Application of AI conversation agent in new frameworks and fields and improvement in sensory aspects

Yaxuan Liang[1,4,7], **Yun Ye**[2,5], **Enhao Zhu**[3,6]

[1]Department of education, Huaibei Normal University, Huaibei,235065, China
[2]Department of computing, Macau University of Science and Technology, Macao, 999078, China
[3]Senior high school, Zhixin International, Canton, 510105, China

[4]alexa_liangyaxuan@163.com
[5]1210022226@student.must.edu.mo
[6]ann759839@gmail.com
[7]corresponding author

Abstract. Taking a big step with the development of AI, the functions of conversation agents are becoming increasingly mature, and people's lives are becoming increasingly dependent on the conversation agent system. Its assistance to people is reflected in many fields. This has sparked people's exploration of some emerging fields, it is necessary to organize and summarize the new technologies, functions, data, and methods for training new data required for studying emerging fields. More and more deep learning technology have been applied to conversation agents that improve the quality of the service significantly. It is still in the initial stage, and needs improvement both at modeling methods and datasets gathering.

Keywords: AI, Anthropomorphic conversational agent, dataset, HCI model.

1. Introduction

First we identify the key components in every AI conversation agent. There are four technical modules required for an AI conversation agent, namely: Automatic Speech Recognition (ASR) Natural Language Understanding (NLU) Natural Language Generation (NLG) Text to Speech (TTS).

The core of the interaction process is intention, after finding the intention, Semantic Analysis is an important step. Then send them to cloud interaction. Context Intention Processing and do the dialogue management, that can help with organizational language.

As the most eye-catching new technology with the most potential for development in the new era, artificial intelligence contains an important driving force for a new round of scientific and technological revolution and industrial change, and its theory and technology are increasingly mature, with constantly expanding application fields. It is also endowed with the expectation of surpassing human thinking. Conversational agent is an important part of artificial intelligence application, which has a wide market and novel content. The scope of application includes but is not limited to: 3D field, character imitation field, crime field, education field and so on. Putting conversational agents into daily use will not only greatly facilitate people's lives, but also push humanity towards a new level of artificial intelligence, inject new vitality into the development of science and technology, and inspire young people's creative motivation and new expectations for the future.

In order to accelerate the development of conversational agent technology as well as give full play to the advantages of conversational agent and make it get proper publicity, we need a literature review of the current research in the field of conversational agent. The literature review plays an important role in summarizing the remarkable achievements of artificial intelligence conversational agents which need a deep understanding of its application model and algorithm so the literature review can apply them to more research and inventions in the future. In addition, significant defects in the field of conversational agent research should also be found in the literature. Point out the research direction for future scholars is also important in the literature. With the help of this literature, the conversational agent has infinite development potential and is likely to become a great scientific and technological progress in human history.

For studying conversational agents and solving existing problems, we searched for papers on customized response systems, emotional perception, the combination of rational and emotional intelligence, and the use of machine learning to enhance the embodiment of perception according to the application of conversational agents in various aspects. These papers highly analyze the closeness of human and conversational agent systems, and show the relationship between changes in system patterns and different human behavior. We analyze the difficulties encountered in the paper and the different solutions which were proposed by the authors of the paper. Following the logic of the paper, we first summarize how they used different high-precision, intelligent scientific algorithms. Then, we begin to describe the methods they used to collect valid data intelligently as well as outline the main contents of the data. Finally, the authors build different models to solve the problem. We briefly describe the basic concept of the model and explain its usage in the paper. We summarize these processes and quote them in our research. In our summary, the studied area of conversational agent coverage is still insufficient, and there are many unknowns waiting for us to explore. First of all, the summary of the algorithm is slightly abstract, which is not easy for readers to understand and make use of it. It needs to be explained more closely to the normal language and illustrated with examples. What’s more, there is a lack of more specific description of data collection in the summary. We need more in-depth research on the effectiveness and singleness of data collection. At last, the contribution of the session agent system to this field is not clearly described in part of the summary, which does not grasp the key point of the paper.

The following structure of the paper are as follow:

Firstly, we divide the literature content related to conversational agents into several important parts, which are as follows: The steps of building a session agent system, the typical problems encountered in the process and the corresponding solutions, explain the classical model and deep learning method used in building the system, the summary of datasets and the advantages and disadvantages of application, and finally show the use of session agents in some major fields.

