The Application of Generative Artificial Intelligence Technology in Conversational Robots

Yu-Yu, Yen[1*], Chen-Yu, Tsai[2] and Jui-Hung Kao[2]

[1] Center of General Education, Shih Hsin University, Taipei, Taiwan
[2] Department of Information Management, Shih Hsin University, Taipei, Taiwan

* melyen@mail.shu.edu.tw

Abstract. As the value of natural language processing in the business world becomes increasingly apparent, many organizations are adopting the technology. Many companies are utilizing big data and natural language processing to develop a variety of chatbots that extend beyond basic question-and-answer functionality. These chatbots act as a communication conduit between the customer and the business, providing real-time, targeted service. GPT-2 has unquestionably attained the pinnacle of language generation modeling in the ongoing evolution of technology. GPT-2's transformer-based Masked Self-Attention mechanism enables it to generate a diverse range of responses contingent on the context of the sentence. These achievements are supported by a substantial quantity of data, which illustrates the pivotal role of data in the contemporary development of AI. About TTS (Text-to-Speech) and STT (Speech-to-Text) technologies, their integration into a dialog model such as ChatGPT not only provides users with a more natural interactive experience but also broadens the possibilities of ChatGPT in a variety of application scenarios to a certain extent. Consequently, this further enriches the manner of communication between humans and machines.

Keywords: Generative AI, Conversational Robots, Artificial Intelligence Applications

1 Introduction

A review of the history of artificial intelligence in NLP must include the two pioneering examples of Google Translate and IBM's Watson, which have provided people around the world with fast and accurate language translation tools through the use of Deep Neural Network (DNN) and Bidirectional Encoder Representations from Transformers (BERT) models. Using Deep Neural Network (DNN) and Bidirectional Encoder Representations from Transformers (BERT) models, Google Translate provides individuals across the globe with expedient and precise language translation tools, thereby facilitating the dissemination of cross-language communication.

IBM's Watson represents a unique convergence of machine learning and natural language processing (NLP) technologies, enabling it to simulate human conversations. This novel approach to interaction is enabled by the combination of these two technologies. Nevertheless, the commercial value has also attracted the attention of numerous enterprises. The use of big data and NLP technology has enabled numerous companies to develop chatbots designed to meet customer service requirements. These robots have evolved beyond their original role as mere question-and-answer tools, becoming a conduit between customers and businesses. They provide real-time, professional service and advice.

2 Materials and Methods

GPT-2 (Generative Pre-trained Transformer 2, GPT-2) has undoubtedly become a masterpiece of speech generation modeling in this long road of technological evolution. With its Masked Self-Attention mechanism, this Transformer-based model successfully generates variable responses based on the meaning of sentences, all supported by a large amount of data, proving once again the importance of data in today's AI era.

However, to talk about the origins of chatbots, we need to go back to 1966 [1]. That was when Joseph Weizenbaum of the Massachusetts Institute of Technology (MIT) developed ELIZA, a first-generation robot that could simulate a psychiatrist's consultation. From ELIZA to the current GPT-2, the development of chatbots has shown a progression from simplicity to complexity, from fixed responses to intelligence generation. These early robots, such as A.L.I.C.E., Watson, and Siri, were powerful but still limited by fixed responses and domain-specific knowledge.

Fortunately, the advent of technology and the growth of social networks have enabled modern chatbots to overcome the limitations previously imposed by the aforementioned issues. Several platforms, including Facebook, Line, and WeChat, now provide self-build tools that enable individuals to develop their own chatbots on demand. Moreover, the financial technology industry is driving an increasing number of financial institutions to adopt these intelligent tools, which provide customers with enhanced convenience.

The advent of chatbots has led to the displacement of certain roles traditionally performed by human workers, including those of online customer service representatives and educators. From the initial stages of rule-based chatbots to the rapidly evolving era of artificial intelligence (AI), the performance of chatbots has continued to improve. In the current era, chatbots can engage in natural language interactions and learn from experience. This project aims to design a chatbot using OpenAI's GPT. This novel technology may alter future research trajectories on chatbots [2].

Recent developments in the fields of artificial intelligence and natural language processing have resulted in the creation of highly efficient and versatile intelligent chatbot models. These models have the potential to achieve a level of linguistic competence that is comparable to that of a human conversationalist. The proposed approach involves the design of a chatbot model that is trained on an open conversation dataset.

