Personalized Dialogue Generation Method of Chat Robot Based on Topic Perception

Junmei Li(✉)

School of Computer Engineering, Jingchu University of Technology, Jingmen 448000, China
chenweiliang7895@163.com

Abstract. Human-Computer interaction system is a significant research direction in the field of human-computer interaction, and the research of open domain chat robot has received extensive attention. There are many problems in the existing chat robot: lack of personalized features, resulting in the process of the same chat, and the conversation has nothing to do with the topic. Therefore, a method of creating personalized conversation based on topic perception is proposed, and a personalized conversation model based on topic perception is designed. Semantic analysis and text similarity calculation are needed to build a conversation model. Based on the dialogue model, the training robot collects the corpus data related to the subject, convolves the corpus data related to the subject, and carries out the topic perception training. Finally, a personalized dialogue mechanism is established to generate personalized dialogue. Through experimental comparison, it is proved that the dialogue generated by this method is more suitable for the chat topic.

Keywords: Chatbot · Personalization · Dialogue generation

1 Introduction

Man-machine dialogue system is a significant research direction in the field of human-machine interaction of chat robot, and various dialogue systems are developing vigorously. Text generation, also known as natural language generation, is a key technology to realize dialogue system. Various types of information, such as text and image, can be used to automatically generate smooth and clear natural language text. BENGIO et al. call neural network language model applied to the task of text generation [1], using neural network language modeling. In order to solve the long term dependency problem in natural language, MIKOLOV uses RNN to build language model, and puts forward RNNLM, which improves the accuracy of language model. Since then, RNN and its variants such as the long short termmemory (LSTM) have become the most commonly used method in natural language processing. However, the recently proposed Transformer model has successfully solved some problems in the RNN model, which has triggered a wave of research. Reference [2] method applies the LSTM algorithm to chatbots. The method extracts the fictional dialogue content in the chatbot film and television database. Taking into account the target program model factors, the fusion of LSTM and BiLSTM

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 549–560, 2022.
https://doi.org/10.1007/978-3-031-18123-8_43

models is used to provide accurate dialogue texts. This method improves the accuracy of the session. The reference [3] approach integrates the contextual bandit algorithm into MathBot to personalize the pacing of the conversation, allowing bandits to insert additional practice questions or skip explanations. Provides valuable experience for teaching course dialogue.

Although the conversation rationality of chatting robot is controlled, it depends on the natural sequence structure of RNN. Although the natural sequence structure of RNN is suitable for the task of natural language processing, its strict linear structure will lead to the problem of gradient disappearance or explosion during the training process, and it is difficult to carry out parallel training, which is a serious problem in large-scale application scenarios. To address these issues, Google introduced a new sequence modeling model, the Transformer Model, in 2017. As soon as this model was put forward, it aroused great repercussions in the field of NLP and abandoned the sequence structure in RNN. The whole model is made up of Attention module, which effectively solves the problems of long distance dependence and poor parallel computing abili. Transformer model can capture the semantic information of text sequence efficiently, and its semantic feature extraction ability, long distance feature capture ability and task comprehensive feature extraction ability are much better than RNN model. Recent popular large-scale pre-training models such as GPT model, BERT model of the basic structure of Transformer, in a variety of natural language processing tasks to create excellent results, its superior ability is obvious.

One of the longest research goals in the field of artificial intelligence is the social chat robot, which is a human-computer dialogue system capable of empathy with human beings. If the chat robot wants to establish emotional contact with the user, it must have several abilities, first of all, the ability to integrate context and context, in the process of chat.

At the same time chat robot must have a consistent personality, such as age, gender, etc., if these features change, it is easy to make users feel stripped. Finally, conversation content must be diverse, not always produce "I don't know", "yes, that's right" such generic replies, otherwise users are very easy to produce boredom. The design of personalized dialogue generation method for chat robot is particularly important [4]. In order to solve the problem of single robot dialogue, this paper constructs a topic-aware-based personalized dialogue content generation model. In the process of dialogue, contextual information and personalized feature information are considered, which can better perceive the dialogue topic and improve the accuracy of human-computer conversation responses. And adopt a variety of optimization methods to increase the diversity of the generated response content, so that the generated context is coherent and consistent with high-quality dialogue content.

