Robot Human-Machine Interaction Method Based on Natural Language Processing and Speech Recognition

Shuli Wang, Fei Long*
Foreign Language School, Harbin University of Commerce, Harbin, 150028, China

***Abstract*—With the rapid development of artificial intelligence technology, robots have gradually entered people's lives and work. The robot human-machine interaction system for image recognition has been widely used. However, there are still many problems with robot human-machine interaction methods that utilize natural language processing and speech recognition. Therefore, this study proposes a new robot human-machine interaction method that combines structured perceptron lexical analysis model and transfer dependency syntactic analysis model on the basis of existing interaction systems. The purpose is to further explore language based human-machine interaction systems and improve interaction performance. The experiment shows that the testing accuracy of the structured perceptron model reaches 95%, the recall rate reaches 81%, and the F1 value reaches 82%. The transfer dependency syntax analysis model has a data analysis speed of up to 750K/s. In simulation testing, the new robot human-machine interaction method has an accuracy of 92% compared to other existing methods, and exhibits excellent robustness and response sensitivity. In summary, research methods can provide a theoretical and practical basis for the improvement of robot interaction capabilities and the further development of human-machine collaboration.**

Keywords—Human-computer interaction; speech recognition; natural language processing; lexical analysis; syntactic analysis

I. Introduction

Natural language processing is an important branch of computer science and artificial intelligence that deals with techniques that enable computers to understand, interpret and generate human language. Speech recognition, on the other hand, is the technology that converts human speech into text. With the continuous development of natural language processing and speech recognition technology, robot human-machine interaction has gradually become a research hotspot [1]. There are three common human-machine interaction methods for language robots, the earliest of which was the use of rule-based human-machine interaction methods, that is, language recognition and response through pre written rules [2]. But this method requires manual writing of a large number of rules and has poor adaptability. Then, statistical human-machine interaction methods are used to identify and understand speech inputs by establishing language and speech models [3]. But this method requires a large amount of corpus for training, which is time-consuming and not accurate. Finally, the human-machine interaction method of deep learning is utilized, which automatically extracts speech and text features through deep learning models, effectively improving the performance of language recognition [4]. With the increasing demand of people, the human-machine environment has become even more harsh, so simple deep learning models can no longer meet high requirements for completing human-machine interaction tasks. In the existing research, the design of robot human-robot interaction system that combines natural language processing technology and language recognition technology is still in the minority, and only proposes, for example, a sentiment analysis robot that can better understand the user's emotions and respond accordingly by analyzing the pitch, speed, volume of speech, and the emotional color of text. Or multimodal interaction robots provide a more natural interaction experience by analyzing the user's voice, text and body language. Despite the success of these human-robot interaction system designs in the existing market, they still face challenges such as dealing with complex and variable natural language understanding challenges and maintaining high accuracy and adaptability in diverse and dynamic environments. Therefore, the research attempts to combine natural language processing and speech recognition to propose a novel approach to human-computer interaction. The method improves the two major steps of lexical analysis and syntactic analysis, and introduces structured perceptual machine and transfer-dependent syntactic analysis for improvement respectively, to enhance the computational performance of each module, and to achieve the goal of enhancing the recognition accuracy of human-computer interaction. The study aims to explore the latest progress of natural language processing and speech recognition technologies in the field of robot human-robot interaction, analyze the limitations of existing technologies, and look forward to the future development trend. Therefore, this study attempts to combine natural language processing with speech recognition and proposes a new human-computer interaction method. This method improves the two major steps of lexical analysis and syntactic analysis to enhance the recognition accuracy of human-computer interaction. This study is divided into five sections. Section I is an introduction to the overall content of the article. Section II is an analysis and summary of research on others. Section III introduces how the improved lexical analysis model and syntactic analysis model are constructed. Section IV tests the performance of the new human-computer interaction system. Section V is a summary of the paper.

Studying the application of natural language processing and speech recognition in robot human-robot interaction is crucial for improving the level of intelligence and user experience of interaction technology. It can not only improve the efficiency of communication between people and robots, but also provide better assistive tools for specific groups (e.g., people with disabilities). In addition, this research is also

*Corresponding Author.

valuable for understanding human language and communication patterns, and can contribute to the development of knowledge in related fields. The results of the above research can reveal the efficacy and limitations of natural language processing and speech recognition technologies in different contexts and provide an empirical basis for theoretical models. For example, by analyzing the performance of robots in different linguistic contexts, the research can help us better understand the impact of language complexity on the performance of the technology. Meanwhile research findings can stimulate new research questions, such as how robots deal with dialects or non-standard languages, or how to deal with metaphors and humor in language more effectively.

II. Related Works

With the advancement of artificial intelligence technology, robots have begun to play an important role in various fields. However, one of the key challenges in making robots more intelligent and humane is how to achieve natural and smooth interaction between robots and humans. Currently, many scholars have conducted in-depth research in this field. Freire Obregon D et al., in order to further improve the recognition performance of biometric verification in human-computer interaction, used independent interest frameworks to improve the accuracy of robot audio recognition. By using high confidence image facial recognition to avoid errors caused by similarity in appearance, its accurate resolution after simulation testing is higher than traditional methods, providing a new method for robot recognition technology [5]. Ko et al. introduced a nonverbal social behavior dataset to improve the recognition and learning efficiency of robots in different scenarios. This dataset includes human body index and bone data, which robots use sensors to identify and analyze, and guide subsequent behavioral operations. The learning rate of robots under this dataset has significantly improved, and their behavior inference and response abilities have significantly improved [6]. Kim et al. conducted opinion interviews with 70 ordinary users in order to further optimize the evaluation indicators of food aid robots and improve user experience. The collection of over 500 suggestions on robot interface design and security provides more solutions for the usability, emotional value, and functional construction of food aid robots [7]. Roda Sanchez L et al. proposed an intelligent system that combines the Internet of Things and human-machine action collaboration to improve the efficiency of product manufacturing processes in the context of digital industry. This system is centered around human-computer interaction, reflecting the natural interaction between IoT inertial measurement unit equipment and robotic arms. The system meets the basic requirements of modern digital industrial manufacturing in terms of real-time performance, success rate, and acceptable level [8].