The objective is to create a model capable of mimicking human conversations while maintaining the emotional nuances present in human speech [3].

3 Methodology

3.1 The field of Natural Language Processing (NLP)

Although NLP may still be an abstract and elusive concept for beginners, it is evident that with the rapid development of AI technology, we have observed its extensive range of applications in various industries and our daily lives. Upon opening a translation tool and entering a sentence, the machine can complete the translation instantly. Similarly, when an article is read, the sentiment analysis tool determines the emotional tendency of the article. Furthermore, the computer's automatic completion of a paragraph indicates the operation of natural language processing technology.

In a traditional classroom setting, educators are responsible for disseminating knowledge to multiple students. One of the inherent limitations of such a system is that students often require individual attention while learning. However, this is nearly impossible when one educator is responsible for the supervision of 25 students. The Educational Chat Robot system employs artificial intelligence (AI) to interact with students and facilitate their comprehension of mathematical concepts by providing explanations analogous to those offered by human educators [4]. The potential of humanoid robots as teaching and social assistants is considerable. However, children frequently anticipate that these robots will interact in ways analogous to those of humans [5].

Recently, there has been a discernible increase in the demand for conversing with dialogue systems. One of the principal difficulties this chat conversation system faces is its inability to respond appropriately to users' spontaneous remarks. This is mainly because these systems frequently generate incongruous responses to previous topics. This is attributed to their inadequate management of conversation history. To address this issue, conversational systems frequently generate cautious, straightforward, and brief responses to maintain consistency within the conversation. However, this approach frequently results in conversations that lack variety and depth. This study aimed to develop a chat conversation system to enrich the input sentences through two main components [6].

To achieve the aforementioned goal, it is insufficient to rely solely on hardware technology; rather, applying mathematical models and algorithms is of greater importance to facilitate machine understanding of language. Over the course of this extensive research endeavor, we have developed a multitude of intricate and highly effective models. To comprehend the fundamental concept of recurrent neural networks (RNNs), it is helpful to consider them a time machine that continually returns to the past to retrieve crucial insights from historical data. This enables an understanding of the present circumstances. The model is inspired by the Hidden Markov Model (HMM), which is a relatively old model. As with the Hidden Markov Model (HMM), the Recurrent Neural Network (RNN) considers the impact of past information on the current state. However, it uses all the past information as a reference point to predict future trends.

In the recurrent neural network (RNN) architecture, time series data, such as $X(t)$, are input into the model, which then produces the output $Y(t)$. It is of paramount importance to recognize that the green line, which represents the state at a specific time point, will influence the input state at the subsequent time point (as illustrated in Fig. 1). This continuous time series training approach enables RNNs to demonstrate excellent performance in the processing of continuous data, such as speech and text.

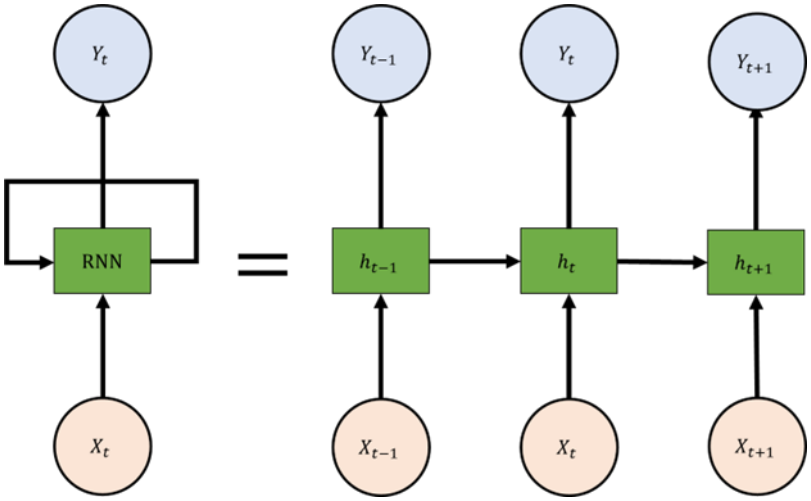

Fig. 1. The Cyclic Neural Network Architecture

3.2 Transformer

Although attention is regarded as a highly promising new structure, it is still considered a mere supplementary tool to be incorporated in RNN-based models. Nevertheless, over time, researchers have begun to recognize the significant potential of the attention mechanism [7]. At the time, Neural Machine Translation (NMT) was introduced to the field of attention mechanisms. Although attention was regarded as a promising new structure, it was still considered a mere supplementary tool that could be integrated into the RNN-based core model [8].