2 Personalized Dialogue Model Based on Topic Awareness

2.1 Semantic Analysis

The chatting system of chatting robot is mainly composed of speech synthesis module and speech recognition module, which transforms text into speech and speech into text respectively. The natural language understanding module uses NLP technology to

express user's intention in the form of data, which is converted into specific semantic data and then handed over to the next dialog management module. The function of the dialogue management module is to coordinate the work of several modules and to maintain the current system state. The natural language generation module is the most important one [5], which is the focus of this paper. The natural language generation framework is shown in Fig. 1:

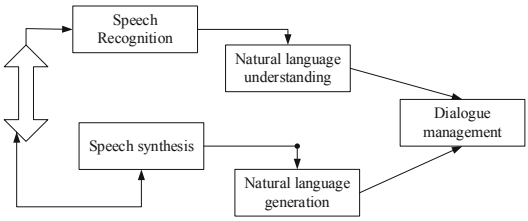

Fig. 1. Natural language generation framework

After generating the natural language, the chat robot should analyze and judge the aim of the syntactic structure analysis, judge whether a text accords with the grammar of the corresponding language, and then analyze the syntactic structure in line with the grammar norms. There are mainly three tasks, one is to perceive the topic, to judge whether the text belongs to a certain language category, the other is to disambiguate the word meaning, and the third is to analyze the sentence structure, context and syntactic relationship [6].

How to get a powerful parser, usually need to solve the following two problems, one is the formal expression of syntax, the other is the description of the entry information. Based on the semantic analysis, we can use the coding model and the Transformer model to construct a non-target-driven dialog system, namely chat robot. And realized with the user to carry on in the open domain the dialog. The Transformer model is essentially a codec architecture, and the overall structure of the model is shown in Fig. 2:

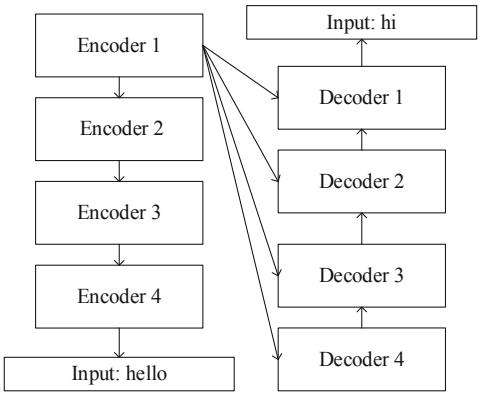

Fig. 2. Overall structure of model

The encoder and decoder model is composed of 4 encoders and 4 decoders. The structure of each encoder and decoder is shown in Fig. 2, respectively. During the encoding process, the data is first weighted by the eigenvector obtained from the Attention module, and then outputted by the feedforward neural network. The decoder has more code-decoding attention module than the encoder, which is used to get the relation between the current time output of the decoding stage and the input of the encoding stage.

2.2 Text Similarity Calculation

At present, most of the corpus-based short text similarity studies use the statistical description method based on context, because the context can provide sufficient semantic information for the definition of words. Lexical Vector Space Model (VSM) is a widely used statistics-based lexical similarity calculation strategy with relatively low algorithm complexity and easy implementation. The Lexical Vector Space Model (VSM) pre-selects a set of feature words [7] and then calculates the relevance of the set of feature words to each word (usually measured by the frequency of the words appearing in the context of the actual large corpus. So each word gets a word vector with the same dimension, and then the similarity between the words is calculated using a formula like this:

$$NDG(x, y) = \frac{\max(\log f(x), \log f(y)) - \log f(x, y)}{\log N - \min(\log f(x), \log f(y))} \tag{1}$$

Among them, *NDG* stands for standard Google distance, and the larger *NDG* stands for higher similarity, ranging from 0 to 1. $f(x)$ and $f(y)$ represent the number of pages containing the words x and y, $f(x, y)$ represents the number of pages containing the words x and y, and N represents the total number of pages referenced.

Distance is used to calculate the similarity of text sequences. It works as follows: there are currently two short text sequences A and B, in which B is the reference sequence [8]. There are three main steps: the short text sequence A deletes a word. Short text sequence A adds a word; short text sequence A replaces a word.

