Deep Learning-driven Enhancement of Chatbot Interaction: A Comprehensive Study on ChatGLM

ZIJIAN ZENG, KURUNATHAN RATNAVELU*
Institute of Computer Science and Digital Innovation, UCSI University,
WP Kuala Lumpur 56000,
MALAYSIA

*Corresponding Author

Abstract: - In the contemporary digital landscape, ChatGLM, powered by advanced artificial intelligence, has risen as a tour de force, particularly excelling in Chinese Q&A scenarios. Its prominence underscores the transformative role of deep learning neural networks in reshaping the chatbot paradigm. This paper offers a holistic exploration of chatbot model designs, building upon seminal research, and delves into the nuances of chatbot development and underlying technologies. We provide incisive analyses poised to guide future advancements in chatbot-related arenas.

Key-Words: - ChatGLM; chatbot, intelligence, deep learning, neural networks, interaction design, Social Characteristics.

Received: January 21, 2024. Revised: July 4, 2024. Accepted: August 3, 2024. Published: September 4, 2024.

1 Introduction

The dawn of the new millennium witnessed the unprecedented development of Internet computer technology and reached its peak in the rapid rise of artificial intelligence in recent years. It is worth noting that the deep learning model stands out as the cornerstone of the breakthrough in the field of natural language processing and sets a new benchmark for the effectiveness of the application. In recent years, under the impetus of "Industry 4.0 Revolution" and artificial intelligence technology, robot technology has developed rapidly. Robot technology has been widely used in industry, life and military defense, and a series of major breakthroughs have been made. As the main research object of robotics, robots, especially intelligent robots, play an increasingly important role in people's lives and work, and people also put forward higher requirements for the way of human-computer interaction. However, at present, the traditional human-computer interaction methods based on keyboard and mouse, teaching equipment and wearable devices often have problems such as high cost, poor real-time performance and low efficiency. With the intelligent chat robot gradually replacing artificial online interaction on mainstream platforms, we find ourselves sailing in a broad era of artificial intelligence. Gone are the days when primary chatbot content was generated by template matching and retrieval. The paradigm of modern chat robots is manipulated by a dynamic deep learning model, which shows real-time adaptability, [1]. Chat robots, which symbolize the potential of artificial intelligence, are carefully designed to decode and process natural languages. They cultivate unparalleled human-computer interaction through text or dialogue. As digital messengers, chatbots understand various human languages, simulate human-like conversations through text, voice, or mixed mode, and are redefining the outline of human-computer interaction.

2 Overview of Relevant Studies

2.1 Development of Chatbots

For the Q&A system, there are three main types of academic research:

(1) The method of counting the number of words in a sentence. A typical scenario for the utilization of this approach is a knowledge retrieval library system. According to the proximity of the statistical frequency of the question, the frequency of the answer that is relatively close to the question is obtained. This approach does not allow a better understanding of the customer.

(2) Machine Learning. They build Q&A systems using probability and statistics. Like SVM, k-mean algorithm, [2]. Machine learning algorithms compute the eigenvalues of the data samples and

then arrange and combine them into a matrix, which is used as input to the model.

(3) The application of deep learning techniques has led to the reduction of the time taken for the creation and pre-processing of datasets. Moreover, it can also be more efficient and reduce the problem of human error. The data obtained is better than the previous data results. In the case of large datasets, the results achieved using deep learning are even better. The method also has the advantage of being relatively easy to implement as it does not require manual feature extraction. The basic application of deep learning algorithms on the natural language level is CNN, which, with its convergence effect, improves data utilization, reduces wastage, and maximizes the acquisition of the corresponding parameters. The use of deep learning techniques has led to the rapid development of question-and-answer systems in the field of natural language processing, [3]. Chatbot models using RNN technology can be used without predefining a knowledge base and it can generate answers directly. The idea of this approach comes from machine translation techniques in the field of natural language. The Neural Conversation model proposed by Google utilizes four different datasets to train the Seq2Seq model.

