Hindawi
Scientific Programming
Volume 2021, Article ID 8900304, 9 pages
https://doi.org/10.1155/2021/8900304

Research Article

Research on Spoken Language Understanding Based on Deep Learning

Hui Yanli

Faculty of Foreign Languages and Business, Jiaozuo Normal College, Jiaozuo 454001, China

Correspondence should be addressed to Hui Yanli; 1295004004@jzsz.edu.cn

Received 9 August 2021; Revised 30 August 2021; Accepted 24 September 2021; Published 27 October 2021

Academic Editor: Bai Yuan Ding

Aiming at solving the problem that the recognition effect of rare slot values in spoken language is poor, which affects the accuracy of oral understanding task, a spoken language understanding method is designed based on deep learning. The local features of semantic text are extracted and classified to make the classification results match the dialogue task. An intention recognition algorithm is designed for the classification results. Each datum has a corresponding intention label to complete the task of semantic slot filling. The attention mechanism is applied to the recognition of rare slot value information, the weight of hidden state and corresponding slot characteristics are obtained, and the updated slot value is used to represent the tracking state. An auxiliary gate unit is constructed between the upper and lower slots of historical dialogue, and the word vector is trained based on deep learning to complete the task of spoken language understanding. The simulation results show that the proposed method can realize multiple rounds of man-machine spoken language. Compared with the spoken language understanding methods based on cyclic network, context information, and label decomposition, it has higher accuracy and F1 value and has higher practical application value.

1. Introduction

In our society, with the development of science and informatization, more tasks have been applied to the field of artificial intelligence, and it has become an irreversible trend. With the rapid development of artificial intelligence, there are countless tasks to process serialized data in our society. For example, speech recognition, natural language understanding, and time series data all need to process serialized data. The spoken language system integrating speech recognition and speech synthesis is the core technology of human-computer interaction, and oral understanding is the core of spoken language system [1]. Therefore, the research on oral comprehension can enable people to apply it more accurately to spoken language system, which is more convenient for people's life and work. It has always been explored and studied to enable machines to communicate with people without barriers. Oral English is used in human communication. If the computer can understand spoken language and make correct answers and can correctly complete various operation tasks required by people according to people's instructions, it can use robots to complete operations in many occasions, especially in some dangerous fields, which can save a lot of resources and reduce the risk. It can be seen that spoken language system is widely used in life and is of great significance. Spoken language system is such that people express their ideas in natural language to communicate with a certain field of computer [2]. This way can make the computer understand human requirements more efficiently and quickly, so as to complete people's demands according to the corresponding model processing. Oral understanding is to convert the natural speech input by people into text through speech recognition, convert the text into corresponding word vector or sentence level vector in the oral understanding system, then send the vector through the encoder or directly into the built model, and finally decode or directly output the sentence. For the whole spoken language system, the key part is oral understanding and dialogue management. If oral comprehension cannot be performed correctly or the performance of oral comprehension is poor, such a situation will lead to the error of query results of subsequent dialogue management, resulting in the poor performance of the whole spoken language system, which cannot complete various operations and tasks required by people [3]. It can be seen that oral understanding plays a key role in spoken language system. The performance of oral comprehension directly determines the performance of spoken language system. Therefore, the study of oral comprehension is of great value and significance.

The academic community has carried out extensive research on spoken language understanding. Zhang and others improved the effectiveness of information feature extraction and oral comprehension performance by adding and storing historical state information [4]. Yang and others constructed the initial representation of the current round of text and context text in combination with the phonetic features and used the context semantic information to assist the intention detection of the current round of text, which improved the detection effect [5]. Xu and Huang transformed label classification into independent classification and introduced external word vector to improve the classification performance of the model [6]. The above research results have improved the performance of spoken language understanding methods, but they still have the problem of poor slot value recognition, which affects the accuracy of dialogue tasks. With the successful application of deep learning, deep neural network has made remarkable achievements [7]. The development of deep neural network has gone through a very long time. Now remarkable achievements have been made in the fields of speech recognition, image processing, text processing, computer vision, and natural language processing. Deep neural network is a successful application in many fields such as natural language processing, compared with the traditional oral comprehension model. The biggest feature of deep neural network is to train a large amount of data and then extract the characteristic information. The characteristic information obtained through the network structure can achieve good results in oral comprehension tasks. Therefore, based on deep learning, this paper proposes a spoken language understanding method to improve the effect of oral understanding tasks, promote the development of human-computer interaction technology, and better meet the needs of practical application scenarios.

2. Spoken Language Understanding Method Based on Deep Learning

2.1. Semantic Text Classification. In man-machine dialogue system, oral comprehension is mainly used to understand what users say and extract important information. Oral comprehension model includes three tasks: domain recognition, intention recognition, and semantic slot filling. Domain recognition task and intention recognition task are used to understand what users say. Both tasks belong to text classification. Semantic slot filling task is to extract important information, which belongs to sequence annotation problem. Domain recognition and intention recognition are to classify the dialogue text entered by the user, and the classified label is the domain involved in the dialogue text or the user's intention. Therefore, this paper first designs a semantic text classification model, uses the algorithm to extract the local features of the text, and matches it with the current task to better classify the text. The model consists of text convolution neural network and bidirectional long-term and short-term memory neural network combined with attention mechanism. The multiconvolution kernel mechanism of text convolution neural network model can better extract the n-gram features of text data, so that the text convolution neural network model can learn more abundant local features [8].