The development of speech recognition technology began in the 1950s, with initial research mainly based on spectral analysis and pattern matching of audio signals. However, due to limitations in computing power and data volume at the time, the accuracy and stability of speech recognition were not satisfactory. With the development of technology, speech recognition technology has made significant breakthroughs in many disciplinary fields. Alsayadi et al. proposed an automatic language recognition system based on convolutional neural networks to address the issue of distinguishing between Arabic language recognition techniques with and without inflections. The system is tested on a standard Arabic single speaker corpus. The results show that recognition techniques with neural networks are superior to traditional recognition techniques, reducing word error rates by 5.24% [9]. Lin's team designed a recognition technique that combines recursive neural network embedding blocks to extract advanced features in order to reduce speech loss caused by radio communication propagation. This technology integrates multi language speech recognition into a single model, thus avoiding class imbalance. The Chinese and English character error rates of this technology are 4.4% and 5.6%, respectively, which are significantly better than other methods [10]. Dong et al. proposed a significant time series method using connectionist time classification to address the issues of delayed response time and emotional noise in continuous emotional speech recognition technology. This method treats sentence labels as a chain of emotional significant events and non-emotional significant event states. This method can continuously improve the performance of emotion recognition, and when the consistency of significant emotional events is high, this improvement is more significant [11]. Yerigeri et al. proposed a mechanical and efficient speech emotion recognition technology that utilizes stress level analysis to explore the impact of stress on people's emotional changes. This technology utilizes learning algorithms to evaluate auditory and visual cues, and uses a pressure speech database for performance analysis. The overall performance of this technology is good, with an accuracy rate of 90.66% for stress related emotion recognition [12].

In summary, many academic teams have conducted extensive research in the field of robot interaction design and recognition technology, and have achieved remarkable results. Overall, the research status of robot human-machine interaction method design is constantly developing and innovating, involving the intersection and integration of multiple disciplines. This study attempts to apply natural language processing technology and speech recognition technology to the design of robot human-machine interaction, exploring how these technologies can be applied to intelligent robots to achieve more natural and convenient human-machine interaction.

III. Design of a Robot Human-Machine Interaction Method Model Combining Natural Language Processing and Speech Recognition

The natural language processing technology in robot human-machine interaction methods enables machines to understand human intentions and instructions by analyzing and understanding human language [13]. Speech recognition technology converts human speech input into text or commands, thereby achieving interaction with machines. The first section of this study will improve and innovate speech recognition technology, and the second section will improve natural language processing methods.

A. Design of Lexical Analysis Model Based on Structured Perception Machine in Natural Language Processing

The most common ways of human-computer interaction are verbal communication and behavioral communication. Speech communication involves speech recognition, while behavioral communication involves image recognition. And verbal communication is nothing more than the simplest way of interaction. The existing processing methods for natural language include lexical analysis, syntactic analysis, semantic analysis, speech recognition, and speech synthesis. Lexical analysis is the most important step in natural language processing. This step consists of three parts, namely Chinese word segmentation, part of speech tagging, and entity naming recognition. For Chinese, the results of lexical analysis will directly affect subsequent natural language processing. Perception machine is a basic binary classification algorithm proposed by American scientist Frank Rosenblatt in 1957 [14]. The goal of the perceptron is to linearly classify input data into two different categories. The perceptron model is shown in Fig. 1.

In Fig. 1, w represents the normal vector and b represents the intercept. The processing of classification problems by perceptron models is called decision boundaries. The neighborhood differentiation of this model is obvious, and the feasibility of linear implementation is high. In a space, from input to output, the perceptron model calculates the formula as shown inEq. (1).

$$f(x) = sign(w \cdot x + b) \tag{1}$$

In Eq. (1), x represents a point in the input space. b represents offset. w represents the weight value. $w \cdot x$ represents the weight of the point. $sign$ represents a symbolic function. This function is shown in Eq. (2).

$$\mathrm{si}\,gn(x) = \begin{cases} +1 \longrightarrow x \geq 0 \\ -1 \longrightarrow x < 0 \end{cases} \tag{2}$$

In Eq. (2), +1 or -1 are usually used as indicators of input and output, and the geometric interpretation of the perceptron model is shown in Eq. (3).

$$w \cdot x + b = 0 \tag{3}$$

In Eq. (3), $w \cdot x + b$ belongs to the set of all linear classification models in the feature center of the perceptual model. It corresponds to a hyperplane in the feature space. Part of speech tagging is a typical type of structured prediction. The structured prediction scoring of the perceptron model is shown in Eq. (4).

$$\hat{y} = \arg\max_{y \in Y} score_{\lambda}(x, y) \tag{4}$$

In Eq. (4), λ represents the prediction model. Y represents all selectable structures. Usually for linear models, structured perceptron is used as a training algorithm, and the classifier can assist in predicting problems such as sequence annotation. The structured prediction is shown in Eq. (5).