This repeats until Short Text Sequence A is converted to Short Text Sequence B, then the middle number of operations is recorded as $ED(A, B)$, and 2-$ED[i]\left[j\right]$ array is the minimum operand, indicating that the first $[i]$ words of Short Text Sequence A are converted to the first $\left[j\right]$ characters of Short Text Sequence B. The recursion formula for $ED[i]\left[j\right]$ is:

$$ED[i]\left[j\right] = \begin{cases} ED[i-1]\left[j-1\right] \\ \min\left(ED[i-1]\left[j-1\right], ED[i]\left[j-1\right], ED[i-1]\left[j\right]\right) + NDG(x, y) \end{cases} \tag{2}$$

3 Training Robot to Discourse Based on Dialogue Model

3.1 Collecting Corpus Data Related to Subjects

Personalized training of robot conversations requires a large amount of corpus data, so Scrapy is used to fetch the data first. Scrapy is a Web content crawling system developed

in Python that is fast, crawls content with less noise, and can be used to crawl a Web site and get data from those pages. When crawling data, due to crawlers and the site itself [9], a considerable amount of garbage data is collected, especially when crawlers crawl through URLs, some of which have no valuable content. Therefore, this article will remove garbage data in the following ways:

(1) Some URLs have no valuable content, based on filtering of worthless URLs, so regular expressions are used to match the URLs, thereby deleting the data.
(2) According to some sensitive words processing corpus data, some sensitive words are fetched in the data, and regular expressions are used to match such statements, thereby filtering the garbage data.
(3) Filter according to the length of the text sequence, some spam data length is very small, text less than 6 bytes will be filtered out.

As for word segmentation, we've already covered it, but we won't go further here. This paper uses the jieba word segmentation tool based on Python to process the corpus word segmentation. There are three patterns of Chinese word segmentation, among which the accurate pattern is the most accurate, mostly used in emotional analysis, syntactic analysis, etc. The full pattern is the fastest, but the accuracy is not high enough.

3.2 Convolution Processing of Corpus Data Related to Topic

On the basis of the corpus collected above, short text information related to the topic needs to be extracted. And construct short text topic graphs through word co-occurrences and document word relationships. In order to improve the accuracy of topic information extraction, it is necessary to use convolutional neural network to convolve the data, and make full use of document node and word node representation to improve text classification results. Convolutional Neural Network (CNN) is currently an advanced technology for subject-related corpus data processing. There is not only one network layer. There will be many different network layers appearing in turn, and the order is not fixed. There are Pooling Layer, Fully Connected Layer, Convolutional Layer, ReLU Layer, etc. The purpose of the convolution operation of the convolutional layer is to classify the picture by discovering the characteristics of a certain part of the picture. The maximum pooling operation is shown in Fig. 4 below, and the average pooling is shown in Fig. 3 below.

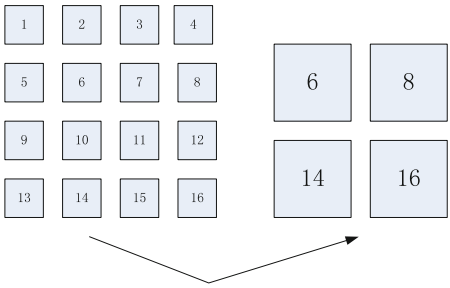

Fig. 3. Average pooling

The last layer in a convolutional neural network is generally a fully connected layer, and the fully connected layer can accept input from the remaining layers, which can be a pooling layer, a convolutional layer, and so on. The input of the fully connected layer is a multi-dimensional vector, and the dimension represents the categories included in the system. This output means that the system contains 8 categories, which are identified as the first. The probability of class I is 0.1, the probability of identifying the second class is 0.3, the probability of identifying the sixth class is 0.1, the probability of identifying the seventh class is 0.4, and the probability of identifying the eighth class is 0.1. On the basis of this convolutional processing, perception training is performed on topic data to improve the accuracy of the robot's response to questions.

3.3 Topic Awareness Training Based on Processed Data

The subject perception training can use the cyclic neural network training method. On the basis of the ordinary multi-layer BP neural network, the horizontal connection between the units of the hidden layer is increased. Through a weight matrix, the value of the neural unit in the previous time series can be transferred to the current neural unit [10]. As a result, the neural network has a memory function, and has good applicability for processing contextual NLP or time series machine learning problems. In the hierarchical expansion of Hidden Layer, $t-1$, t, and $t+1$ represent time series. X represents the input sample. S_t represents the memory of the sample at time t:

$$S_t = f(W * S_{t-1} + U * X_t) \tag{3}$$

where W represents the weight of the input, U represents the weight of the input sample at the moment, and V represents the weight of the output sample. At t = 1, the input $S_o = 0$ is generally initialized, and W,U,V are initialized randomly, where W,U,V are equal at each time (weight sharing). Carry out the following formula calculation:

$$h = U * X_1 + W * S_o \tag{4}$$

$$S = f(h) = f(U * X_1 + W * S_o) \tag{5}$$

$$O = g(V * S) \tag{6}$$

Among them, f and g are both activation functions. The state S at this time is used as the memory state of the previous time to participate in the calculation of the next time, and so on, as shown in the following formula.