The bidirectional long-term and short-term memory neural network model can effectively extract the context information of semantic text and concentrate the more important information of the current task [9]. There are 160 neurons in the hidden layer of bidirectional long-term and short-term memory neural network model, the size of four-layer convolution nuclei is 2–5, the number of convolution nuclei is 32, and the size of attention mechanism is 128. The two-channel neural network model is used to extract text information at the same time, which makes the model more efficient. The first channel is the word vector of the word. First, the text is segmented, converted into word vectors, and input into the first channel. Unsupervised stacked deconvolution neural network is used to learn from the word vector of the text to obtain the feature mapping matrix. The feature mapping matrix is used as the convolution kernel of the deep convolution neural network to convolute and pool the word vector layer by layer [10]. Then, the important information of text context is extracted through the two-way long-term and short-term memory model. The second channel is the input word vector. Convert a single Chinese character in the text into a word vector. The local semantic features of the text are extracted through the convolution model of multiconvolution kernel. The hierarchical attention mechanism is used to select the important sentences in the text at the network layer, and the network is extracted layer by layer to obtain the text feature vector [11]. Finally, the left and right channel outputs are spliced into vectors, and then text classification is carried out through the maximum layer of fully connected neural network. In text classification, each output is related to the context of the input and context at that time [12, 13]. If the output vector is added each time and the average value is taken directly, the contribution of each output to text classification is the same, but this is not the case. Keywords in the classification should have greater weight. Therefore, when outputting vectors, we want to focus on vectors that are more important to the current task, so we introduce the attention mechanism. Note that the mechanism model can be expressed as

$$\alpha = \tanh(w\beta + \varepsilon). \tag{1}$$

In formula (1), α represents the output value of the attention mechanism; tanh represents hyperbolic tangent function; w and ε represent the weight and bias of attention mechanism; β represents the splicing of outputs in both positive and negative directions at each time. The outputs of forward and reverse are spliced to obtain the randomly initialized attention mechanism column vector. Finally, a normalized attention mechanism weight is obtained by softmax operation. The calculated output weights at each time are weighted and summed to obtain the output of the model; that is, the text classification is completed. Figure 1 shows the semantic text classification model.

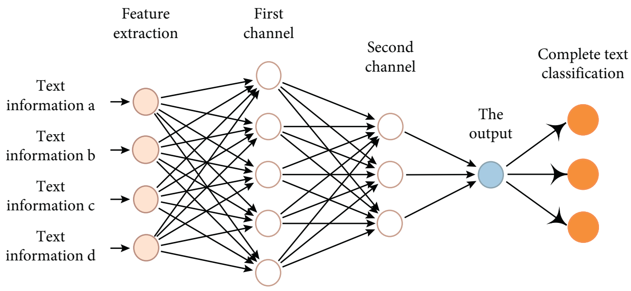


Figure 1: Semantic text classification model.

2.2. Design Intent Recognition Algorithm. In oral comprehension task, intention recognition task is a classification problem, which is used to extract users' specific intention [14]. Semantic slot filling task is a sequence labeling problem, that is, each word in a given sentence is labeled respectively. There is a certain relationship between intention recognition task and semantic slot filling task. Therefore, these two tasks can be completed in the same model [15]. In this paper, Albert pretraining model and convolutional neural network are used to complete the task of intention recognition, and Albert pretraining model and conditional random field are used to complete the task of semantic slot filling. Albert model backbone network adopts transformer encoder framework and GELU nonlinear activation function. The dimension of the input embedded vector dimension is smaller than the output vector. The word level embedding vector has no context dependence, but the output of the hidden layer includes not only the meaning of the word itself, but also some context information. Therefore, the expression of the hidden layer contains more information [16]. In natural language processing tasks, the word embedding matrix is usually large. Due to the large number of parameters and the process of back propagation, the updated content is also sparse. Combined with the above two points, Albert model adopts a factorization method to reduce the amount of parameters. Firstly, the unique heat code is mapped to a low dimensional space and then mapped to a high-dimensional space. Through this decomposition method, the amount of word embedding parameters can be reduced [17]. On this basis, the parameters of the full-connection layer and the attention mechanism layer are shared; that is, all the parameters inside the encoder are shared. In order to retain only consistency tasks and remove the impact of subject prediction, the positive samples of SOP are obtained in the same way, and the negative samples reverse the order of positive samples. That is, SOP only focuses on the order of sentences and has no influence on the subject [18]. Albert model input needs to add [CLS] at the beginning of the text, and the output corresponds to the input [CLS] vector containing the information coding of the whole sentence, which can be used for text classification tasks [19]. The remaining eigenvectors are used for the sequence annotation task. Use [CLS] vector and feature vector to identify semantic intention. Conditional random field is a probability graph model and belongs to discriminant model. In the field of natural language understanding, linear chain conditional random fields are often used to solve the problem of sequence annotation [20]. In the linear chain random field, the sequence and label meet the following conditions:

$$\theta\left(B_i \mid A, B_1, B_2, \cdots, B_n\right) = \theta\left(B_i \mid A, B_{i-1}, B_{i+1}\right). \qquad (2)$$