$$\hat{y} = \arg\max_{y \in Y} (w \cdot \phi(x, y)) \tag{5}$$

In Eq. (5), x and y represent independent variables. $\phi(x, y)$ represents a characteristic function. After the product of the new feature vector and weight points, the highest output structure is used as the decoding of the sequence annotation problem. The decoding process of this method is described in Eq. (6).

$$\delta_{t,i} = \begin{cases} w \cdot \phi(s_{0,} s_i, x_1), i = 1, \cdots, N \\ \max_{1 \leq j \leq N} (\delta_{t-1} + w \cdot \phi(s_j, s_i, x_t)), i = 1, \cdots, N \end{cases} \tag{6}$$

In Eq. (6), N represents the corresponding state. s represents the score. j belongs to any one of the state sets. t represents the time. The maximum score calculation is shown in Eq. (7).

$$S = \max_{1 \leq i \leq N} \delta_{T,i} \tag{7}$$

In Eq. (7), S represents the maximum score. T represents the corresponding time set under this score. i represents the optimal path. When there is a local optimal solution, the maximum score at this point corresponds to the label under that path, which is the i-value. In summary, this study integrates the structured perceptron model into the natural language processing of human-computer interaction, and proposes a new natural language processing framework for human-computer interaction, as shown in Fig. 2.

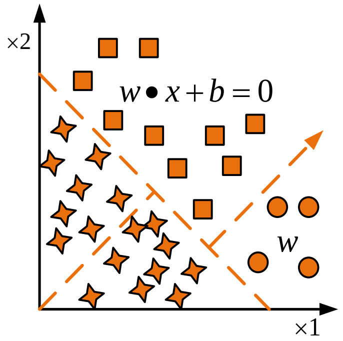


Fig. 1. Perceptron model.

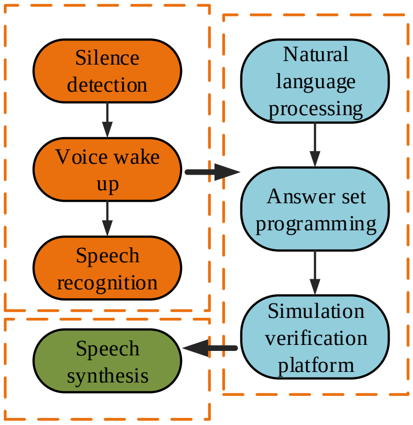


Fig. 2. Human-computer interaction natural language processing framework.

In Fig. 2, the natural language processing framework is mainly divided into three parts. Firstly, the language detection and recognition section provides information on the sounds emitted by external users. The second step is to convert the extracted sound information into linear features, which are labeled by programming data. Finally, the structure aware machine algorithm performs lexical analysis on these labeled data and converts them into communicative text.

B. Design of Key Information Speech Recognition Method Based on Dependency Syntactic Analysis

Natural language is constantly changing and cannot be represented by simple linear symbols. Meanwhile, due to the extremely complex construction of natural language, robot acquisition analysis is still too abstract. Therefore, this study continues to focus on constraint transformation of structured prediction results and remove abstraction. Natural language processing is generally manifested as the relationship between words in a sentence. Dependency syntax analysis defines this relationship as a binary nonequivalence relationship, namely the master-slave relationship [15]. To more vividly display the dependency relationships between words, the dependency tree is obtained by dividing the words in order, as shown in Fig. 3 [16].

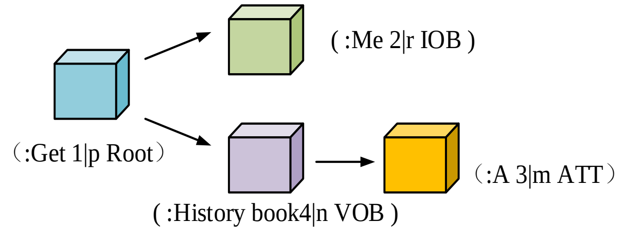


Fig. 3. "Get me a history book" depends on the syntax tree.

In Fig. 3, there are four phrase relationships in the sentence "Help me get a history book". Among them, "get" has the strongest interdependence with "me" and "history Book", while "a" has the lowest interdependence. There are two common implementation methods for dependency syntax analysis, namely the combination graph or transfer analysis method. By combining the dependency syntax analysis of graphs, using independent assumptions and establishing a model, the optimal branch solution can be found in the entire dependency tree model [17]. As shown in Eq. (8).

$$Score(x,d) = \sum_{p \subseteq d} Score_{subtree}(x,p) \tag{8}$$

In Eq. (8), w represents the weight vector. p represents a branch that conforms to the hypothesis. The analysis method of combining graphs completely depends on the maximum number of dependencies allowed in the tree. The strength of obtaining effective information in a simultaneous graph model depends on the number of features used. Under normal operation, the graph model utilizes a feature extractor to extract features for each word and transmit them to the classification scorer as scores for dependency relationships.

Based on the characteristics of dependency parsing, an improvement was made on its state machine, and conditional analysis was introduced to obtain the transfer dependency parsing algorithm [18]. The corresponding machine state under this algorithm is shown in Eq. (9).

$$s(x)_0 = ([x_0]_\sigma, [x_1 \cdots x_k]_\beta) \tag{9}$$

In Eq. (9), σ represents the stack. β represents the queue. x represents a sentence. k represents the tail element of the queue. $s(x)_0$ represents the initial state. The set of state transitions in transfer dependency syntactic analysis roughly includes move in actions, left reduction, and right reduction. The move in action is shown Eq. (10).

$$(a, x_i | \beta, A) \Rightarrow (a | x_i, \beta, A) \tag{10}$$

In Eq. (10), A represents the set of constructed dependent edges. x_i represents the state of being pushed onto the stack. Ensure that the queue is in a non-empty state and push the elements in the queue onto the stack. The calculation formula for left reduction is shown in Eq. (11).