2.2 Technical Components of Chatbots

Chatbots have many different categories based on different division criteria. Based on application scenarios, chatbots can be divided into multiple categories of chatbots, such as online customer service, entertainment, education, personal assistant, and intelligent Q&A; based on the technical implementation method, they can be divided into retrieval and generative chatbots; based on the functional positioning, they can be divided into four categories of chatbots, namely, Q&A system, task-based dialog system, idle chat system and active recommendation system, [4].

When analyzing the technical aspects of chatbots, there are five important modules that are part of the system: automatic speech recognition, natural language understanding, conversation management, natural language generation, and speech synthesis. In the process of its operation, it usually follows the following basic flow: the user inputs text or voice, and then preprocesses it into text patterns for natural language understanding. Then, it enters into the dialog management based on the semantic representation and context and extracts the answers of the current dialog model. Finally, the final synthesized reply is output to the user. On the whole, different types of chatbots include the three modules of natural language understanding, natural language generation, and conversation management, but the technical modules favored by different types of chatbots and the technical details used in them are different.

The new generation of chatbots represented by ChatGLM relies on key features such as large-scale language models, naturally flowing conversations, multi-tasking, contextualization, personalization, and open source, and adopts the technological logic of "Big Data + Big Computing Power + Big Algorithms = Intelligent Models", which greatly improves its performance in the areas of natural language interpretation and digital content generation. The use of "big data + big computing power + big algorithm = intelligent model" technology logic, so that its performance in natural language interpretation and digital content generation is greatly improved, but "there are still technical limitations in logical reasoning, reliability, knowledge learning, and risk resistance", [5].

In this paper, the design of a chatbot and its interaction system is constructed on the basis of a deep learning model, in order to improve the performance of the interaction system and provide certain references for the development of chatbots.

3 Chatbot Design

3.1 Sequence-to-sequence Model Seq2Seq

The end-to-end Seq2Seq framework is best applied in the field of machine translation and is also widely used in the fields of text summarization and dialog generation. Therefore, the Seq2Seq framework is applied to a library chatbot to automatically answer user questions and generate coherent and diverse responses. The Seq2Seq framework is as follows:

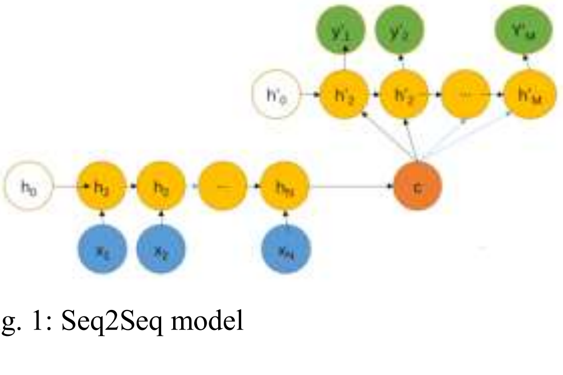

Fig. 1: Seq2Seq model

As shown in Figure 1, the basic principle of the Seq2Seq framework is to translate the input sequence to the output sequence by the deep neural network and consists of an input encoder and output decoder. In dialog generation, the role of encoder and decoder is to generate reply statements, [6]. The user input statement X contains n words, i.e., the reply statement Y generated by the system has m words. The specific expressions of X and Y are as follows.

$$X = \{x_1, x_2, x_3, \ldots, x_n\}$$
$$Y = \{y_1, y_2, y_3, \ldots, y_m\}$$

In the above equation, x_n, y_m refers to the vectors corresponding to the words in the original sentence. The intermediate vector C can represent the state of the last moment output of the $LSTM$ neuron h. After C is nonlinearly transformed, the following can be obtained.

$$\mathrm{h} = \mathrm{z} \odot \tan(\mathrm{C})$$
$$C = f\{x_1, x_2, x_3, \ldots, x_n\}$$

In the formula, the f denotes the nonlinear transformation done to the input sequence.