In formula (2), θ represents probability; A represents sequence; B_n represents the tag sequence; n indicates the number of labels; i indicates the marked serial number. The parametric form of linear chain conditional random field is as follows:

$$\theta\left(B_i \mid A\right) = \frac{1}{\delta} e^{\left(\lambda_1+\lambda_2\right)}. \qquad (3)$$

In formula (3), $\theta(B_i|A)$ represents the parametric form; δ represents normalization factor; e is the natural constant; λ_1 and λ_2 are local characteristic function and node characteristic function, respectively. λ_1 is only related to the current node and the previous node, and λ_2 is only related to the current node. λ_1 and λ_2 values can only be 0 or 1.

Each network layer of Albert model has two subnetwork layers: the first layer is multihead self-attention mechanism layer. The second layer is the common feedforward network layer, which is used to integrate the position information of words. In addition, each subnetwork layer contains an add label layer, which is used to add and normalize the input and output of this layer, and then the residual connection is used between the two subnetwork layers [21].

Let L be the additional weight matrix to compress the spliced matrix dimension into the sequence length, Q, K, V is the vector of each corresponding label in the input sequence, and Q^{ij}, K^{ij}, V^{ij} is the weight matrix of Q, K, V; DK represents the vector dimension of each label, and Dm is the normalized activation function [22]. σ is vector point multiplication; r is the hidden layer of the network layer. The calculation formula of dynamic word vector W is as follows:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= L\sigma\left(Q^{ij}, K^{ij}, V^{ij}\right), \\ Dm &= \frac{DK\left(Q^{ij}, K^{ij}, V^{ij}\right)}{\sigma}, \\ W &= \frac{Dm}{\sigma r}(Q, K, V). \end{aligned} \qquad (4)$$

The algorithm uses the Chinese Albert pretraining model to obtain the dynamic word vector with context and then uses the conditional random field model which can effectively deal with the sequence annotation problem to complete the idiom meaning slot filling task [23]. At the same time, the multicore convolutional neural network is used for training, so that each datum has a corresponding intention label, so as to complete the intention recognition task [24]. Figure 2 shows the structure of Albert model.

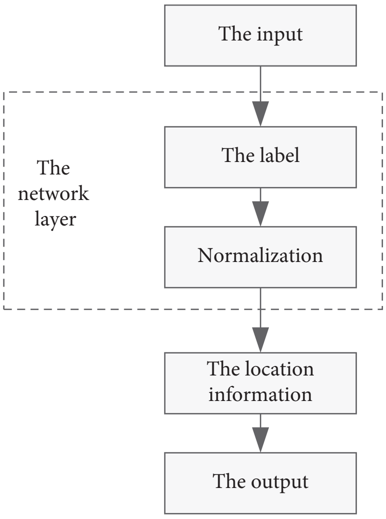


Figure 2: Albert model structure.

2.3. Conversation State Tracking Based on Slot Feature. In each round of conversation, the user input is used as an important information source for conversation status tracking, which directly contains the slot or slot value pair related to the user's needs. In the process of interaction, users are allowed to modify or improve their needs at any time. Therefore, the spoken language system needs to update the dialogue status according to the user's current round input [25]. In order to improve the recognition accuracy of new slot values and rare slot values, a multiround dialogue state tracking model based on local slot features is proposed, including coding module and state evaluation module. Since each state in state tracking consists of multiple sets of slot value pairs and the dataset of state tracking is often small [26], many slot value pairs rarely appear in the dataset, so wrong inference of rare slot value pairs often leads to poor state tracking results. The slot information word vector is spliced into the word vector of the text to be encoded to strengthen the connection between the slot information and the text sequence [27]. At the same time, the attention mechanism is applied to the slot information to obtain the weight of the hidden state, so as to obtain the feature representation of the whole text for the slot. The encoder is composed of a forward LSTM (the input is the original sequence input) and a backward LSTM (the input is the reverse sequence input) [28]. It can be expressed as

$$y_t = L(c_t). \tag{5}$$

In formula (5), y_t represents the bidirectional feature vector corresponding to the word in the input text; L represents bidirectional LSTM structure; c_t represents the input sequence; t indicates time. For the input sequence and slot set, in order to calculate the sequence representation of any slot, the coding vector of the slot is spliced with the sequence, and finally the feature vector corresponding to the word in the input text is generated, which can be expressed as

