Midoriko Chatbot:

LSTM-based Emotional 3D Avatar

Yu-Ting Wan, Cheng-Chun Chiu, Kai-Wen Liang, and Pao-Chi Chang
Department of Communication Engineering
National Central University
Taoyuan City, Taiwan
pcchang@ce.ncu.edu.tw

***Abstract*—In this work, we developed a new model of chat robot with facial expression interaction. Different from the voice assistant that only interacts with text and voice, we let the chat robot interact with the user with a specific role play. A text-sentiment-analysis network is added to this system to recognize the emotion of the response sentences. Moreover, we created a 3D-constructed anime character "Midoriko" to have the image of the chat robot more vivid. Our goal is to provide better experience to users. The technical parts of this work is based on neural network techniques.**

Keywords—Neural Network, Avatar, Chatbot, Unity, LSTM, Companionship, 3D module

I. INTRODUCTION

Thanks to the rapid development of artificial intelligence technology, many applications bring a lot of convenience and fun to human life. Task-oriented artificial intelligence assistants such as voice assistants have been launched by many large companies and are widely used in our daily life.

In some applications, we let a robot do some easy tasks through text or voice interactions. Although their ability to recognized the meanings of words is amazing, we believe that the interaction between the robot and the user can be more meaningful on specific perspectives. In addition to text and sound, chatbots can have expressions and body movements through graphical interfaces as role-playing. It is not just a task-oriented artificial intelligence assistant, but a companionship. In this age of rapidly evolving technology, people's addiction to the Internet is increasing day by day in the modern society, whether it is the elderly, children or many single people may lack sufficient interpersonal interaction so as to result in psychological problems. We believe that partners created with artificial intelligence may alleviate the problems caused by alienation in society [1]. Therefore, our goal is to create a good friend, lover and even a life partner, which can bring more companionship to this society.

II. SYSTEM ARCHITECTURE

A. Proposed system

The system module is illustrated in Fig.1. The Chatbot Network, a neural network, first accepts the input string from the user and then outputs the response string to the Text-Sentiment-Analysis Network. The receiver calculates the emotional score of the response. Finally, both the response and sentiment scores will be sent to the 3D module. The function of this 3D module is to create actions and expressions for the character through emotional scoring. The character, Midoriko, then will be display by the responses
.

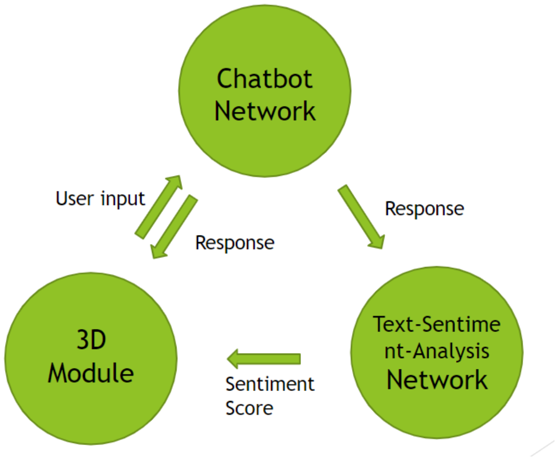

Fig. 1. System Modules

B. Server

Because of the large amount of the computational consumption for a sequence neural network class, a Client-Server architecture, as shown in Fig.2, is used in order to make our service available to any device. We define a specific communication protocol such that all requests under this protocol will be responded by the server. It is efficient to allocate most computational cost on the server to make it possible to extend our service to any platform or environment, such as Android, iOS, Windows. Of course, if you are willing, Raspberry Pi is OK as well.

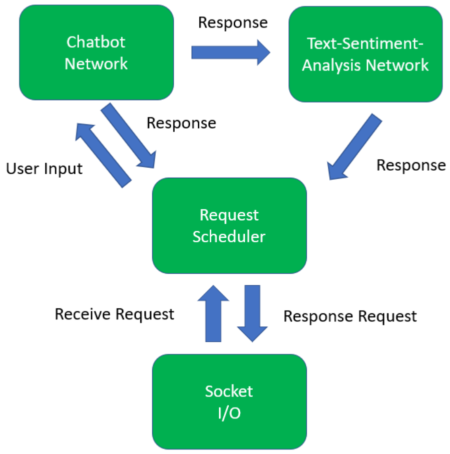

Fig. 2. Server Functions

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

978-1-7281-3575-5/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: National Central University. Downloaded on July 15,2020 at 07:29:04 UTC from IEEE Xplore. Restrictions apply.