$$h_t = U * X_t + W * S_{t-1} \tag{7}$$

$$S_t = f(h_t) = f(U * X_t + W * S_{t-1}) \tag{8}$$

$$O_t = g(V * S_t) \tag{9}$$

However, the training process will force the model to make non-zero or one predictions to distinguish between real data and generated content, reducing the generalization performance of the model. Topic perception solves this problem by acting like a regular term to reduce the model's confidence in its prediction results. Use a prior distribution that is not related to the current input parameters to smoothly predict the distribution function of the target, usually using a uniform prior distribution of all words. Label smoothing is equivalent to adding a divergence term on the basis of the negative log-likelihood function, that is, calculating the distance between the prior distribution and the predicted output probability of the model, that is, by preventing the model from over-concentrating the predicted value on the higher probability In terms of categories, it reduces the probability of general replies and increases the diversity of generated replies.

The cluster search algorithm is an algorithm commonly used in the decoding stage of the Seq2seq model. Its parameters are that the word with the highest probability is selected as the output at each moment in the decoding process. By selecting the word with the highest probability at each moment, the algorithm finally generates the sequence of sentences with the highest probability. And maximize it to generate more reasonable results and improve the quality of generated sentences. Since the probability of a sentence sequence is obtained by multiplying the probabilities of multiple words, the longer the generated sentence sequence, the smaller the probability value obtained by the multiplication, so the cluster search algorithm tends to generate shorter sentences. Google proposed a length penalty method to solve this problem. By reducing the probability value of short sequences and increasing the probability value of long sequences, the model has more opportunities to generate a longer sequence P, namely:

$$P = \frac{f(U * X_1 + W * S_o)}{(5 + g(V * S))} \tag{10}$$

Another problem in the cluster search algorithm is that the generated sentences have little difference and low diversity. By grouping the generated results, a similarity penalty is added between the groups to reduce the similarity of multiple results, forcing the model to generate more diversified content, and reducing the appearance of general responses.

4 Establishing Personalized Dialogue Mechanism of Robot

The robot's personalized dialogue mechanism is to imitate the process of humans observing a certain thing. When humans observe a thing, they must only pay attention to a part of the thing, and their attention moves with the movement of the focus. In other words, when human beings observe a thing, the attention given to various parts of the thing is inconsistent. It must be that one part gets more and the rest gets less. Therefore, the robot's personalized dialogue mechanism is very suitable for natural language processing.

The Encoder and Decoder in the traditional model exchange data through an intermediate semantic vector, and the length of this semantic vector is fixed, which will bring the problem of long-distance dependence to the model, that is, for long sequences of text. As the input progresses, the information in the latter part of the sequence will overwrite the information in the former part of the sequence. Therefore, Attention, by

retaining part of the output results of the Encoder on the input sequence, the training model selectively learns these output results, and will eventually be associated with the corresponding output sequence. In other words, the output and input will be selectively associated together.

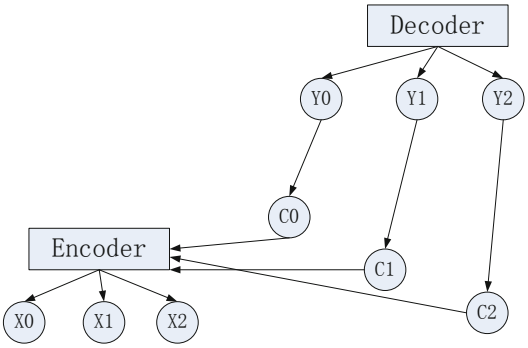

Fig. 4. Mechanism abstract

In the traditional robot personalized dialogue mechanism, the output of the decoder is directly based on the output with the highest probability, but sometimes the output with the highest probability is often the most common words in the corpus, usually "I don't know", "Hello", "Haha" and other meaningless replies, we call them safe replies. Therefore, how to avoid this kind of safe response is also a key issue of the seq2seq model. In order to solve this problem, the Beam Search algorithm can be used. Beam Search (cluster search) reduces the search space and improves time efficiency through "pruning" and multi-layer search. The so-called "pruning" is to filter out some nodes with a small probability every time you search downwards, and will not use this node as the parent node to continue the downward search. This reduces the number of searches and improves space utilization and time efficiency. Suppose that machine translation is used as an example to illustrate. The task is to translate "I am Chinese" in Chinese and "I am Chinese" in English. Assuming that the vocabulary is only three words in size, it is "I", "am", and "Chinese". So if the beam size is 2. In the decoder process, after having the beam search method, in the first output, we select the two words "I" and "am" with the highest probability, instead of selecting only the word with the highest probability.

