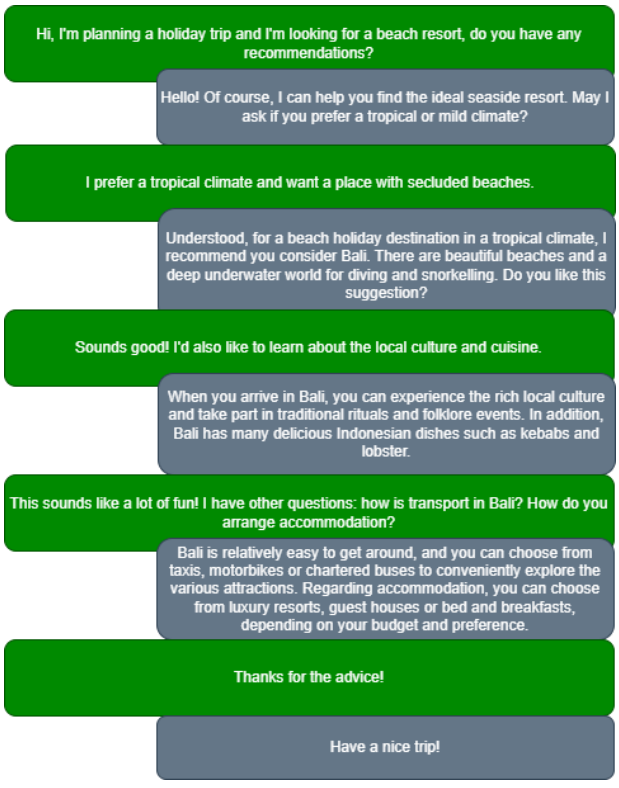

Figure 9: Dialogue comprehension test

As shown in figure 9, the robot's questions and answers during the dialogue are able to follow the logic smoothly and meet the requirements of the dialogue scenario. This indicates that the robot is able to correctly analyze the content of the dialogue and respond accordingly.

The robot's ability to correctly understand the user's needs and questions, such as looking for a beach holiday resort, preferring a tropical climate, etc., shows that the robot has some information

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

extraction capabilities.

The robot is able to provide appropriate recommendation suggestions based on the user's preferences and needs, for example, recommending Bali as a tropical beach holiday location and introducing the local culture, cuisine, and transport.

The fluency and naturalness of the dialogue between the user and the robot can influence the overall interaction experience, which can be considered good if the user is able to clearly express needs and the robot is able to accurately understand and provide valuable responses.

The boat is able to keep the dialogue on topic as the user continues to ask questions, and can provide sound advice and information, such as explaining transport options and arranging accommodation.

We tested the robot on business logic and dialogue comprehension, and it performed relatively well in all tests. This shows that our robot performs relatively well in both business logic and dialogue comprehension, indicating that the robot meets the expected standard of competence in these two key areas, and is able to effectively process and understand user needs and commands.

4. Experimental Results

We conducted a total of 2000 rounds of testing on the robots in terms of business logic and dialogue understanding, and the results of the tests showed that the robots were able to answer the questions smoothly 1914 times, with a pass probability of 95.7%, and 86 times in the process of answering the questions with delay, long waiting time and non-answers to the questions, with a fail probability of 4.3%.

Table 2: Experimental results statistics

Category	Test quantity	Pass quantity	Probability of passing
Logic Test Questions	400	386	96.5%
Riddle Test Questions	400	395	98.8%
Cognitive Ability Test	400	389	97.3%
Paradox questions	400	347	86.8%
Dialogue comprehension test	400	397	99.3%
Total	2000	1914	95.7%

As can be seen from Table 2, the worst performing section of our model is the Paradox Questions test, with the lowest pass rate and a pass probability of 86.8%; the best performing section is the Dialogue Comprehension test section, with the highest pass rate and a pass probability of 99.3%. Improving the pass rate of the paradox test can therefore be a focus of our follow-up work. However, the overall performance of the model is good.

5. Conclusion

We propose an innovative architecture that combines a retrieval-based chatbot and a generative chatbot to improve the interaction capabilities and user experience of chatbots. By leveraging the knowledge base of retrieval-based bots to quickly find information and combining the capabilities of generative bots to generate more natural dialogues, we successfully improve the overall performance of chatbots. Experimental results show that this converged architecture performs well in both business logic and dialogue comprehension tests, effectively improving chatbot interaction capabilities and user satisfaction. This finding not only provides strong support for our research work, but also contributes useful explorations and practical experiences to the development of the chatbot field.

References

[1] Wang Z,Chen A,Tao K, et al. MatGPT: A Vane of Materials Informatics from Past, Present, to Future [J]. Adv Mater. 2024;36 (6):e2306733. doi:10.1002/adma.202306733
[2] Luo L,Ogawa K,Peebles G, et al. Towards a Personality AI for Robots: Potential Colony Capacity of a Goal-Shaped Generative Personality Model When Used for Expressing Personalities via Non-Verbal Behaviour of Humanoid Robots [J]. Front Robot AI. 2022; 9:728776. doi:10.3389/frobt.2022.

ISSN 2616-5775 Vol. 7, Issue 11: 1-8, DOI: 10.25236/AJCIS.2024.071101

728776
[3] Borges, R.M. A Braver New World? Of chatbots and other cognoscenti [J]. J Biosci, 2023, 48, 10. https://doi.org/10.1007/s12038-023-00334-6
[4] Nishimura Y,Nakamura Y,Ishiguro H. Human interaction behavior modeling using Generative Adversarial Networks [J]. Neural Netw. 2020;132:521-531. doi:10.1016/j.neunet.2020.09.019
[5] Prescott TJ,Camilleri D,Martinez-Hernandez U, et al. Memory and mental time travel in humans and social robots [J]. Philos Trans R Soc Lond B Biol Sci. 2019; 374 (1771):20180025. doi:10.1098/rstb. 2018. 0025
[6] Rizvi SKJ, Azad MA,Fraz MM. Spectrum of Advancements and Developments in Multidisciplinary Domains for Generative Adversarial Networks (GANs) [J]. Arch Comput Methods Eng. 2021;28 (7):4503-4521. doi:10.1007/s11831-021-09543-4
[7] Lowe R, Pow N, Serban I, et al.The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems[C] // Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Prague, Czech Republic, 2015: 285 - 294.
[8] ZHANG S, DINAN E, URBANEK J, et al. Personalizing Dialogue Agents: I have a dog, do you have pets too? [C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. ACL, 2018 : 2204-2213.
[9] WELLECK S, WESTON J, SZLAM A, et al. Dialogue natural language inference [C] // Proceedings of the57th Annual Meeting of the Association for Computational Linguistics.ACL, 2019 : 3731-3741.
[10] WU Y, WU W, XING C, et al. Sequential Match Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots [C] // Proceedings of the 55th Annual Mee-ting of the Association for Computational Linguistics. 2017 : 496-505.
[11] Zupanc K, Štrumbelj E. A Bayesian hierarchical latent trait model for estimating rater bias and reliability in large-scale performance assessment [J]. PLoS One. 2018;13 (4):e0195297. doi:10.1371/ journal. pone.0195297.
[12] Yu L, Zhang W, Wang J et al. Seqgan: sequence generative adversarial nets with policy gradient [C]. In: Proceedings of the AAAI conference on artificial intelligence, 2017.