$$([x_0 \cdots x_k, x_j, x_i], \beta, A) \Rightarrow ([x_0 \cdots x_k, x_i], \beta, A \cup (i,, j) \tag{11}$$

In Eq. (11), i and j each represent a new element. When the element in the stack is greater than 1, the two elements at the top of the stack are introduced into the set (j,i), and then the i element is re pushed onto the stack. The calculation formula for rightward reduction is shown in Eq. (12).

$$([x_0 \cdots x_k, x_j, x_i], \beta, A) \Rightarrow ([x_0 \cdots x_k, x_i], \beta, A \cup (j,i) \tag{12}$$

In Eq. (12), i and j each represent a new element. When the elements in the stack are greater than 1, the two elements at the top of the stack are introduced into the set (j,i), and then the j element is re pushed onto the stack. For example, in the short sentence "You drink water", there is a subject verb relationship between the words "you" and "drink", and a verb object relationship between "drink" and "water". Therefore, machines need two steps to establish a syntactic dependency tree during learning.

The analysis process of transfer dependency syntax analysis roughly includes transfer system, feature extraction, and action sequence transformation. The transfer system mainly covers some executable actions and their conditions. Feature extraction selects object features through manually set feature templates that combine single words, two words, and three words. The transformation of action sequences is commonly divided into static specification transfer and dynamic specification transfer [19]. To improve the persuasiveness and feasibility of machine learning, a training process for machine learning was proposed by selecting dynamic norm transfer and combining it with a structured perceptron model, as shown in Fig. 4.

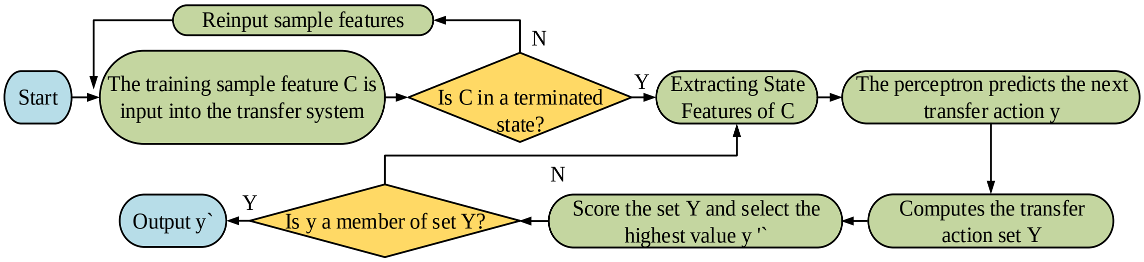

Reinput sample features
N
Start
The training sample feature C is input into the transfer system
Is C in a terminated state?
Y
Extracting State Features of C
The perceptron predicts the next transfer action y
N
Y
Output y`
Is y a member of set Y?
Score the set Y and select the highest value y '`
Computes the transfer action set Y

Fig. 4. Machine learning transfer dependency parsing training flow.

In Fig. 4, from the perspective of data structure, state judgment has been added in both the training sample input to the transfer system stage and the selection of the highest scoring feature stage, indicating that the entire process of machine learning is robust and feasible. Because after selecting features and scoring in the model, the results are usually directly output. But the process introduces attribution judgment for actions, thereby reducing the chances of self-error and increasing accuracy.

C. Construction of a Robot Human-Machine Interaction Model Combining Natural Language Processing and Speech Recognition

The first two chapters have already introduced the implementation steps of lexical analysis and syntactic analysis. This time, we will elaborate on the design of the human-machine part of the robot in the human-machine interaction model [20]. Assuming that the position of the robot is composed of an initial inertial coordinate system and a carrier reference coordinate system, the coordinate representation of the mobile robot is shown in Fig. 5.

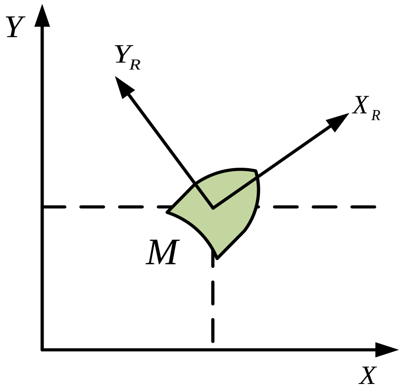


Fig. 5. Robot moving coordinate representation.

In Fig. 5, XOY represents the inertial coordinate system. X_RMY_R represents the carrier coordinate system. The representation of robots in three-dimensional coordinates changes over time. If the position at time k is inferred from the position at time $k+1$, the motion model is shown in Eq. (13) [21].

$$X_r(k+1) = f(X_r(k), u(k)) \tag{13}$$

In Eq. 13, u represents the navigation weight value. f represents the state transition function. In the initial inertial coordinate system, the position vector at time $k+1$ is represented as the motion model of the robot, as shown in Eq. (14).

$$\begin{bmatrix} x(k+1) \\ y(k+1) \end{bmatrix} = \begin{bmatrix} \cos\theta - \sin\theta \\ \sin\theta, \cos\theta \end{bmatrix} \begin{bmatrix} x(k) \\ y(k) \end{bmatrix} \tag{14}$$

In Eq. (14), θ represents the heading. x and y represent the horizontal and vertical coordinates of the mobile robot, respectively. At this point, the input variable of k is $[v_r, w_r]$, where v_r represents speed. w_r represents angular velocity. After constructing the robot movement model, a new robot human-machine interaction method flow is proposed by combining the improvement scheme of natural language processing module and speech recognition module, as shown in Fig. 6 [22].