3.2 Chatbot Structure Design

In order to realize an automatic Q&A system for open-domain deep learning, it is proposed to design an open-domain oriented chatbot, through which the chatbot interacts with the library users in the text to increase the interest and attractiveness of the system. According to the characteristics of the chatbot and the user's question and answer, it is proposed to input the user's statement into the Seq2Seq model based on $LSTM$, and then output the reply statement after matching and analyzing the question and answer, [7]. This realizes the automatic generation of Q&A responses for the chatbot. Figure 2 represents the structure of the chatbot designed in this paper.

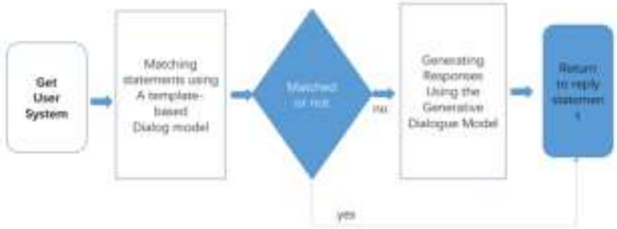
Get User System
Matching statements using A template-based Dialog model
Matched or not
no
Generating Responses Using the Generative Dialogue Model
Return to reply statement
yes

Fig. 2: Chatbot overall structure design

AIML (Artificial Intelligence Markup Language) is used to realize template-based dialog matching model construction. Before replying to the user's question, the chatbot first uses the $AIML$ model to obtain the reply, and if it cannot match the reply structure, it uses the generative model to generate the reply. This method ensures that the chatbot realizes accurate responses to the questions extracted by the user.

3.3 Chatbot with Attention Mechanism and Cluster Search

Among the traditional chatbots designed with the Seq2Seq model and the template dialog model, the decoder is to obtain the input sequence information, it needs to obtain it in the form of intermediate vector states, which makes the decoder generation structure more general. To address this problem, we propose to add bottom-up unconscious attention, called the salient attention mechanism, to the Seq2Seq model. The structural properties of the bottom-up attention mechanism are shown in Figure 3.

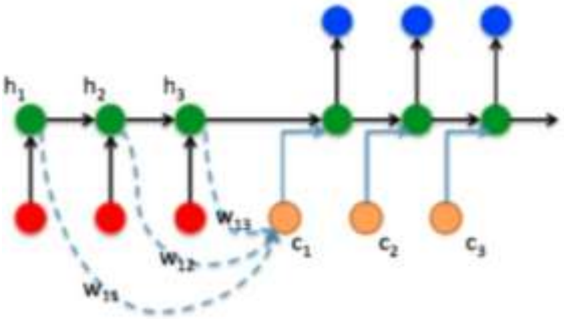

Fig. 3: Schematic diagram of salience attention mechanism

Referring to Figure 3, after the introduction of the salient attention mechanism in the model, the implied state of the output of the previous sequence and the context vector of the current generated word can be obtained. from this, we get the specific computational equation of the decoder at the moment of t is:

$$P(y_t | y_1, \ldots, y_{t-1}, x) = g(y_{t-1}, s_t, c_t)$$

In the formula, y_t and s_t denote the output and implied state of the decoder and neuron at moment t, respectively; $g(\ )$ denotes a series of nonlinear transformations. Among them, C_t has a large influence on the output at moment t. The specific formula for C_t is as follows.

$$C_t = \sum_{j=1}^{T} at_j * h_j$$

In the formula, h_j denotes the implicit state of the encoder neuron in processing the input sequence value at moment j; at_j denotes the degree of influence of the encoder input at moment j on the decoder output at moment t .

The generating sequence vocabulary and the input sequence vocabulary are associated through the saliency attention mechanism, so that the generating vocabulary can pay attention to some of the words in the input sequence, thus increasing the accuracy of the question-and-answer results, [8]. In the chatbot dialog process, there may be inconsistencies in the bot's responses, which reduces the user's chat experience. To address this problem, we propose to add a cluster search algorithm, BeamSearch, to the Seq2Seq model to constrain the information of the robot itself. Cluster analysis is an exploratory analysis method, which can analyze the internal characteristics and laws of things and group things according to the similarity principle. This is a common technology in data mining, [9]. BeamSearch is a heuristic search algorithm that combines the greedy strategy with the exhaustive strategy, which has been widely used in the field of NLP and achieved good results. The search process of the BeamSearch algorithm is shown in Figure 4.

