Journal of Physics: Conference
Series

PAPER • OPEN ACCESS

Research on the Design of Intelligent Chatbot Based on Deep Learning

To cite this article: Chuwen Kuang 2022 *J. Phys.: Conf. Ser.* **2170** 012017

Research on the Design of Intelligent Chatbot Based on Deep Learning

Chuwen Kuang*

Huizhou Economics and Polytechnic College, Huizhou, 516001, China

*Corresponding author's e-mail: kuang@hzcollege.com

Abstract. At present, chat robots still have the problems of insufficient utilization of multi-modal information, weak emotion expression ability, and lack of model fusion mechanism. With the application of deep learning in the fields of natural language understanding, word vector representation, machine translation, sentiment analysis, and Chinese word segmentation, people have begun to study the key technologies of chat robots and apply deep learning to chat robots. In view of the problem that the traditional chat robot Seq2Seq model training will produce too safe, common, and repetitive answers to the speech corpus, which affects the interactive experience with users, this paper combines the idea of mutual information to improve the objective function of the model so that the model can produce richer and more diverse answers and give users a better experience. Experiments show that the model has improved many indicators compared with the single-modal model in the microblog data set experiment.

1. Introduction

In recent years, with the explosive growth of Internet data scale and the advancement of artificial intelligence technology, chatbots, as a product with good human-computer interaction, have gradually become the research focus of academia. Chatbot applications are also called dialogue systems, dialogue agents, etc., which are computer programs that can interact with users in a complete way using natural language based on user input [1]. Compared with traditional search engines, chat bots can also extract the information users need from a large amount of information, but they emphasize the stickiness of interaction with users. That is, the user does not want to leave as soon as possible, so it has a better interactive effect. The accumulation of massive data and the improvement of computing power have made deep learning models, it has unprecedented breakthroughs in the field of natural language processing [2]. As a subdivision of natural language processing, deep learning technology can be realized whether it is based on retrieval methods or generation methods. The interactive effects of chat robots have improved. This paper attempts to apply deep learning technology to the realization of chat robots, in order to improve the ability of using multi-modal information and emotional expression of chat robots to promote the development of intelligent robots.

2. Deep learning model and chatbot model

2.1. Deep learning model

Deep learning is a method of automatically learning features. In essence, it is a machine learning model that has built a large amount of training data and a large number of hidden layers to learn more useful features to achieve the purpose of improving classification or prediction accuracy. To achieve the purpose of feature learning. Different from traditional shallow learning, deep learning: 1) The depth of the model structure, usually with 5 layers, 6 layers, or even 10 layers of hidden nodes; 2) The importance of feature learning is clearly emphasized, that is, through the layer-by-layer feature transformation transforms the feature representation in the original space to the new feature space, making model classification and prediction easier [3]. There are two ways to learn to construct features: based on manual rules and using big data, the latter can better reflect the information inherent in the data. It is an end-to-end structure that can go from data training directly to the final result, without additional feature extraction [4].

The commonly used models of deep learning mainly include: sparse coding, autoencoder, conviction network, restricted Boltzmann machine, and convolutional neural network. In deep learning networks, there are generally four training methods: guided learning (supervised learning), unsupervised learning (unsupervised learning), semi-supervised learning and reinforcement learning [5]. Among them, autoencoders, restricted Boltzmann machines, sparse coding, and deep belief networks all belong to unsupervised learning, and convolutional neural networks belong to supervised learning.

The purpose of unsupervised learning is to discover the special structure of a given data. It is necessary to clarify the inherent properties and laws of the training data through the learning of unlabeled training samples, so as to provide a basis for further data analysis [6]. The framework of unsupervised deep learning is generally shown in Figure 1.