The Transformer model is the result of a series of cognitive processes. In this model, the attention mechanism is no longer regarded as an additional component; rather, it becomes the fundamental element of the model, responsible for the extraction and transformation of information. In particular, the base structure of the Transformer (illustrated in Fig. 2) allows for a detailed observation of its components. The left-hand side of the diagram depicts the encoder component, which comprises N=6 identical layers. Each layer has a multi-head self-attention mechanism and a fully connected feedforward network. The multiread attention mechanism enables the model to simultaneously capture multiple correlations in different contexts. This is achieved by applying distinct linear transformations to Q, K, and V and synthesizing the outcomes of these disparate attentions.

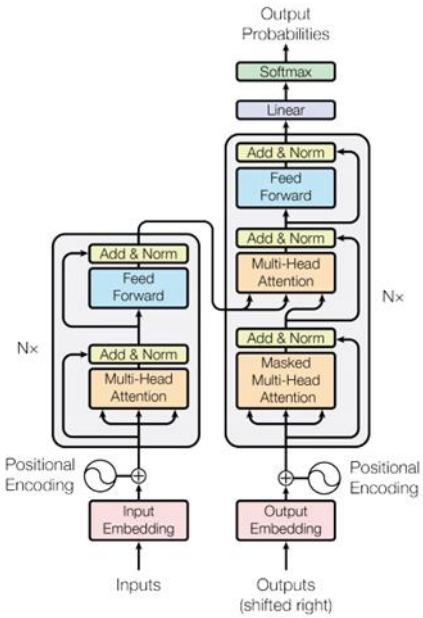

Fig. 2. Transformer architecture.

The Transformer structure is designed to facilitate the efficient capture of temporal and contextual relationships in linguistic data, in contrast to the cyclic structure of RNN. Indeed, the six-layer structure of the Transformer enables the capture of highly detailed contextual information.

4 Discussion and Conclusion

In the context of dialogue robots, retrieval modeling represents one of the most prevalent technologies. For instance, Microsoft's Little Bing was founded on this technology. In response to a user's query, this dialogue robot employs a search process to identify the most suitable response from its extensive repository of previous interactions. Subsequently, the robot presents this response to the user. However, it remains unclear how to determine which answer is the most appropriate. There are a multitude of potential methodologies and algorithms for this process to be undertaken. One of the most commonly employed methodologies is based on cosine similarity calculation. This is a measure of the similarity between the directions of two vectors, which can be employed to assess the similarity between a user query and a question within a corpus. In this manner, the robot can identify the most analogous question within the corpus to the user's query and subsequently provide the corresponding answer.

References

1. J. Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine,” Communications of the ACM, vol. 9, no. 1, pp. 36-45, 1966.
2. M. M. SWATHI, and K. GOPALREDDY, “AI-BASED CHATBOT WITH GPT-3,” Journal of Engineering Sciences, vol. 14, no. 08, pp. 644-654, 2023.
3. G. Chettiar, A. Shukla, P. Nalwaya, K. Sethi, and S. Prakash, “Impersonated Human Speech Chatbot with Adaptive Frequency Spectrum,” 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), pp. 1-7, 2023.
4. A. Attard, and A. Dingli, “EDUCATIONAL CHATBOT USING A HYBRID-RESPONSE APPROACH,” ICERI2021 Proceedings, pp. 4192-4201, 2021.
5. A. Lekova, P. Tsvetkova, T. Tanev, P. Mitrouchev, and S. Kostova, “Making humanoid robots teaching assistants by using natural language processing (NLP) cloud-based services,” Journal of Mechatronics and Artificial Intelligence in Engineering, vol. 3, no. 1, pp. 30-39, 2022.
6. T. Kinouchi, and N. Kitaoka, “A response generation method of chat-bot system using input formatting and reference resolution,” 2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), pp. 1-6, 2022.
7. F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” Neural computation, vol. 12, no. 10, pp. 2451-2471, 2000.
8. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 3th International Conference on Learning Representations, pp. 1-15, 2014.