Then we analyze the differences between the conversational agent systems demonstrated by various research organizations and expectations and reality, and summarize them in several directions. First of all, some of the models in the system are still not mature enough especially to build anthropomorphic models. Secondly, it is about the ethics and security of session agents. Finally, there are some concentration problems encountered during the development of session agent applications in specialized fields.

The last part of the article is a summary of our discussion of conversational agents and the future prospects of this field.

2. Literature review

2.1. Procedure for developing conversational agent

In the process of building the conversational agent system, almost all the research teams follow the three steps, which means from generating question and answer text to synthesizing audio and finally realizing virtual character. Of course, the third step may be different due to the final shape of the session agent implemented. This pattern requires strict adherence because the implementation of the following steps depend entirely on the previous results which play a role as input.

In the first stage of the system, researchers typically use models and algorithms like the Azure Custom Answering Service to train the conversational agent system to answer questions and then provide large data sets for the system to learn autonomously to improve text quality [1]. This stage is the key part of the whole system, and the text content generated by it is the reflection of the system to the user, and its accuracy is the standard of the quality of the system. What's more, with the addition of CNN and some tools which effective analysis and understanding of context such as Formal Concept Analysis and CORK model, make the system more human and respond after understanding context and emotion that let the system get a further upgrade.

Next stage is speech model which is the easiest steps in the whole system. Almost all conversational agents in this part tend to use either text-to-speech model or automated audio synthesis methods. In the research that I know of, Tortoise TTS is a system that takes text as input and generates an audio signal from it which is a good way to get a precise voice [2]. At the same time, some research groups want to recreate real human voices like Albert Einstein, cloning technology and databases which include various of videos and recordings can help the speech model to simulate human voices. The completion of this model can make the session agent not only communicate directly with the chat box, but also add the function of voice communication, and further realize the personification of the conversation agent.

The final step is to create the digital character. This step requires the use of various technologies related to neural networks and face capture, the main purpose of which is to use the input audio to automatically generate 3D avatars with synchronized facial expressions and mouth movements. To make virtual characters behave more like real people, VOCA model and Temporal GAN will provide effective help which synchronizes the mouth movements with the audio and new technology is used to capture the expressions of people in recorded videos to drive the transformation of virtual faces.

2.1.1. Making conversational agent more human like

In different papers, researchers have encountered some difficulties when constructing the conversational agent system. These difficulties are specific and abstract depending on the application field of the system, and solutions which varied and multifaceted are also provided.

In solving the problem of perception-embedded dialogue agent, the researcher Dolça Tellols also uses an effective and appropriate method. In order to make dialogue agents have human-like perceptions and make them respond to human feelings, researchers propose the following solutions. First, they introduced the SECA [3] database and redesigned the personality, requirements, and dialogue modules. Next, new knowledge, memory, empathy and NLP modules are created to refine the structure. Then, NLP is preprocessed with further operational analysis and user-friendly text entry of the standardized text. More importantly, researcher Dolça Tellols integrated our SECA as a virtual tutor into the child's application for verification, using scientific methods to test the participants' satisfaction with the system. In the end, they got a very satisfactory result, the SECA being endowed with surrogate human characteristics such as personality, needs, and empathy, which increases user connections [3].

2.1.2. Preservation of voice signal

While studying the implementation of the system as a conversational agent for virtual characters, the researcher Nitin Sakhare found it difficult to clone good voices [1]. The cloning of speech is a challenging downstream field in TTS that aims to guide the generation of speech and improve the fit between its audio and reference speech. To address these issues, researchers have used different models, such as the Turtle TTS model, CVVP model, and CLVP model. Firstly, the TTS model is used to clone the speech of digital characters, while the CLVP model plays an important role in evaluating the speech closest to the character. Afterwards, Tortoise provided a judgment method for extracting and inferring speech sounds such as pitch and pitch, resulting in a sharp reduction in the search space for possible speech output information about a given text. Researchers trained the model by a contrastive language-speech pertaining converter (CLVP) and paired discrete speech tags with text tags to reorder multiple AR outputs [2]. Finally provide the database like videos for the generator to learn, let it generate the sound more relevant you want to achieve with constant training. With the help of this method, the researchers succeeded in matching the system's facial expressions, voices and historical figures.

2.1.3. Making the system more transparent

It is very important for the system to be transparent, especially in the criminal arena. In order for conversational agents to be successfully used in the criminal arena, we need to improve the transparency of the system, make it easier for people to understand how it collects data, and strengthen trust in human systems. The researchers used formal concept analysis (FCA) when building the conversational agent platform to define the concepts that it responds to different intents, and then combined RPD model attributes to build functional models, thereby increasing vulnerability and enabling data visualization [4]. Finally, the researchers evaluated model use scenarios through surveys of analysts and their practices, which greatly improved the possibility of conversational agents in case resolution [5-6].