$$R = L[\omega(c_t, u) + \kappa]. \tag{6}$$

In formula (6), R represents the feature vector corresponding to the word; ω represents hidden state weight; u represents the slot set; κ represents the slot value vector. The feature vector is the set of word feature vectors in the input text and the semantic vector of the whole sentence under the condition of different slot information. In the new round of state generation, the purpose is to predict the actual intention of users in this round. Intuitively, judge whether the slot value pair in the candidate is expressed in the user's statement. There are three main contribution sources in the dialogue information, namely, the user's current round of input, the previous round of system reply, and the previous round of system action. This means that the status evaluation module will use these three information contribution sources to score the slot value pairs in each candidate [29]. As the most important information source, the user's current input can directly state the goal or request. Score a specific slot value set based on the current user's words. The calculation formula is

$$f = hg + v. \tag{7}$$

In formula (7), f represents the degree of expressing the specific slot value under consideration in the user's discourse; h represents the discourse of the current user; g represents the context representation generated by the attention mechanism; v represents a specific set of slot values. Considering that the last round of system reply can enrich or enhance the text information input by the user in detail, the user input text and the last round of system reply are jointly modeled in the evaluation module, so as to score the candidate slot value. The calculation formula is

$$f' = \sum_j h_j g_j. \tag{8}$$

In formula (8), j represents all candidate slot value sets; f' indicates the degree of expressing the candidate slot value under consideration in the user's discourse. After the score weighted sum is obtained by the scoring module, the score is mapped to the range of [0, 1] by using the activation function. The result is used as the basis for selecting the current candidate slot value, and the threshold is set. When the score exceeds the threshold, it indicates that the candidate slot value is characterized in the user requirements, and the slot value is used to update the dialog status.

2.4. Establishing Spoken Language Understanding Model Based on Deep Learning. In multiround and multitask oral comprehension, because the user conversation may switch multiple times in different tasks, not all historical information will be useful. It is very important to select the historical information related to the current conversation.

When the current conversation encounters task jump; that is, all historical information talks about other tasks, its historical information has a negative impact on the intention identification and slot filling of the current conversation, resulting in the historical information not correctly helping to understand the user's current conversation [30]. In order to alleviate this problem, this model introduces current dialogue into historical coding and combines historical dialogue with current dialogue to reduce the negative impact of irrelevant historical context. That is, an auxiliary gate unit is constructed to learn the relationship between historical context and slot position. The input of the auxiliary gate unit is the output state and history information coding of the decoding layer LSTM, which are jointly input into the auxiliary gate unit for the following calculation:

$$z = \mathrm{sigmoid}\left(\vartheta_1 o + \vartheta_2 m\right). \tag{9}$$

In formula (9), z represents a weight feature of the historical information and the output state of each step; ϑ_1 and ϑ_2 represent weights; o represents the output state of the decoding layer LSTM; m represents the historical dialogue information code; sigmoid is the activation function. Learning the relationship between the historical information and the output state of each step of the current dialogue, we can better use the historical information to find the key information in the current dialogue. The intention of user statement can affect the generation of slot. Introducing the context vector and intention representation vector of slot into a gate structure at the same time can improve the performance of slot filling task [31]. The function of the input layer is to convert the text in the dialog box into a pattern that can be understood by the computer. In the neural network channel, word vector and text vector are used to represent the classified text information. Through the expression method of information distribution, the transformed text information is mapped into high-dimensional space by using the method based on deep learning, and the semantic relationship can be inferred by using distance [32]. In this paper, Word2Vec toolkit is used to train on the prediction set. Finally, the word vector obtained by training is used as the word coding of the model. Assuming that a training sample is a given conversation, the conditional probabilities of intention label and slot label predicted by the model can be expressed as

$$\begin{cases} p_1 = \max\left(M_1\right), \\ p_2 = \prod_{1}^{v} \max\left(M_2^v\right). \end{cases} \tag{10}$$

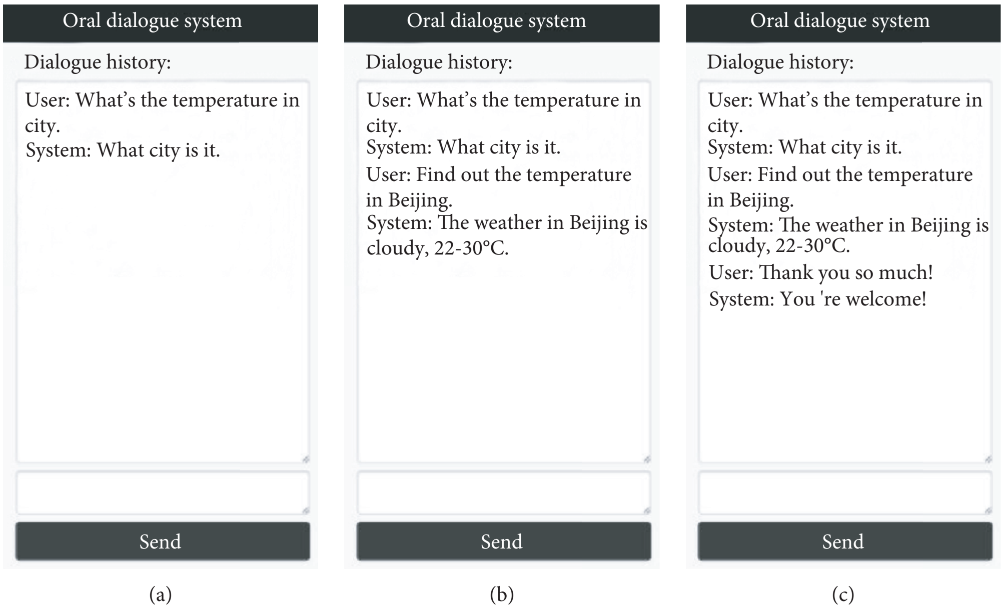
In formula (10), p_1 and p_2 represent the conditional probability of intention label and slot label; M_2^v represents the probability vector of softmax transport slot label corresponding to each word; v represents the number of corresponding slot labels; M_1 represents the intention probability vector output by softmax. The probability of the corresponding category in each dimension and the maximum probability are taken as the intention category predicted by the sample [33]. Intent recognition and slot filling share the same encoder. In the process of training the model, the two loss functions are added, and the parameters of the joint model are updated through back propagation [34]. In this model, the cross direct function is used as the loss function of the model. The model proposed in this paper introduces the current dialogue and attention mechanism into the historical information coding, learns the relationship between the historical information and the slot through the auxiliary gate unit mechanism, and effectively uses the historical dialogue information to improve the effect of the model.