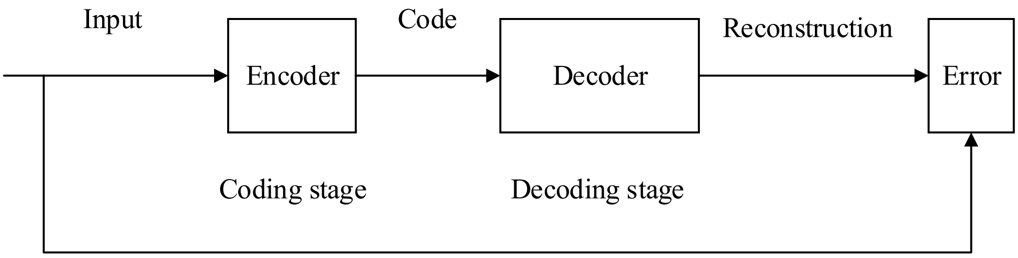

Figure 1. Unsupervised deep learning framework

Semi-supervised learning is a hybrid deep learning network that combines supervised learning and unsupervised learning. Generally, supervised learning is used to inventory models, unsupervised learning is used to generate features, and a large amount of unlabeled data and a small amount of labeled data are combined in the training phase. The framework of semi-supervised learning is shown in Figure 2.

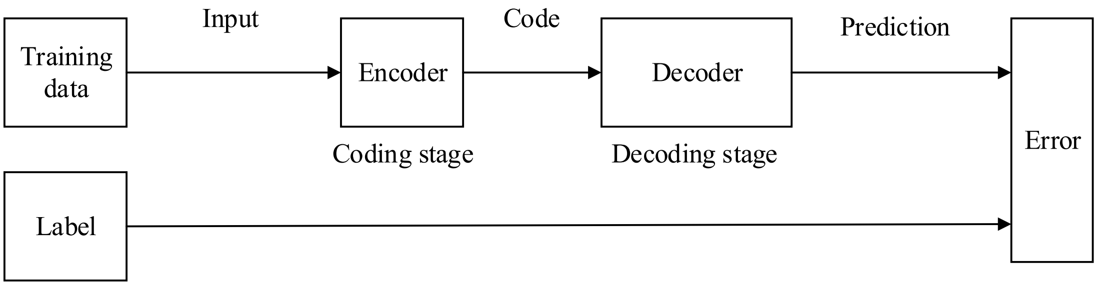

Figure 2. Semi-supervised learning framework

2.2. Chatbot model

According to the mainstream technology, chat bots can be divided into chat bots based on artificial templates, chat bots based on retrieval, chat bots based on machine translation, and chat bots based on deep learning.

The chat robot technology based on manual templates generally sets chat backgrounds artificially, and writes targeted chat templates for each background. The chat templates include questions that users may ask and answers to the questions. The advantage of artificial template-based chatbot technology is that it can accurately respond to the question model raised by the user, but its disadvantage is that this technical route requires a lot of manual work, and the model has poor scalability, and requires a background artificial expansion.

The chat robot technology based on machine translation applies the technology in the field of machine translation to the development process of chat robots. The core idea of this technology is to regard the process of a chat robot generating reply sentences based on user input sentences as a process of translating sentences in one language into sentences in another language [7].

Most of the chatbot technologies based on deep learning are built under the deep learning model framework Encoder-Decoder. It is worth mentioning that in the field of machine translation, if deep learning is involved, it is mostly built based on the Encoder-Decoder model. This paper uses deep learning technology to develop a chatbot model.

The model structure of a generative chatbot based on deep learning is shown in Figure 3. The text entered by the user is converted into a language understandable by the machine through the text preprocessing module. The text preprocessing module includes the following tasks: obtaining text, data cleaning, word segmentation, and text representation. The quality of the text directly affects the final performance of the system [8]. The natural language generation module is the most important part of the generative chatbot system and is also the focus of this paper. The natural language generation module learns how humans conduct conversations through complex deep learning models. The trained model will generate interesting responses based on user input.