2.2. Key Techniques & Classical Models

In order to achieve a more accurate and perfect advanced conversational agent, some deep learning techniques are used to solve some problems of the system. CNN is a supervised learning method based on deep neural networks, by building a wide and deep neural network, using the temporal locality principle of audio signal partitioning, to determine the user's mood based on the audio characteristics [2]. Time GAN relies on frames, sequences, and synchronous discriminators, with automatic lips driven by three types of machines, and then synchronized speech head generation, which depends on the input audio content [1].

The following contents are the classical models that can be adopted in the system construction process, which have certain representativeness and have some advantages over ordinary models while realizing basic functions.

The CORK model built by a research team for the first phase of the system, takes a client-server software architecture exploiting the Model-View-Controller MVC architectural pattern to divide the system into three logical layers, so it separates the internal representation of information from the way it is presented and presented to the user which is different from some models [2].On this basis, the modular framework is used to subdivide the system into ten independent modules, each includes only a small portion of the functionality refined. This is beneficial for rapid system maintenance.

The speech step needs the model to exchange text to speech. The Tortoise is a common text-to-speech model often used in this regard which has several subfields related to sound synthesis and the probability of application is quite high. Also, some teams will use extra modes like MOS, CLVP and CVVP to evaluate and select the best audio.

The third step requires models with multiple functions to generate expressions, mouth movements and other gestures of virtual characters respectively. The FLAME model is a classic model for this part, which generates both realistic and custom expressions [1], and techniques such as the VOCA and GAN methods mentioned above help the system improve the recognition accuracy of facial expressions.

2.3. Analysis of Datasets

Table 1. Analysis of Datasets

	Data type	content	Advantage	Disadvantage	purpose
RPD	character	inputting the person, time, cause, location, and their necessary relationships.	It is in textual form, the difficulty of extracting information is relatively low.	The content that needs to be chatted with needs to have the information and signals searched for before corresponding.	Retrieving and linking information based on past experience in similar areas, and further advancing the investigation.
Tortoise TTS	audio	collected audio clips of the historical figures from various sources,like movies, speeches, and interviews	It is commonly used to test the accurate response and improvement ability of the system.	The content that needs to have the information and signals searched for before corresponding, which has certain requirements for the quality of the content.	It should be ensured that the sound generated by the Turtle TTS model does not have a high degree of similarity to the actual sound of historical figures, due to the consideration of model training
FLAME	image	A dataset of 3800 3D human head scans This shape space captures the variations in legal shapes.	The dataset can make the model more three dimensional without deviation in the subsequence process.	Large database volume, training model has certain difficulty and cost.	To recognize the facial expression of the user, to improve customer satisfaction.
Emovo	harmonic features	feature extraction and Classification into emotions	Because it is in textual form, the difficulty of extracting information is relatively low.	There may be deviations in the recognition of sound signals.	To train the emotional recognition model.

The first one is Identify the initiation decision RPD, according to a paper. It transfers specific input details for confirmation and further information search based on the search table [4].

Each process or function can be defined as a separate attractor of a related type, inputting the person, time, cause, location, and their necessary relationships, retrieving and linking information based on past experience in similar areas, and further advancing the investigation. For another scenario, analysts can access and visualize data in the same way, however, they are also encouraged to enter the triggered intent to check and validate the functional process. The focus of the study is to understand the provision of system transparency. The method of use is to first ask a series of questions, and respond to both data and system processes in response to a specific situation.

According to a study, Tortoise TTS model dataset: To create the dataset used to train the Tortoise TTS model, we collected audio clips of the historical figures from various sources, including movies, speeches, and interviews. The carefully curated dataset of audio clips that we used to train the model ensured that the voices generated by the Tortoise TTS model were very similar to the actual voices of the historical figures.

Some 3D datasets are created to train FLAME. What is FLAME? Some algorithm even trained to recognize the facial expression of the user, to improve customer satisfaction. The dataset for this purpose are: FLAME. The linear shape space in FLAME is constructed from a dataset of 3800 3D human head scans, this shape space captures the variations in legal shapes [1].