3. Simulation Experiment

3.1. Dialogue Effect Test. In order to verify the performance of this method, the following simulation experiments are carried out. The experimental platform is MATLAB simulation platform. The computer used is Windows 10 system, equipped with i7 processor and running memory of 16 G. In this simulation experiment, the API of iFLYTEK is adopted for speech recognition and speech synthesis and transplanted to the ROS operating system. The spoken language understanding part is the content of this paper. Each module is encapsulated and organized as a node of ROS. The speech recognition node collects the user's audio and publishes the recognition results in the form of topic after speech recognition. Using spoken language understanding node, and after the question preprocessing, question and answer module, and dialogue management module based on slot features discussed above, the processing results are released again in the form of topic. Finally, the speech synthesis node subscribes to the topics published by the semantic analysis node and feeds back to the user in the form of audio after synthesis to complete a round of interaction. In practical application, the spoken language system runs on the ROS system and the microphone collects the user's speech and finally feeds back to the user with audio. Generally, there is no graphical interface. For demonstration, this section uses flask as the background framework to realize the API of semantic recognition and Apache as the static server of web resources, and the front end uses HTML, JavaScript, and CSS to realize a web application. Taking weather query as an example, the multiround interaction of the spoken language understanding method is shown in Figure 3.

According to the demonstration results in Figure 3, the spoken language understanding method proposed in this paper can ask and answer common questions and has the effect of real-time interaction of multiple rounds of dialogue. On this basis, the performance of the proposed method is further tested.

3.2. Experimental Environment. The KVRET dataset used in this paper is from Stanford Natural Language Processing Group. Task-oriented dialogue focuses on participating in the dialogue on specific topics initiated by users. Generally speaking, if researchers want to do task-oriented dialogue and the training model dataset is not large and diverse enough, the next work is likely to be blocked. To help



Oral dialogue system

Dialogue history:

User: What's the temperature in city.
System: What city is it.

Oral dialogue system

Dialogue history:

User: What's the temperature in city.
System: What city is it.
User: Find out the temperature in Beijing.
System: The weather in Beijing is cloudy, 22-30°C.

Oral dialogue system

Dialogue history:

User: What's the temperature in city.
System: What city is it.
User: Find out the temperature in Beijing.
System: The weather in Beijing is cloudy, 22-30°C.
User: Thank you so much!
System: You 're welcome!

(a) (b) (c)

Figure 3: Schematic diagram of multiwheel interaction. (a) First round of dialogue. (b) Second round of dialogue. (c) Third round of dialogue.

alleviate this problem, the Stanford natural language processing group published a corpus. This dataset contains more than 3000 rounds of conversation data, mainly distributed in schedule, weather retrieval, and navigation. Because there is only one task in each conversation in KVRET dataset, in order to fit the reality, this paper reorganizes the KVRET dataset and obtains the conversation dataset containing multiple tasks. The reorganization method is as follows: two dialogue paragraphs of schedule, weather retrieval, and navigation are randomly selected for cross splicing, so that the spliced dialogue paragraphs contain two different tasks. The learning rate of shallow neural network is set to 0.064, the size of context window is set to 8, the dimension of word vector is set to 150, and the number of hidden layer neurons is 120. The number of training steps is set to 10 and the number of iterations is 100. In order to compare the performance of this method, it is compared with the spoken language understanding methods based on circular network, context information, and label decomposition. The experimental evaluation criteria are the accuracy and F1 value, which are widely used at present.