On the basis of the above selection mechanism and the self-attention mechanism, the multi-head attention mechanism is proposed in the model. Each self-attention module is called a attention head. In the calculation of multi-attention mechanism, the input sequence is first computed through h different attention heads, then the h different feature matrices are assembled into a feature matrix by column, and then compressed into a matrix with the same dimension as a single attention head through a fully connected layer to obtain the output results of multi-attention modules. The general conversation model can carry out several rounds of conversations without considering personalized information. Chatbots without personalized features may have semantic inconsistencies in their conversations, such as being a student and working at work. So that its users have a sense of stripping, easy to make users aware of their own with a fake robot chat, it

is difficult to establish long-term emotional contact. In order to solve this problem, this paper designs a personalized dialogue model, which can generate appropriate replies according to its own characteristics and improve users' interactive chat experience.

This model is based on the codec framework in Fig. 2 above. The encoder and decoder are composed of encoder and decoder of Transformer model respectively, and the number of layers is the same as the general model. In contrast, personalized dialogue model needs to consider personalized information, and add personalized information in the input part of the model to encode. In this paper, personalized features refer to a set of sentence information that describes the character, and the personalized feature vector after coding together with the current input and the historical dialogue content guides the reply generation process in decoding stage. In the decoding process, the code-decoding attention module can determine the impact of user input, historical conversations and personalized information on the output of the current moment. By integrating contextual information and personalized information, personalized dialogue model can generate replies that are consistent with context and accord with specific personalized features.

5 Test Experiment

5.1 Set Up an Experimental Environment

This article uses Google's open source deep learning framework, TensorFlow. TensorFlow is an open source software library for high performance numerical computing. TensorFlow has a flexible architecture that allows users to easily deploy computing tasks to multiple platforms, devices, and even mobile devices. Tensorflow is a processing framework based on data flow graph. The nodes in the graph represent numerical computation, and the edges represent the data interaction between computing nodes. The hardware and software environment for the test is shown in Table 1.

Table 1. Experimental environment

Serial number	Software/hardware	Parameter
1	Operating system	Linux
2	Framework platform	TensorF1ow
3	Development language	python
	Hardware	Parameter
1	Graphics card	GTX-1060
2	CPU	Xeon-E5
3	RAM	8G

From the perspective of data transmission and processing, the data flow diagram graphically expresses the logical function of the system, the logical flow of data within the system and the logical transformation process. It is the main expression tool for structured

system analysis methods and a graphical method for representing software models. Data flow diagram can be more intuitive representations of TensorFlow. Data flow graphs represent mathematical computations through directed graphs of "nodes" and "lines". The "nodes" in the diagram are used to represent the mathematical calculations being performed. They can also represent the start or end of the data input or output, or the end of reading in and writing out persistent variables. Lines represent data interactions between Nodes. A "line" can transmit a multidimensional array of data, known as a "tensor," that is "resizable." The tensor flows through the graph, which is why this tool is called "Tensorflow." As long as all the tensors at the input are ready, the "nodes" are allocated to each computing device to perform asynchronous parallel operations.

5.2 Experimental Data

In order to realize the model of chat robot, we need a lot of Chinese conversation corpus. So the most appropriate one is the dialogue in movies and TV plays. In contrast, there are more dialogues in foreign films and TV series, so this paper selects the open subtitles corpus, because this article is to verify the Chinese chat robot, but the open subtitles is English dialogue, so the use of translation tools to translate into Chinese. The corpus size is shown in Table 2.

Table 2. Corpus size

Serial number	Corpus	Open subtitles corpus
1	Development set	2000
2	Test set	2000
3	Training set	44063050
4	Data set	44067050

Using the model to construct the chat robot model, the model is simply a translation model, translating one language sequence into another language sequence. The whole process is to map one sequence as input to another output sequence 62 by using long and short memory network or recursive neural network.

5.3 Experimental Result

In order to judge the practicality of the method, this paper compares the method based on LSTM with the method designed in this paper. It should be noted that the hidden layer of the model has 512 nodes, the word vector dimension is 64, and both Encoder and Decoder use the LSTM model, the size of the batch is 128. Table 3 shows some experimental results.