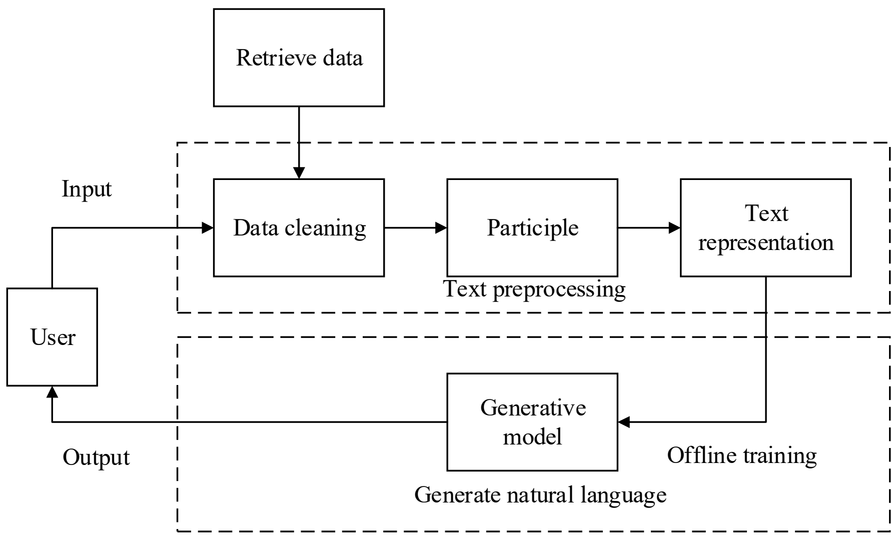

Figure 3. Chatbot system structure

3. Multi-modal dialogue generation model based on deep learning

3.1. Seq2Seq model

In the field of natural language processing, the Seq2Seq model is mainly for time-series data. After deep neural network learning, it is calculated from a variable-length input sequence and outputs another sequence of varying lengths. The Seq2Seq model is one of the most critical technologies in the field of artificial intelligence. Figure 4 shows a simple Seq2Seq model, which contains three parts, namely the encoder, the decoder, and the intermediate semantic vector C used to connect the encoder and the decoder [9]. Among them, the encoder is used to learn the training input sequence, and encode the sequence into a fixed-length intermediate semantic state vector C, and the decoder will output another sequence after learning and training the intermediate semantic state vector C.

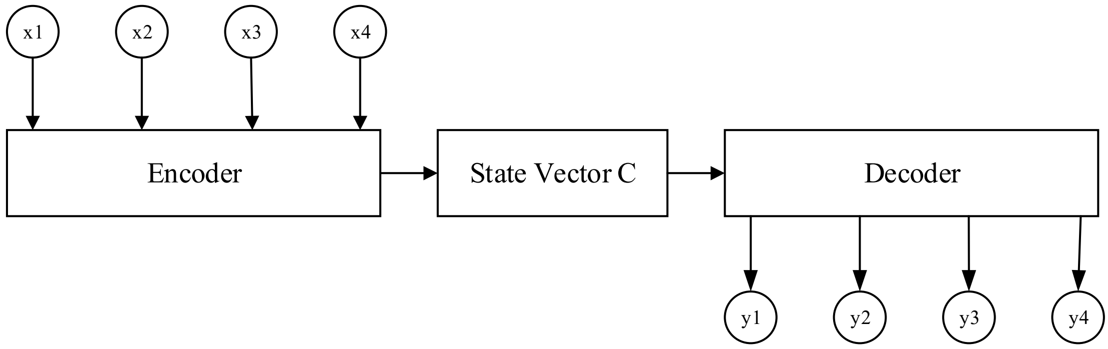

Figure 4. Simple Seq2Seq model

After the conversion from the encoder to the intermediate semantic vector C, the decoder obtains the relevant information from the semantic vector, and then decodes the vector through the label to obtain the appropriate sentence through the algorithm, which simplifies the data set to a certain extent. The data set is converted into a fixed-length vector representation by the encoder, and the vector is introduced into the network as a data set for training. The emergence of the Seq2Seq model based on deep learning has improved the efficiency and accuracy of natural language processing tasks.

3.2. A two-way GRU+Attention chat robot model based on mutual information improvement

This paper uses the two-way GRU + Attention model as the chat robot model, and the network structure. On the basis of the two-way GRU model, the Attention mechanism is added, and the Attention layer is used to calculate the weight of each time sequence, and then the vector weights of all time sequences are summed and output.

Mutual information is a concept of information theory. It is used to describe whether two random events are related or not. It is a measure of the interdependence of two events, that is, under the condition of understanding one of the information Y, it is uncertain to eliminate the other message X. The amount of information provided by sex, that is, the amount of information about X contained in the value of Y.