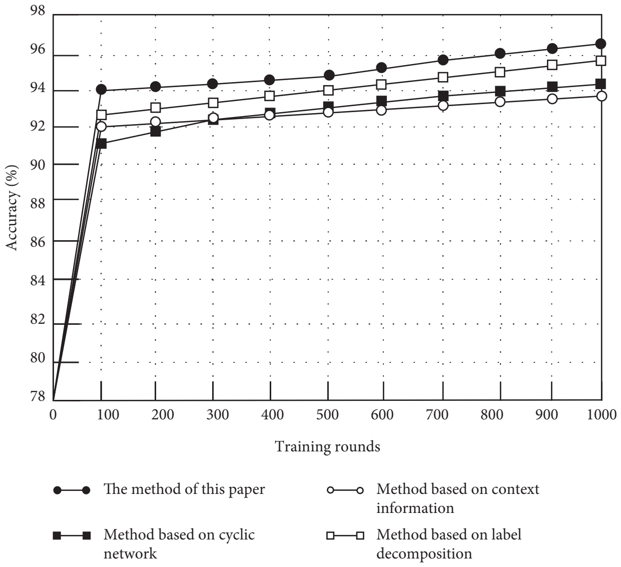


Figure 4: Comparison results of accuracy test.

3.2.1. Experiment of Measuring Accuracy. According to the experimental environment, taking 1000 rounds of training as an example, the accuracy of the four methods is calculated. The precision experiment comparison diagram of Figure 4 is obtained.

As can be seen from Figure 4, the test accuracy of these four methods is basically relatively stable and the fluctuation is small. The test accuracy of the three methods based on cyclic network, context information, and oral understanding of label decomposition is always higher than 90%, and the test accuracy is relatively stable. The test accuracy of the method studied in this paper remains above 94%, up to 97%, which is always higher than the other three methods, indicating that the performance of this method is better. According to the accuracy comparison results in Figure 4, the comparison diagrams of the highest accuracy and the lowest accuracy of different methods can be drawn, as shown in Figure 5.

As shown in Figure 5, the highest accuracy of the method based on cyclic network is 93%, and the lowest accuracy is 91%. The highest accuracy of context information method is 93.5%, and the lowest accuracy is 92%. The highest accuracy of label decomposition method was 96%, and the lowest accuracy was 92.5%. The highest accuracy of this method is 96% and the lowest is 94%. Compared with the oral

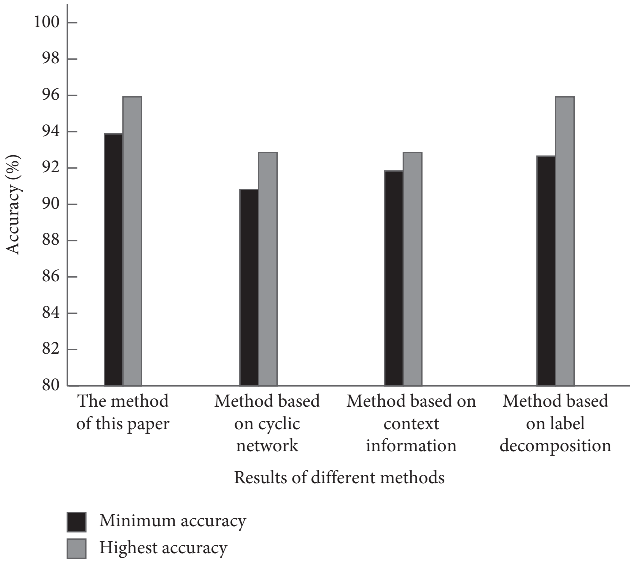


Figure 5: Comparison of maximum and minimum accuracy of different methods.

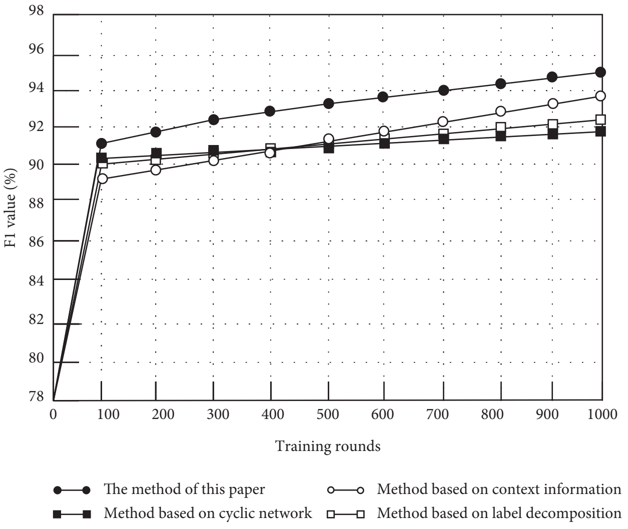


Figure 6: Comparison results of F1 value test.

comprehension method based on circular network, context information, and label decomposition, the accuracy is 3%, 2.5%, and 3.5% higher, respectively. It shows that this method has high accuracy and good practical application performance.

3.2.2. Test of F1 Value. In order to better test the actual performance of the method in this paper, the F1 value of the four methods is tested. F1 value is an index used to measure the accuracy of binary classification model in statistics. Its maximum value is 1 and its minimum value is 0. The F1 value is tested in the form of percentage, and the comparison test results are shown in Figure 6.