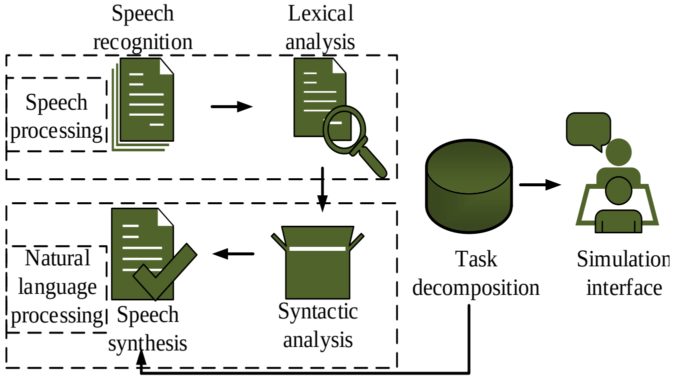


Fig. 6. New human-computer interaction process.

In Fig. 6, speech recognition and natural language processing are the two main parameter analysis modules. The speech recognition module serves as the main input part of the entire human-computer interaction process. The speech synthesis and simulation interface is the output part. In each stage, semantic recognition and synthesis are responsible for recognizing and expressing sounds, while simulation interfaces are responsible for verifying the effectiveness of interaction results [23].

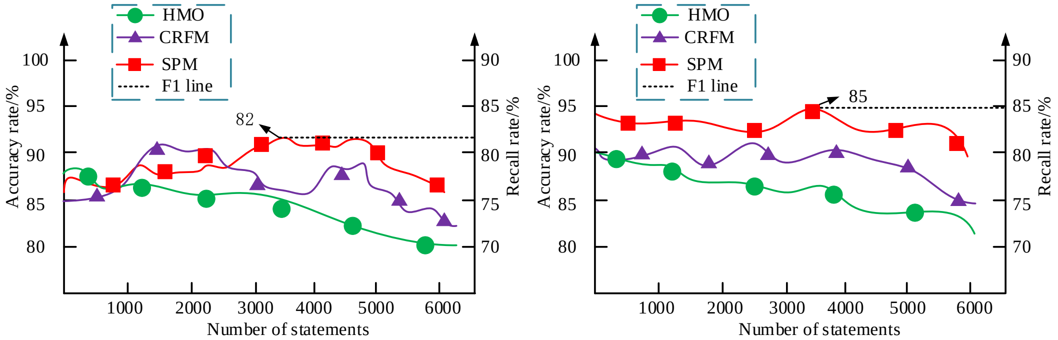
IV. Performance Testing of a Robot Human-Machine Interaction Model Combining Natural Language Processing and Speech Recognition

The first step is to establish a suitable experimental environment, set various experimental parameters and indicators, and create a reliable experimental corpus. Further extracting key feature information into robot human-machine interaction systems through natural language processing of structured perceptron models and speech recognition through transfer dependency syntax analysis. The accuracy (P), recall (R), and comprehensive evaluation index F1 value are used as evaluation indicators to conduct performance tests on the human-machine interaction model.

A. Performance Testing of Natural Language Processing Models and Speech Recognition Models

To verify the performance of the proposed new human-computer interaction model, natural language processing and speech recognition modules were tested separately. The computer system used in the experiment is Window10, the development environment is Pycharm, the language is Pyuthon3.7, and the CPU is (Intel ® Core TM i7-9700CPU@3.00GHz × 8), GPU is (NVIDIA GeForce RTX 3060 SUPER). The data source is the BCC Modern Chinese Corpus of Beijing Language and Culture University. This corpus includes over 30000 pieces of corpus information from various fields. According to an 8:2 ratio, the corpus information is divided into a training group and a testing group.

In natural language processing experiments, P value, R value, and F1 value are used as reference indicators to compare the performance of existing hidden Markov models, conditional random field models, and research models. The hidden Markov model is represented by HMO, the conditional random field model is CRFM, and the structured perceptron model is SPM. The experimental test results are shown in Fig. 7.



(a) Training set (b) Test set

Fig. 7. Test results of different lexical analysis models.

Fig. 7 (a) and Fig. 7 (b) show the test results curves of three models in the training and testing sets. In Fig. 7, the performance of the training and testing sets of the three models shows a slow downward trend. However, the test P-values, R-values, and F1 values of the hidden Markov model perform the worst in both the training and testing sets. The combination of structured perceptron models proposed in the study performed the best, with an accuracy of up to 95%, a recall rate of up to 81%, and an F1 value of up to 82% in the test set.

Lexical analysis, as the most crucial step in natural language processing, provides the foundation for all subsequent speech processing steps. The accuracy of its analysis directly affects the efficiency of subsequent processes. Therefore, tests are conducted on the P-value, R-value, and F1 value of the three subtasks in the sequence annotation of the structured perceptron model, namely word segmentation, part of speech annotation, and named entity recognition. The results are shown in Fig. 8.

In Fig. 8, after inputting instructions from the BCC modern corpus, the average score of part of speech tagging in the structured perceptron model is the highest at 97.9%. Compared to word segmentation, its F1 value is not significantly different, but it surpasses word segmentation in terms of accuracy and recall. Due to the relatively single corpus tested, the three scores for named entity recognition are generally low.

Regarding the speech recognition process, considering that the sentences involved in human-computer interaction design in the BCC corpus are relatively fixed to be closer to life and simulate real-life communication more realistically, this study used the CTB8.0 corpus for model training and comparison. Using P value, R value, and F1 value as reference indicators, comparative tests are conducted on the probabilistic context free grammar (RCFG), semantic feature analysis (SFA), and transfer dependency syntactic analysis models. The transfer dependency parsing model is represented by TDSA, and the test results are shown in Fig. 9.