Table 3. Experimental result

Number of experiments	Question	Answer (LSTM)	Answer (context bandit algorithm)	Answers (in this article)
1	What do you like to eat?	Thank you!	I do not know	There are many favorite foods
2	What do you like to eat?	Sorry!	Thank you!	I don't know
3	What do you like to eat?	Thank you!	There is no favorite food	There is no favorite food
4	What do you like to eat?	Thank you!	There is no favorite food	What kind of food do you like?
5	What do you like to eat?	I have no idea	I do not know	No favorite food
6	What do you like to eat?	I do not understand you	Glad you like it, too	The food is very nutritious

As can be seen from Table 3, compared with the traditional method based on LSTM and context bandit algorithm, the dialogue answer of the robot based on topic perception is more personalized and diversified, but the traditional method has a single answer. And there are many repeated sentences in 6 experiments. And based on the "What do you like to eat?" Obviously, the method designed in this paper to generate personalized conversations between chat robots is more in line with the topic of chat - favorite food.

In order to compare the number of dialogue rounds of the dialogue generation model, we use the test set to simulate the dialogue process, and count the number of dialogue rounds before the answer with no clear meaning like "I don't know". The results are shown in Table 4:

Table 4. Comparison of different model dialogue turns

Method name	Number of dialogue rounds
LSTM	3.7
Context bandit algorithm	4.2
The method of this paper	4.6

It can be seen from Table 4 that this method has more dialogue rounds than the other two methods, probably because the context management of this method records the historical dialogue information, which can be answered according to the context topic. This method improves the diversity of dialogues, reduces the probability of meaningless answers, and increases the number of dialogue rounds to some extent.

6 Conclusion

The chatbot dialogue model proposed in this paper improves the problems of traditional models to a certain extent. This paper innovatively applies a topic-aware approach to generate personalized bot chat conversations. On the basis of semantic analysis, a human-machine dialogue model is constructed, and then the text similarity of chat machine short texts is calculated. Based on the dialogue model, the robot is trained on the dialogue data, the corpus data related to the subject is collected, and the convolution training is performed on the corpus data related to the topic. Finally, the robot personalized dialogue is generated through the robot personalized dialogue mechanism.

But the current dialogue model of chat robot still faces many problems. At present, the conversation model based on topic perception needs a large number of standard Chinese pairs. The more the number of question and answer pairs, the less the noise in the data, the better the model will be in theory. However, there are few open source corpus for Chinese dialogues, so how to collect large scale standardized Chinese corpus is an urgent problem to be solved. From the point of view of chat robot on the market at present, the development of chat robot is still in its infancy, and breakthrough is needed in technology and Chinese corpus. So I hope that in the near future, to find a better technology to achieve a breakthrough and development of chat robot.

References

1. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-Learning and e-Education. Mob. Netw. Appl. **26**(5), 2123–2126 (2021). https://doi.org/10.1007/s11036-021-01777-7
2. Anki, P., Bustamam, A.: Measuring the accuracy of LSTM and BiLSTM models in the application of artificial intelligence by applying chatbot programme. Indones. J. Electr. Eng. Computer Sci. **23**(1), 197–205 (2021)
3. Cai, W., et al.: Bandit algorithms to personalize educational chatbots. Mach. Learn. **110**(9), 2389–2418 (2021). https://doi.org/10.1007/s10994-021-05983-y
4. Zhao, L.N., Li, W., Kang, B., Zhang, K.: Python-based intelligent robot multi-channel knowledge base push simulation. Comput. Simul. **37**(3), 328–332 (2022)
5. Chen, W., Chen, X., Sun, X.: Emotional dialog generation via multiple classifiers based on a generative adversarial network. Virtual Real. Intell. Hardw. **3**(1), 18–32 (2021)
6. Yang, M., Huang, W., Tu, W., Qu, Q., Lei, K.: Multitask learning and reinforcement learning for personalized dialog generation: an empirical study. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 49–62 (2020)
7. Lee, C.H.: Dual policy network for speaker-specific dialog generation in deep reinforcement learning. J. Inst. Electron. Inf. Eng. **56**(4), 44–49 (2019)
8. Stein, B.: Zum titelbild: intergenerationeller dialog gezeichnet von leonard von der stein. Psychotherapie im Alter **16**(2), 217–219 (2019)
9. Liu, S., Wang, S., Liu, X., Lin, C.T., Lv, Z.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2020)
10. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mob. Netw. Appl. **26**, 1891–1903 (2021). https://doi.org/10.1007/s11036-021-01777-7