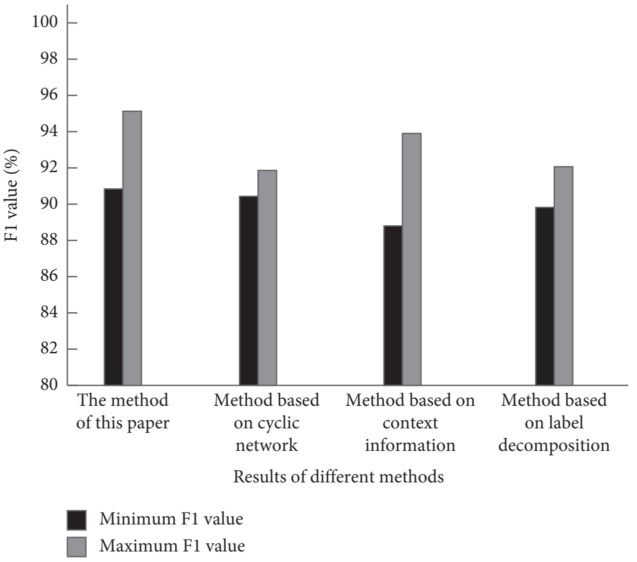


Figure 7: Comparison of the highest and lowest F1 values of different methods.

It can be seen from Figure 6 that the test F1 value of the method studied in this paper is higher than that of the oral comprehension method based on circular network, context information, and label decomposition. Basically, the F1 values of the four methods are relatively stable and have little fluctuation. According to the F1 value comparison results in Figure 6, the comparison diagrams of the highest F1 value and the lowest F1 value of different methods can be drawn, as shown in Figure 7.

As can be seen from Figure 7, the highest F1 value is 92% and the lowest F1 value is 91% based on the cyclic network method. The highest F1 value of context information method is 94%, and the lowest F1 value is 89%. The highest F1 value of label decomposition method is 92%, and the lowest F1 value is 90%. The highest F1 value is 95%, and the lowest F1 value is 91%. The highest F1 value is 3%, 1%. and 3% higher than the oral comprehension methods based on circular network, context information. and label decomposition, respectively. F1 value is an effective evaluation standard for comprehensive accuracy and recall, which can comprehensively reflect the performance of this method. Experiments can prove that this method has good application performance and has certain advantages.

Through experiments, it can be concluded that the method proposed in this paper has the highest accuracy of 97%, the highest accuracy of 96%, and the highest F1 value of 95%, which can realize man-machine oral English understanding and has a good application prospect.

4. Conclusion

Man-machine dialogue system is the concentrated embodiment of the level of artificial intelligence. As the core part of man-machine dialogue system, oral comprehension model is the focus and difficulty of research. This paper proposes an spoken language understanding method based on deep learning. The test results show that this method can significantly improve the accuracy and F1 value and has high practical value. The application research of oral comprehension is a complex and far-reaching topic, and there are still deficiencies in this paper. Due to the limitation of hardware equipment, too many rounds will lead to too large model parameters and failure to run. However, in actual situations, such as the communication between online customer service and users, there may be dozens or even hundreds of rounds of dialogue between them. How to solve the difficulty of multiple rounds of dialogue needs further research. The data of oral comprehension in this paper focus on the fact that each sentence contains only one intention. However, in real life, a sentence may contain multiple intentions, which needs further exploration.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

References

[1] L. Hou, Y. Li, and C. Li, "Review of research on task-oriented spoken language understanding," *Computer Engineering and Applications*, vol. 55, no. 11, pp. 7–15, 2019.

[2] P. Qiao, "Automatic language generation simulation of two-way interactive robot," *Computer Simulation*, vol. 36, no. 4, pp. 310–314, 2019.

[3] A. Vanzo, D. Croce, E. Bastianelli, R. Basili, and D. Nardi, "Grounded language interpretation of robotic commands through structured learning," *Artificial Intelligence*, vol. 278, pp. 1–21, 2020.

[4] J. Zhang, H. Huang, Y. Hu, and Y. Wushour, "Modified recurrent neural networks in spoken language understanding," *Computer Engineering and Applications*, vol. 55, no. 18, pp. 155–160, 2019.

[5] X. Yang, J. Wang, L. Qi-yuan, and L. Shou-shan, "Intention detection in spoken language based on context information," *Computer Science*, vol. 47, no. 1, pp. 205–211, 2020.

[6] Y. Xu and H. Huang, "Spoken language understanding model based on label decomposition," *Computer Engineering*, vol. 45, no. 7, pp. 237–241, 2019.

[7] Q. Dong, H. Li, G. Cao, and L. Xia, "An exploratory posts detecting method for MOOC forums based on deep learning," *Library and Information Service*, vol. 63, no. 5, pp. 92–99, 2019.

[8] G Wang and X Huang, "Convolution neural network text classification model based on word2vec and improved TF-IDF," *Journal of Chinese Computer Systems*, vol. 40, no. 5, pp. 1120–1126, 2019.

[9] D Yang, Y Wu, and F. Chun-xiao, "Chinese short text keyphrase extraction model based on attention," *Computer Science*, vol. 47, no. 1, pp. 193–198, 2020.

[10] E. Lv, X. Wang, and Y. Cheng, "Deep convolution neural network learning based on deconvolution feature extraction," *Control and Decision*, vol. 33, no. 3, pp. 447–454, 2018.