After experiencing the two-way GRU + Attention training and testing to obtain the model, we found that although the model has matured to generate reply language and achieve normal chat effects, it still has obvious flaws. For example, we are having a series of dialogues. Later, the chatbot will always generate safe and grammatical replies like "Me too", "I love you", "I don’t know", "We can’t be here anymore", etc. These replies are sometimes not only repeated, but also affect for the quality of chat, we should try to avoid such situations. Analyzing the reasons, although the generation of this type of response will be affected by the quality of the corpus content, it is mainly caused by the lack of a better objective function for the training model. The traditional Seq2Seq model uses the log-likelihood function as the objective function for training, and is more inclined to generate safe and ordinary responses. Therefore, the higher the frequency of the sequence in the training corpus, the maximum probability of the final generation, and its optimized objective function It is the maximum likelihood method (MLE) to model, that is, given a user input sequence, through training to maximize the probability of generating a response output:

$$\hat{O} = \arg max_{O}\{logp(O|I)\} \tag{1}$$

Among them, / is the input sequence, and 0 is the output sequence. The traditional model pays more attention to the impact of input on output, but ignores the impact of output on input sequence. It was originally used to process a single round of dialogue, and is most often used for tasks such as machine translation. Therefore, the response results lack diversity and range of choices. It can be seen that the loss of the value loss during the training process will cause high-frequency replies in the training set, and the probability of these high-frequency replies appearing in the test is very high. In this paper, in the text generation Seq2Seq model in a specific field, the traditional Seq2Seq model is easy to be more sensitive to high-frequency data, and it is easy to generate short-length, repeated responses, and meaningless answers. The method based on entropy as the loss value is modified. Advance, combined with the idea of maximizing mutual information, adding penalty to mutual information instead of the objective function, adding consideration to the output sequence, as shown in equation (2).

$$log\frac{p(I,O)}{p(I)p(O)} = log\frac{p(O|I)}{p(O)} = logp(O|I) - logp(O) \tag{2}$$

Equation (2) adds logp(O) as a penalty item more than the maximum likelihood probability, and adds the λ parameter at the same time:

$$\hat{O} = \arg max_o\{(1-\lambda)logp(O|I) - \lambda logp(O)\} \tag{3}$$

The objective of maximizing mutual information optimization is not only to maximize the probability of the model from the input sequence/generating response 0, but also from the opposite direction, that is, to maximize the probability of the response 0 generating the input sequence. Through the X parameter adjustment, control which of the two sequences is more It is important to consider optimization in the opposite direction, which can reduce the generation probability of those common responses. Since the appearance of each word in the sequence is related to the previous W-1 words, the final probability is the product of the probability of each word:

$$p(O) = \prod_{m=1}^{N_o} p(O_m|O_1, O_2, \cdots, O_{m-1}) \tag{4}$$

Because in the Seq2Seq model, the last output word largely determines the output of the next word, and the top words in the 0 sequence have a great influence on the entire sequence, and the influence gradually changes as the sequence progresses. Generally speaking, for generating a long sentence, the latter part of the sequence is more prone to grammatical errors, so choose to set the coefficient to the standard normal distribution function.

4. Test analysis

Due to the openness of the generative chatbot model, there is no unified standard to evaluate the pros and cons of its structure and the quality of the training effect. It is similar to the common metrics used to evaluate machine translation such as BLEU, ROUGE, METEOR, and task completion of task-based robots. Evaluation indicators such as degree are not suitable for this model. Our goal is to enable the chatbot model to generate more diverse responses under the condition of generating a grammatical sequence. Therefore, this section judges whether the response generated by the model is correct, whether it conforms to the grammar, and whether the diversity is guaranteed from the training and effect of the model.

The improved two-way GRU network + Attentin model based on mutual information obtains the final chat robot model after 40 rounds of learning. The improved model increases the calculation of the objective function, so the training time is significantly increased. Each round takes about 15 minutes and 30 seconds. The loss value during training varies with the round as shown in Figure 5. Experiments show that the improved model can produce more interesting answers while ensuring that the output conforms to the grammatical sequence. Such a model also increases the user's interactive experience.