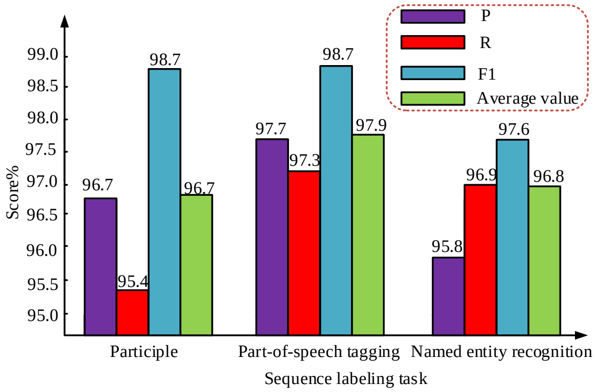
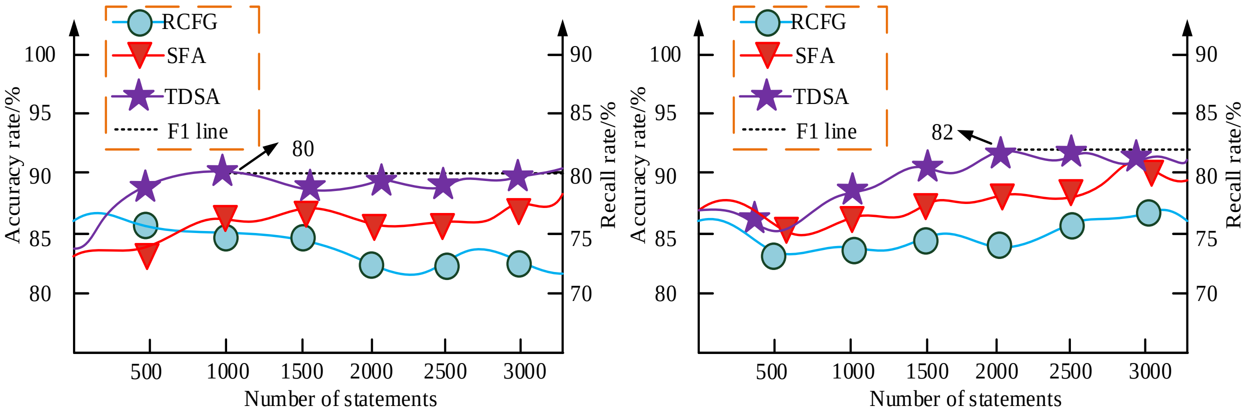


Fig. 8. Sequence annotation results of structured perceptron model.



(a) Training set

(b) Test set

Fig. 9. Test results of different syntactic parsing models.

Fig. 9(a) and Fig. 9(b) show the test results of three syntactic parsing models in the training and testing sets. According to Fig. 9, in a specific corpus training environment, the accuracy of the research model is close to 90%, and compared to other models, this method has the best training results. In the test set results, the maximum P-value of the transfer dependency parsing model is close to 93%, the maximum R-value is close to 83%, and the F1 value is close to 82%.

To delve deeper into the actual performance of syntactic analysis of these three types of models, accuracy, analysis speed, and running memory are used as reference indicators, and the CTB8.0 corpus continues to be used as the test object. Table I shows the specific test results.

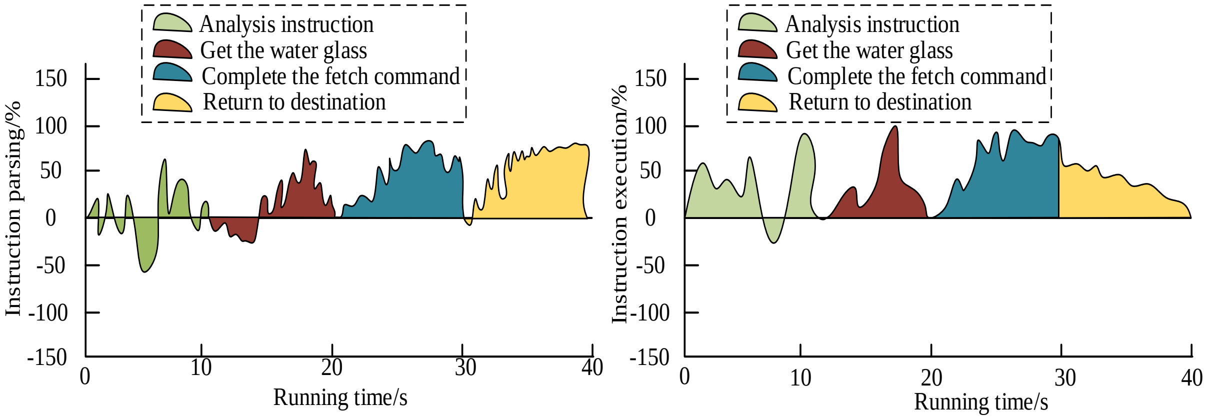
In Table I, among the three syntactic analysis models, the RCFG model has low accuracy, analysis speed, and running memory. Compared to RCFG, the accuracy and analysis speed of SFA have been improved. However, its running memory is small and not suitable for statement analysis in larger information environments. The accuracy of the research model can reach up to 96.77%, with a data analysis speed of up to 750K/s and a running memory of 126M. This can indicate that the transfer dependency syntactic analysis model proposed in this study has good practical application performance.