Another study shows that,to train the emotional recognition model ,The dataset from the harmonic features of the audio is performed by an original machine learning model all by us. The process is organized in two subsequence steps:

(1) feature extraction

(2) Classification into emotions

The form extracts temporary and spectral characteristics of the audio signal; Where as, the classification has been implemented using a supervised learning approach based on deep neural networks In granular, a wide and deep (convolutional) neural network has been built to explore the principal of temporary locality across audio signal partitions and increase the discriminatory strength of the model In order to properly train the model an open source and free Italian dataset called Emovo [2].

2.4. Application area for conversational agent

At present, people's expectations for conversational agents are rising, with the hope of using technology in various fields in the future.

Conversational agent systems can be applied to medical scenarios to alleviate alexithymia and its effects, and to strengthen the caregivers' emotional awareness of their own feelings [7]. ECA can be used in high-risk, high-reward mental health services. Part of the system can identify stress, anxiety and depression in children who have suffered trauma. For example, Nora and ECA, since Nora is used for trauma-informed care, it needs to recognize the signs of trauma and reduce the situation through emotional regulation techniques. The technology could also be used in education to create engaging and lifelike digital characters that can answer questions visually. The contribution of the system is to create more natural and engaging ways for users to interact with digital assistants or educational tools[8]. It can also assess the level of student participation. Such a system should contribute to student success as well as the design, delivery and evaluation of courses. Similarly, consumer brands can monitor customer satisfaction and brand recognition by analyzing conversations about the brand on social media. In terms of entertainment, AI can identify the artwork that visitors appreciate as they walk through the museum and create a conversation around the artifacts [9]. One could build a customer service chat robot that analyzes chat text to detect a customer's happiness or frustration. Consumer sentiment determines voter preferences, stock market forecasts and movie reviews. Finally, conversational agents are also used in Criminal Investigations [4]. By transcoding clues, investigators can effectively identify the focus of the case and correct the investigation strategy, thus improving the efficiency of solving the case.

3. Analysis

3.1. Current model is far from human like agent

Due to the limitations of technology, time and resources, some models of session agent systems are still immature. Among them, the problems about the construction of anthropomorphic models are more concentrated.

The basic model of the conversational agent (the conversational model) relies too much on sets and rules, which means that the system answers in a rigid way and does not respond correctly to abnormal input and questions from users as humans do. To solve that can simulate the communication mode between people in reality to re-establish the model, and introduce a huge database for the system to carry out machine learning and promote the iterative upgrade then the performance of the system has been greatly improved [2].

The other model is for generating virtual characters. To make human like agent, 3D portrait model is essential, however, current model used in creating digital character of the system is still at a relatively elementary level, so the generated animation or 3D model is rough which shows that the expression of virtual characters is insufficient and rigid. In the face of this dilemma, we can flexibly use more relevant models to customize more ways of expression and constantly train the model to make each part of behave more smoothly.

3.2. Ethics and privacy issues of conversation agent

The use of sensory conversation agents in some fields is limited.

Ethically, emotion is a very personal feeling that users may not want to disclose. Therefore, users should be able to choose not to conduct any emotion analysis.

In terms of methodology, the system should be able to distinguish between right and wrong (but isn't this subjective?), legal and illegal (varies from country to country), moral and immoral (also subjective) [10].

For example, if the user asks the agent about the method of suicide, the agent should give up the answer. Instead, it should assess users' stress and seek help when needed.In these studies, there is no research on the dataset that makes the sensory conversation agent have the concept of ethical right and wrong. If an AI system can access the hidden state of users' emotions, users may feel vulnerable and betrayed. This invasion of privacy will violate the norms of trust and make the system unattractive. In addition, when the agent deletes all collected data at the end of user system interaction, the user can exercise the right of being forgotten.

Since right and wrong, morality and immorality are subjective, we should first establish the prototype of the dataset with the law and recognize public order and good customs as the controller, then improve it, and then iterate on the agent. When the user is touched with a hidden private emotional state, set a prompt to prompt the conversation agent to stop the conversation.

3.3. Lack of awareness about the use of AI in crime

The use of conversational agents in the criminal field is limited. At present, it only helps analysts to find the focus of the case, and in the future, more functions of conversational agents should be studied in this field. In addition, the use and promotion of conversational agents in the field of crime is too low, and many people do not even know that artificial intelligence has this function, and the popularity and use are insufficient [4].

3.4. Application focused data are hard to get

The models and methods used in the process of building a conversational agent like Pan for use in the criminal field have highly human involvement; it cost lots of money and time, which is very adverse to research groups with insufficient budgets. Insufficient data sets for training conversational agents applied to the criminal field, too few investigators led to one-sided results. In the future, more intelligent models and methods should be created to solve this problem.