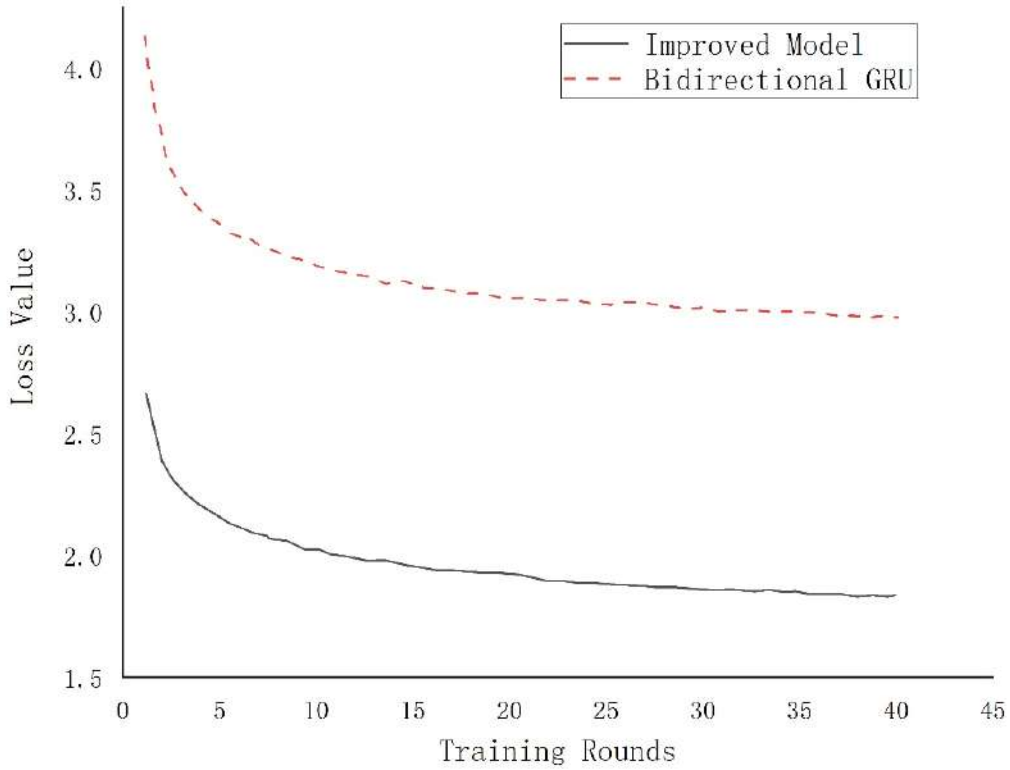

Figure 5. The improved model and the two-way GRU model first 40 rounds of training loss value

5. Conclusion

The research of chatbot dialogue based on seq2seq model is the main research content of this paper. This paper proposes an improved two-way GRU + Attention model based on the idea of mutual information, and examines the quality of the model from the final response effect. After experiments, it is found that the improved Chinese chat model can generate grammatical responses while adding a lot of interaction. Sexual answers increase the variety of responses.

References

[1] Jia X B, Li R H, Chang J C, et al. A Survey of Intelligent Dialogue System Research. Journal of Beijing University of Technology, 2017,43(09):1344-1356.
[2] Yuan R X. Design of chat robot based on intelligent terminal. Digital Technology and Application, 2019,37(10):144-145.
[3] Yu Z F. Robot application research based on deep learning. Digital Technology and Application,2019,37(05):53-54+58.
[4] Li J K. Intelligent chat robot for students. Electronic Production,2019(16):37-40.
[5] Wang S L, Li D P, Zhao B X, et al. Analysis of chat robot technology. Journal of Wuhan University (Information Science Edition),2021,46(02):296-302.
[6] Zhao H Y. Research and implementation of intelligent chat robot based on deep learning. Intelligent Computers and Applications,2019,9(06):308-311.
[7] Zhang S L, Hai L, Wang Y Y. Research on chat robots based on big data analysis. Digital Communication World,2020(01):138+147.
[8] Liu J C, Wang W J, Sun X Y. A campus intelligent chat robot based on Java technology——Xiaotong. Software,2020,41(03):69-73.
[9] Zhang H N, Ma J L. Research on Psychological Counseling Dialogue System Based on seq2seq Model. Information Recording Materials,2021,22(03):207-208.