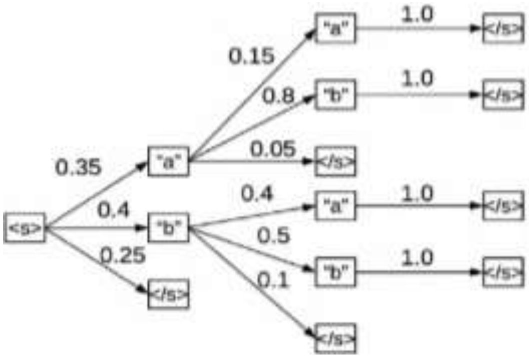

Fig. 4: Search process of Beam Search algorithm

As can be seen from Figure 4, the basic principle of cluster search is to prune the search space when searching. In the Seq2Seq model, adding the beam search algorithm can greatly enhance the diversity of model-generated replies, keep the questions and answers consistent, thus enhancing the user experience of the Q&A effect.

4 Interaction Design Ideas for Chatbots based on Social Characteristics

4.1 Intelligent Dialog

4.1.1 Proactivity

Proactivity indicates that the chatbot system takes conversational actions autonomously instead of the user, which can help the user reduce the energy and time for thinking and replying. In the process of human-robot interaction, proactivity enables the chatbot to share with the user, using natural forms to carry out conversations. For example, chatbots can initiate topics, propose exchanges, provide information, and follow-up comments to actively communicate with users.

4.1.2 Due Diligence

Due diligence is the ability of a chatbot to show dedication to the conversation at hand. Based on the dialog flow setup, the context is scientifically interpreted and each sentence is taken seriously as something that makes sense throughout the conversation.

4.1.3 Communicability

The primary way that users want to achieve their communication goals is by exchanging messages with the system, interactive software is inherently communicable. Its ability of the software to communicate underlying design intent and interaction principles to the user. Providing communicability can help users improve the learnability of the system. The problem with chatbot communicability is the nature of its human-robot interaction. Chatbots take turns demonstrating their functionality through dialog, and the lack of communicability may lead to users abandoning the use of chatbots when they are unable to understand the available features and how to use them.

4.1.4 Natural Interaction

Natural interaction means that the user uses multiple sensory channels and interacts with the interface in a non-mechanical, parallel way. Therefore, natural interaction is mainly characterized by: (1) naturalness: which reduces the learning cost, the user can operate with the help of experience; (2) habituation: which conforms to the user's habits and common sense; (3) non-precision: the user interacts with the interface through rough behavior; (4) high efficiency: the interaction method is more intuitive and improves the interaction efficiency. Combined with the chatbot interaction design described in the article, it can be found that the natural interaction method can significantly improve the user experience, so the integration of natural interaction into the chatbot interaction design can improve the interactivity of chatbots to a certain extent, [10].

4.2 Intelligent Socialization

4.2.1 Damage Control

Damage control specifically refers to the fact that chatbots may encounter conflict problems during human communication and are unable to handle the problem well. Users interacting with chatbots may result in users being harassed. Test the chatbot's abilities and knowledge, and get frustrated with mistakes. When a chatbot fails to respond correctly, it may engage in abusive language behavior, frustrate another user, etc., which ultimately leads to the failure of the conversation, among other things. Therefore, it is necessary to set up damage control for chatbots so that they can recover as quickly as possible, quickly find a way to handle the situation in a way that is acceptable to the user and reduce the harm created by inappropriate handling.