[11] Q. Zhao, X. Cai, and L. Bo, "Text feature extraction method based on LSTM-Attention neural network," *Modern Electronics Technique*, vol. 41, no. 8, pp. 167–170, 2018.

[12] T. Xue, Y. Wang, and M Nan, "Convolutional neural network based on word sense disambiguation for text classification," *Application Research of Computers*, vol. 35, no. 10, pp. 2898–2903, 2018.

[13] H. Wei and L. Fan, "Feature extraction of binary documents using neural network," *Communications Technology*, vol. 12, pp. 2881–2887, 2019.

[14] J. Liu, Y. Li, and M. Lin, "Review of intent detection methods in human-machine dialogue system," *Computer Engineering and Applications*, vol. 55, no. 12, pp. 1–7, 2019.

[15] L. Hou, Y. Li, and M. Lin, "Joint recognition of intent and semantic slot filling combining multiple constraints," *Journal of Frontiers of Computer Science & Technology*, vol. 14, no. 9, pp. 1545–1553, 2020.

[16] S. C. Akkaladevi, M. Plasch, M. Hofmann, and A. Pichler, "Semantic knowledge based reasoning framework for human robot collaboration," *Procedia CIRP*, vol. 97, no. 5, pp. 373–378, 2021.

[17] A. K. Bosen and E. Buss, "Short-term audibility is a better predictor of vocoded speech-in-speech recognition than long-term target-to-masker ratio," *Journal of the Acoustical Society of America*, vol. 148, no. 4, p. 2465, 2020.

[18] Q Zhou and Z. Li, "BERT based improved model and tuning techniques for natural language understanding in task-oriented dialog System," *Journal of Chinese Information Processing*, vol. 34, no. 5, pp. 82–90, 2020.

[19] V. Kadyan, M. Dua, and P. Dhiman, "Enhancing accuracy of long contextual dependencies for Punjabi speech recognition system using deep LSTM," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 517–527, 2021.

[20] M. Wang, D. Yu, Y. Rui, H Wenpeng, and Z. Dongyan, "Chinese multi-turn dialogue tasks based on HERD model," *Journal of Chinese Information Processing*, vol. 34, no. 8, pp. 78–85, 2020.

[21] Y. Xu and H. Huang, "Spoken language understanding method based on recurrent neural network with persistent memory," *Computer Engineering and Applications*, vol. 12, pp. 145–148, 2019.

[22] Z. Matteo, B. Luca, S. Ivan, and G. Alfonso, "Evaluating different natural language understanding services in a real business case for the Italian language," *Procedia Computer Science*, vol. 176, pp. 995–1004, 2020.

[23] M. Mcshane and S. Nirenburg, "Context for language understanding by intelligent agents," *Applied Ontology*, vol. 14, no. 4, pp. 1–34, 2019.

[24] Z. Yang, L. Wang, and Y. Wang, "Application research of deep learning algorithm in question intention classification," *Computer Engineering and Applications*, vol. 55, no. 10, pp. 154–160, 2019.

[25] J. Hu and H. Tao, "Design and implementation of domain question answering system based on deep learning," *Journal of Chengdu University Of Information Technology*, vol. 34, no. 3, pp. 232–237, 2019.

[26] N. Zhu, J. Dong, and Z. Zhang, "A human-computer dialogue model for digital reference consultation in library," *Library and Information Service*, vol. 63, no. 6, pp. 5–11, 2019.

[27] H. Akay and S.-G. Kim, "Measuring functional independence in design with deep-learning language representation models," *Procedia CIRP*, vol. 91, pp. 528–533, 2020.

[28] Z. Chen and H. Yang, "Yi language speech recognition using deep learning methods," in *Proceedings of the 2020 IEEE 4th*

Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, Chongqing, China, June 2020.

[29] M. Ali, M. L. Rahman, J. Chaki, N. Dey, and K. Santosh, "Machine translation using deep learning for universal networking language based on their structure," *International Journal of Machine Learning and Cybernetics*, vol. 3, pp. 1–12, 2021.

[30] A. H. Zadeh, Z. Poulos, and A. Moshovos, "Deep learning language modeling workloads: where time goes on graphics processors," in *Proceedings of the 2019 IEEE International Symposium on Workload Characterization (IISWC)*, November 2019.

[31] R. W. Filice, "Deep-learning language-modeling approach for automated, personalized, and iterative radiology-pathology correlation," *Journal of the American College of Radiology: JACR*, vol. 16, no. 9, pp. 1286–1291, 2019.

[32] L. Wu, D. Han, and Z. Du, "Assembly language and assembler for deep learning accelerators," *High Technology Letters*, vol. 25, no. 4, pp. 42–50, 2019.

[33] Y. Kim, J. H. Lee, S. Choi et al., "Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records," *Scientific Reports*, vol. 10, no. 1, Article ID 20265, 2020.

[34] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Deep learning for natural language processing in radiology-fundamentals and a systematic review," *Journal of the American College of Radiology*, vol. 17, no. 5, pp. 639–648, 2020.