B. *Simulation Performance Testing of Robot Human-Machine Interaction Model*

Combining, Fig. 5, Eq. (13), and (14) for the design of the human-machine part of the robot, a human-machine interaction system using speech recognition has been built this time. It is tested using a universal Chinese textbook corpus and simulated with simple texts from daily family life. This study takes instruction parsing and execution as reference indicators, and uses the text "Give me a water cup" as the initial instruction for analysis and testing. A positive score indicates correct analysis and execution of instructions, while a negative score indicates error analysis and execution of instructions. The simulation results are shown in Fig. 10.

TABLE I. ACTUAL TEST RESULTS OF THREE SYNTACTIC ANALYSIS MODELS

Task		Accuracy/%	Speed/K/s	Memory/M
RCFG	Training set	94.31	580.00	77.00
	Test set	96.77	640.00	82.00
SFA	Training set	95.69	620.00	45.00
	Test set	96.83	650.00	68.00
TDSA	Training set	96.31	710.00	118.00
	Test set	97.28	750.00	126.00



（a）Robot command parsing situation

（b）Robot command execution

Fig. 10. Robot human-computer interaction instruction execution.

Fig. 10(a) and Fig. 10(b) show the analysis accuracy and execution accuracy curves of the "Give Me a Water Cup" command executed by the human robot. In Fig. 10, the robot first performs command analysis after receiving voice commands. After disassembling the analysis results, it goes to retrieve the water cup. After completing the retrieval of the water cup, it returns to its destination. Due to the complex environment in the home environment and the presence of various interference factors in robot analysis, error analysis occurred in the instruction analysis stage for approximately two seconds. At the same time, within 10 seconds, an error action with a negative execution rate occurred. But overall, the performance during the subsequent 30 seconds of picking up the water cup and returning was relatively satisfactory.

In order to further explore the practical application performance of the robot human-machine interaction method, execution speed, execution accuracy, sensitivity, and robustness are used as reference indicators this time. This study discusses the proposed human-machine interaction method and compares it with existing robot little i human-machine interaction systems, robot Echo human-machine interaction systems, and robot Hanna human-machine interaction systems on the market. Table II shows the experimental results.

TABLE II. EXECUTION DATA OF DIFFERENT HUMAN-COMPUTER INTERACTION SYSTEMS

Human-computer interaction method	Execution time /s	Execution accuracy rate /%	Response sensitivity /%	Robustness /%
Robot little i	32	67	81	69
Robot Echo	52	88	84	54
Robot Hanna	45	74	76	83
The method proposed in this study	35	92	86	92

In Table II, the human-machine interaction execution time of little i robot is the shortest, and the execution accuracy of Echo robot is the highest. Based on the above data, it is found that the proposed human-machine interaction method combining structured perceptron and transfer dependency syntax analysis has much higher robustness and response sensitivity than other systems. Its accuracy is 92%, indicating that the combination of lexical analysis and syntactic analysis human-machine interaction method has the best execution effect in a certain speech recognition environment.

V. CONCLUSION

In order to further improve the accuracy and effectiveness of robot human-machine interaction systems, this study proposed a new method based on traditional speech recognition interaction systems. It combined a structured perceptron model and a transfer dependency syntactic analysis model for a new type of human-computer interaction. The experiment showed that the testing accuracy of the structured perceptron model in this method was as high as 95%, the recall rate was as high as 81%, and the F1 value was as high as 82%. In the testing of the transfer dependency syntactic analysis model in speech recognition, the maximum P value was close to 93%, the R value was close to 83%, and the F1 value was close to 82%. At the same time, the data analysis speed was up to 750K/s and the running memory was 126M. In the simulation testing experiment of the proposed new human-computer interaction method, although there were brief erroneous data analysis, the overall task execution rate was high. Compared to other robot human-machine interaction systems on the market, the accuracy of this method could reach 92%, and its robustness and response sensitivity were excellent. It can be seen from the above test results that, compared with the same type of language recognition interaction models, the new HCI model proposed by the study, which combines the structured perceptual machine model and the transfer dependency syntactic analysis model, is more adaptable to complex and randomly changing interaction scenarios, and the advantages of the model of this method are not only manifested in the aspects of very high accuracy, recall and F1 value, but also has an absolute leading advantage in the analysis speed and running memory. The model of this method not only shows its advantages in terms of very high accuracy, recall and F1 value, but also has absolute leading advantages in analysis speed and operation memory. Therefore, it can be said that the proposed model in the study is in the leading position in all the indexes, and can bring great impetus to the field of speech recognition human-computer interaction. In summary, the proposed new human-computer interaction method could timely and accurately respond to user instructions after correctly recognizing them, and could automatically detect and avoid erroneous voice data analysis. In addition, the high accuracy and efficiency data provided by the current study can be used as a benchmark to help future researchers optimize existing models. By analyzing and understanding the advantages of structured perceptual machines and transfer-dependent syntax, future research can build on these success factors to further improve the model. In a large number of corpus datasets, the test results of lexical analysis and syntactic analysis performed better. However, this study only focused on optimizing and improving the field of speech recognition, and had not yet introduced image analysis technology in human-computer interaction. Further research can be conducted on this basis, combined with image recognition technology, for in-depth exploration. This research provides new insights into robotic human-robot interaction systems in the field of natural language processing and speech recognition, and the results pave the way for future research. Future research can build on the current high accuracy and efficiency data for further model optimization, and delve into the causes of error data to reduce the occurrence of errors. In addition, explorations incorporating image recognition technology will open up a more comprehensive human-computer interaction experience. Cross-domain applications, such as healthcare, education or customer service, are also important directions for future research. Meanwhile, the ability to adapt to different cultural and linguistic environments, improve user experience and interaction design, and test the system's durability and application performance in long-term and real-world environments are all areas of interest. Finally, as technology evolves, research on ethical, legal, and privacy issues is indispensable to ensure the safe and responsible use of technology. Through these multiple perspectives, we are able to advance not only on the technological level, but also on the application, ethical, and legal dimensions that drive the overall development of the field.