4.2.2 Linguistic Permeability

Linguistic fluency refers to the ability of a chatbot to fluently use various languages and communicate well with users. In a traditional user interface, visual capabilities such as buttons, menus, or links may be used for user communication. The most crucial tool in the process of human-robot interaction is language. Therefore, a chatbot should have the appropriate language style and be able to compute the words that are expected to be answered. Therefore, chatbots should consistently use language that depicts the desired style. When chatbots use inconsistent language or unintended language patterns, it can lead to miscommunication between users and them, affecting their satisfaction.

4.2.3 Social Etiquette

Social etiquette refers to the ability of chatbots to engage in polite behavior and conversational habits during chats. Although politeness may be perceived differently by people with different personalities and cultural backgrounds, Politeness creates goodwill and leads to better relationships. In doing so, chatbots should endeavor to control the rapport during human-robot interactions. Chatbots can behave politely by adopting language such as greetings and apologies, [11].

4.2.4 Moral Constraints

Binding morality refers to the concepts of right and wrong and values that chatbots possess, knowing the laws and regulations, moral codes, etc. that society needs to abide by. Chatbots should have correct values with strong moral constraints, when chatbots have this personalization feature, it can prevent them from reproducing hate speech or abusive speech and increase user goodwill.

4.2.5 Emotion Management

Emotion management can help the robot assess the user's emotions and choose the appropriate response to help the user regulate the emotional response and better solve the user's problems through emotions and so on. Even though chatbots have no real emotions, there are many opinions expressed by users about the advantageous role of chatbots emotional expression. Chatbots with intelligent emotions can recognize and influence the user's feelings, demonstrating respect and empathy, among other things, helps users achieve rapport with chatbots, [12].

4.2.6 Personalization

Personalization mainly focuses on the chatbot system interface and access content and adjusts the relevant content through technology to help the chatbot better understand the characteristics and relevant information of specific users. Personalization is achieved by understanding the context of the chat process and making dynamic adjustments to ensure that it is more adaptable to the user's individual needs and to achieve the goal of chatbot socialization. When chatbots are equipped with personalization features, there are the obvious advantages of (1) enriching interpersonal relationships and increasing anthropomorphic authenticity, and (2) providing users with private and unique services.

4.3 Anthropomorphization

Anthropomorphism refers to the assignment of non-human factors to personal traits, including appearance, emotional state, etc. In the HCI domain, an important way to realize natural interaction is to apply anthropomorphic characters in the user interface. Chatbots should have at least one human-like characteristic.1) Sense of belonging. Belonging specifically refers to an individual's demonstrated ability to exist within a particular social group. While chatbots cannot autonomously choose which trait group to belong to, designers will intentionally or unintentionally, assign identities to them, and when they define the way chatbots talk or behave, the partner's identity (even if it is only perceived) creates new processes that have the effect of anticipating and influencing the outcome of the interaction, [13]. Communicate information about the chatbot's personality, such as name, gender, and age.2) Prediction in terms of personality. Specifically refers to the chatbot's prediction of the user's thoughts, feelings, and behavior. A robot with personality prediction can better communicate with the user and solve the user's problems in a targeted way. If the chat robot has emotional recognition obstacles, it will also make users feel confused and dissatisfied, [14]. Therefore, personality prediction ensures that the chat robot participates in the dialogue behavior that meets the user's expectations in a given scene.

5 Conclusion

In this paper, through an in-depth analysis of the development of chat robot technology and related models, a design model of chat robot is constructed. This paper puts forward the concept of chat robot interaction and holds that chat robots should have the functions of intelligent conversation, intelligent socialization, personification, and interactive technology. In the process of designing the chat robot, the Seq2Seq framework is applied to the library chat robot, which automatically answers the user's questions and produces coherent and diverse answers. At the same time, the user's sentence is input into the Seq2Seq model, which realizes the automatic generation of question-and-answer responses of the chat robot. AIML is also used to realize the construction of dialogue dialogue-matching model based on a template. Ensure that the chat robot can accurately answer the questions extracted by users. As a result, more clear ideas are put forward, such as responsiveness, due diligence, communicability, natural interaction, intelligent socialization, language permeability and so on. Therefore, I hope to provide some references for the research and development of chat robots. However, the deep learning technology applied to chat robots is still in the early stage of development. No matter the technical method or the actual system performance, there is great room for improvement. I hope that in future research work, we can find better and more effective technologies to realize the chat robot with better performance.