3.5. Limited use in education

At present, the conversational agent system can only play an auxiliary role in education, but cannot completely replace teachers. The teaching method of AI is rigid, they are not very mature to make personalized plans for students and the opinions provided to students are not targeted as well.

3.6. The risk of aggravating the disease by AI

While using AI to treat human beings, there is also a risk of aggravating the disease. With reference to many cases where patients' illness is aggravated by exposure to the Internet, how to better develop the conversational agent treatment system is also an issue that we need to be concerned about.

4. Conclusion

In general, the conversational agent systems we know have certain similarities in the construction process, which means that the model development of today's conversational agent has been relatively mature, and is developing towards more detailed and special function realization, as shown in this paper about the conversational agent applied in the professional field and the conversational system with quasi-human nature after refinement. These summaries of the model structure can provide ideas for beginners who want to develop session agents and get started faster.

In order to give you a better understanding of the development of conversational agents, this paper uses a lot of space to explain the recent good technology and typical models, typical problems and current bottlenecks. These include the results of current research developments where the proposed methods can be directly adopted for example: the TTS system is a commercial tool for anyone, and the FLAME model can be used directly. In addition, for the summary of the difficulties that the session agent still cannot solve, on the one hand, this study want to let everyone understand that the structure of the session agent system and the application function realization is not as simple as imagined, and there are many details under the general model, on the other hand, this study also encourage the later research team to focus on the realization of a more accurate intelligent system, and help the development of the field.

In any case, the development prospects of conversational agents are huge, whether it is everyday intelligent conversational systems or some professional applications have plenty of room for progress.

Finally, this paper would like to talk about the content of this article. In the process of writing and revising the paper, we gradually realized my lack of knowledge and experience. The amount of literature used in this paper may be very limited, and it cannot bring you a deeper and broader discussion, hoping you can understand. Thanks to all those who have read this article, hoping this article can provide you with some help.

Acknowledgement
Yaxuan Liang, Yun Ye, and Enhao Zhu contributed equally to this work and should be considered co-first authors.

References

[1] Sakhare, N., Bangare, J., Ajalkar, D., Walunjkar, G., Borawake, M., & Ingle, A. (2023). Intelligent Conversational Agents Based Custom Question Answering System. International Journal of Intelligent Systems and Applications in Engineering, 11(6s), 337-344.

[2] Catania, F., Spitale, M., Fisicaro, D., & Garzotto, F. (2019). CORK: A COnversational agent framewoRK exploiting both rational and emotional intelligence. In IUI'19: Proceedings of the International Conference on Intelligent User Interfaces (pp. 1-8).

[3] D. Tellols, M. López-Sanchez, I. Rodríguez, P. Almajano, Sentient embodied conversational agents: architecture and evaluation, Artif. Intell. Res. Dev. 308 (2018) 312.

[4] Hepenstal, S., Zhang, L., Kodagoda, N., & Wong, B. L. W. (2021). Developing conversational agents for use in criminal investigations. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1-35.

[5] Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., ... & Cowan, B. R. (2019, May). What makes a good conversation? Challenges in designing truly conversational agents. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-12).

[6] Nass, C., Isbister, K., & Lee, E. J. (2000). Truth is beauty: Researching embodied conversational agents. Embodied conversational agents, 2000, 374-402.

[7] J. Feine, S. Morana, U. Gnewuch, Measuring service encounter satisfaction with customer service chatbots using sentiment analysis, in: 14th Internationale Tagung Wirtschaftsinformatik (WI2019), 2019, p. 1115.

[8] Kongthon, A., Sangkeettrakarn, C., Kongyoung, S., & Haruechaiyasak, C. (2009). Implementing an online help desk system based on conversational agent Authors. Published by ACM 2009 Article, Bibliometrics Data Bibliometrics. Published in: Proceeding, MEDES'09 Proceedings of the International Conference on Management of Emergent Digital EcoSystems, ACM New York, NY, USA.

[9] Ghosh, S. (2023). Sentiment-aware design of human–computer interactions: How research in human–computer interaction and sentiment analysis can lead to more user-centered systems?. In *Computational Intelligence Applications for Text and Sentiment Data Analysis* (pp. 209-224). Academic Press.

[10] Tellols, D., Lopez-Sanchez, M., Rodríguez, I., Almajano, P., & Puig, A. (2020). Enhancing sentient embodied conversational agents with machine learning. *Pattern Recognition Letters*, *129*, 317-323.