ACKNOWLEDGMENT

The research is supported by: 2021 “Innovation” Project Support Plan for Teachers of Harbin University of Commerce. (No.20YYB020) (No.20YYC145); 2022 “Innovation” Project Support Plan for Teachers of Harbin University of Commerce. (No.22YYB252); Ministry of Education Humanities and Social Sciences Youth Foundation for Research Funds: A Study on Discourse Semantic Representation System and Resource Construction for Language Intelligence (No.23YJCZH150).

REFERENCES

[1] Qiu S, Liu Q, Zhou S, Huang W. Adversarial attack and defense technologies in natural language processing: A survey. Neurocomputing, 2022, 492(1):278-307.

[2] Peer D, Stabinger S, Engl S, Rodriguez-Sanchez A. Greedy-layer pruning: Speeding up transformer models for natural language processing. Pattern recognition letters, 2022, 157(5):76-82.

[3] Shi J, Hurdle J F, Johnson S A, Ferraro J P, Skarda D E, Finlayson S G, Samore M H, Bucher B T. Natural language processing for the surveillance of postoperative venous thromboembolism. Surgery, 2021, 170(4):1175-1182.

[4] Li Z, Ming Y, Yang L, Xue J H. Mutual-learning sequence-level knowledge distillation for automatic speech recognition. Neurocomputing, 2021, 428(7):259-267.

[5] Freire-Obregon D, Rosales-Santana K, Marin-Reyes P A Penate-Sanchez A, Lorenzo-Navarro J, Castrillon-Santana M. Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. Pattern recognition letters, 2021, 149(9):179-184.

[6] Ko W R, Jang M, Lee J, Kim J. AIR-Act2Act: Human–human interaction dataset for teaching non-verbal social behaviors to robots:. The International Journal of Robotics Research, 2021, 40(4-5):691-697.

[7] Kim H K, Jeong H, Park J, Kim W, Kim N, Park S, Park N. Development of a Comprehensive Design Guideline to Evaluate the User Experiences of Meal-Assistance Robots considering Human-Machine Social Interactions. International journal of human-computer interaction, 2022. 38(16):1687-1700.

[8] Roda-Sanchez L, Olivares T, Garrido-Hidalgo C, Luis de la Vara J, Fernandez-Caballero A. Human-robot interaction in industry 4.0 based on an internet of things real-time gesture control system. Integrated Computer-Aided Engineering, 2021, 28(2):159-175.

[9] Alsayadi H A, Abdelhamid A A, Hegazy I, Fayed Z T. Arabic speech recognition using end-toned deep learning. IET Signal Processing, 2021, 15(8):521-534.

[10] Lin Y, Yang B, Guo D, Fan P. Towards multilingual end-to-end speech recognition for air traffic control. IET intelligent transport systems, 2021, 15(9):1203-1214.

[11] Dong Y, Yang X. Affect-salient event sequence modelling for continuous speech emotion recognition. Neurocomputing, 2021, 458(11):246-258.

[12] Yerigeri V V, Ragha L K. Speech stress recognition using semi-eager learning. Cognitive Systems Research, 2021, 65(3):79-97.

[13] Bitterman D S, Miller T A, Mak R H, Savova G K. Clinical Natural Language Processing for Radiation Oncology: A Review and Practical Primer. International Journal of Radiation Oncology Biology Physics, 2021, 110(3):641-655.

[14] Ocquaye E N N, Mao Q, Xue Y, Song H. Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. International Journal of Intelligent Systems, 2021, 36(1):53-71.

[15] Hidayat I, Ali M Z, Arshad A. Machine Learning-Based Intrusion Detection System: An Experimental Comparison. Journal of Computational and Cognitive Engineering, 2022, 2(2):88-97.

[16] Yang B, Wang L, Wong D F. Context-Aware Self-Attention Networks for Natural Language Processing. Neurocomputing, 2021, 458(10):157-169.

[17] Mustafa H A, Al-Wesabi F N, Abdelzahir A. A Hybrid Intelligent Text Watermarking and Natural Language Processing Approach for Transferring and Receiving an Authentic English Text Via Internet. The Computer Journal, 2021,65 (2):423-435.

[18] Perboli G, Gajetti M, Fedorov S. Natural Language Processing for the identification of Human factors in aviation accidents causes: An application to the SHEL methodology. Expert Systems with Applications, 2021, 186(7):115694-115695.

[19] Jeon J H, Xu X, Zhang Y. Extraction of Construction Quality Requirements from Textual Specifications via Natural Language Processing. Transportation Research Record, 2021, 2675(9):222-237.

[20] Le T, Huang D, Apthorpe N J. SkillBot: Identifying Risky Content for Children in Alexa Skills. ACM Transactions on Internet Technology (TOIT), 2022, 22(3):79-110.

[21] Chen X, Zhang F, Zhou F, Marcello B. Multi-scale graph capsule with influence attention for information cascades prediction. International Journal of Intelligent Systems, 2022, 37(3):2584-2611.

[22] De Lope J, Grana M. An ongoing review of speech emotion recognition. Neurocomputing, 2023, 528(4):1-11.

[23] Rosenbaum T, Cohen I, Winebrand E. Differentiable Mean Opinion Score Regularization for Perceptual Speech Enhancement. Pattern recognition letters, 2023, 166(2):159-163.