References:

[1] Wang J, On the Limit of Machine Intelligence. *International Journal of Intelligence Science*, 2013, pp. 170-175. http://dx.doi.org/10.4236/ijis.2013.34018.

[2] Naveed S, Sajid U, Imrab S, Faiz Ai, Mrim M, Chat-GPT; validating Technology Acceptance Model (TAM) in education sector via ubiquitous learning mechanism, *Computers in Human Behavior*, Vol.154, 2024, pp. 0747-5632. https://doi.org/10.1016/j.chb.2023.108097.

[3] Zalimkhan N, Olga N, Murat A, Kantemir B, Sultan K, The symbol grounding problem in the system of general artificial intelligence based on multi-agent neurocognitive architecture, *Cognitive Systems Research*, Vol.79, 2023, pp. 71-84. https://doi.org/10.1016/j.cogsys.2023.01.002.

[4] Ruchi G, Kiran N, Mahima M, Blend I, Seema, Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda, *International Journal of Information Management Data Insights*, Vol.4, No.1, 2024, pp. 100232. https://doi.org/10.1016/j.jjimei.2024.100232.

[5] Evgeny T, Igor P, Machine learning algorithms for teaching AI chat bots, *Procedia Computer Science*, Vol.190, 2021, pp. 735-744. https://doi.org/10.1016/j.procs.2021.06.086.

[6] William V, Adrián A, Xavier P, Proposal of an Architecture for the Integration of a Chatbot with Artificial Intelligence in a Smart Campus for the Improvement of Learning, *Sustainability*, Vol.12, No.4, 2020, pp. 1500. https://doi.org/10.3390/su12041500.

[7] Maria V, George A. Evangelia-Aikaterini T, VIRTSI: A novel trust dynamics model enhancing Artificial Intelligence collaboration with human users – Insights from a ChatGPT evaluation study, *Information Sciences*, Vol.675, 2024, pp. 0020-0255. https://doi.org/10.1016/j.ins.2024.120759.

[8] Hancock PA, Billings DR, Schaefer K, Chen J, Visser E, A meta-analysis of factors affecting trust in human- robot interaction. *Hum. Factors*, Vol.53, No.5, 2011, pp. 517-527. https://doi.org/10.1177/0018720811417254.

[9] Vallverdú J, Approximate and situated causality in deep learning, *Philosophies*, Vol.5, No.2, 2020, pp. 1-12. DOI: 10.3390/philosophies5010002.

[10] Liao X, Zheng Y, Shi G, Bu H, Automated social presence in artificial-intelligence services: Conceptualization, scale development, and validation, *Technological Forecasting and Social Change*, Vol.203, 2024, PP. 0040-1625. https://doi.org/10.1016/j.techfore.2024.123377.

[11] Amelie A, Soumyadeb C, Sachin K, A shared journey: Experiential perspective and empirical evidence of virtual social robot ChatGPT's priori acceptance, *Technological Forecasting and Social Change*, Vol.201, 2024, pp. 0040-1625.

https://doi.org/10.1016/j.techfore.2023.123202.

[12] Michael D, Brian L, ChatGPT for (Finance) research: The Bananarama Conjecture, *Finance Research Letters*, Vol.53, 2023, pp. 1544-6123. https://doi.org/10.1016/j.frl.2023.103662.

[13] Soumyadeb C, Pawan B, Prasanta K, Sian J, Amelie A, AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework, *Journal of Business Research*, Vol.144, 2022, pp. 31-49. https://doi.org/10.1016/j.jbusres.2022.01.069

[14] Ai-Hsuan C, Silvana T, Yu L, Emotion and service quality of anthropomorphic robots, *Technological Forecasting and Social Change*, Vol.177, 2022, pp. 0040-1625. https://doi.org/10.1016/j.techfore.2022.121550.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)