C. Client

In the Client end, as shown in Fig. 3, the user's input is send to the server and then both the response sentence and the emotional score will enable the 3D module to make the appropriate action. In order to establish a user friendly interface, we provide user to input sentences through voice with some existing voice assistants that can be acquired on the Internet. At the same time, thanks to the multi-platform features of the Unity engine, our existing code and parts of the 3D module can be exported to other platforms and environments, such as Android. VR devices such as HTC Vive can also be used. ◦

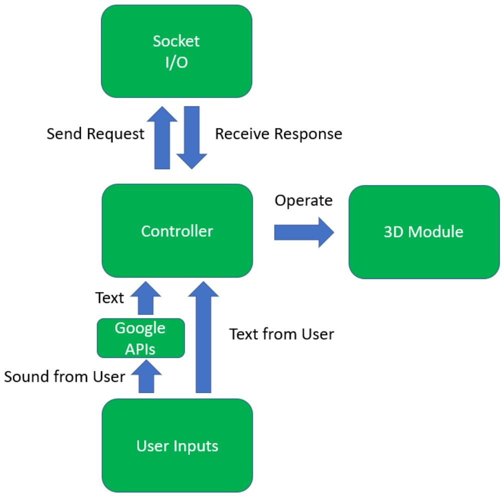

Fig. 3.Client Functions

III. NEURAL NETWORK AND DATA COLLECTOIN

We use LSTM neural network with back-propagation in our ChatBot network module as shown in Fig.4. The training set is collected from Cornell Corpus and the Twitter dialog. In addition, we convert words into vectors with 1024 dimensions by Embedding. The converted vectors are used as input data and then the network will outputs results with the same dimension. That is, a sentence to sentence neural network.

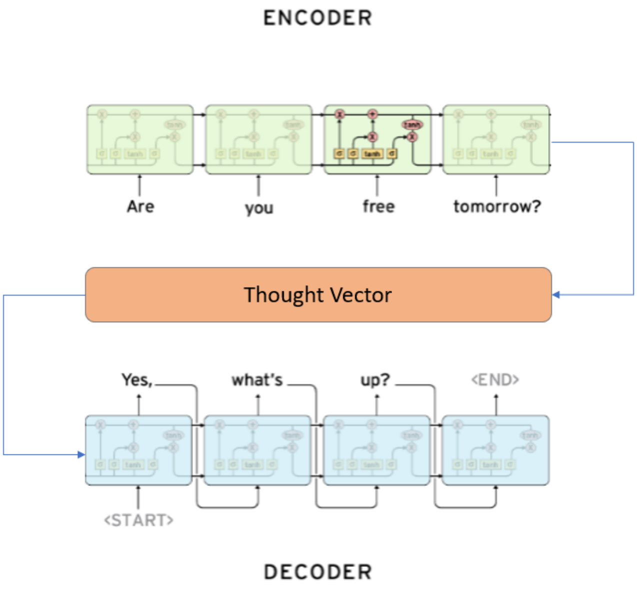

Fig. 4. Sequence to Sequence Structure

A. Word to Vector Model

The objective of the word vector model is to assign each word a coordinate in 128-dimensional vector space[4], and words that have similar meanings will be close to each other in this coordinate system. In this work, we use Google's nnlm-en-dim128 pre-training model, which meets the requirement mentioned above.

B. Stemming

Stem extraction is a method of dealing with words. Mainly used in dealing with issues about different types and tenses of the same word.

Ignoring the change of different types and tenses will increase the number of word vectors significantly. Therefore, words are changed to their root type, omitting the plural type, possession, verb change, etc. , by a predetermined algorithm. By using a good algorithm, we can also confirm the root of a word by its pronunciation that is beneficial to provide much more information and help to accurately normalize the word vector.

IV. 3D MODULE CONSTRUCTION

A. Technical tools & methods

Unity, the adopted framework, provides a C# interface that creates event loops and supports VRM files. With virtual cameras, it provides a suitable display angle. We also use dynamic object loading to save memory space.

B. Emotion animation & action module embedding

The work flow of emotion animation is shown as Fig.5. The embedded BlendShape creates the emotion of our character, Midoriko, by pulling each endpoint in the model structure proportionally. In addition, the 3D Morphing improves the animation change, for example, from happy to sad, smoothly. The animation controller records animation states, which are controlled in the Emotion Layer.

Fig.5 also shows the Motion layer structure. The "Entry" is the start point for an interaction when a response is received. The workflow of the animation will follow the solid arrow. The state will stay in "Idle" until the received emotional score reaches "Anystate", the animation will enter "Anystate" (green dotted line) and change to the state we set (Happy, Sad, etc.) according to the score, and finally returns to the "Idle" state.

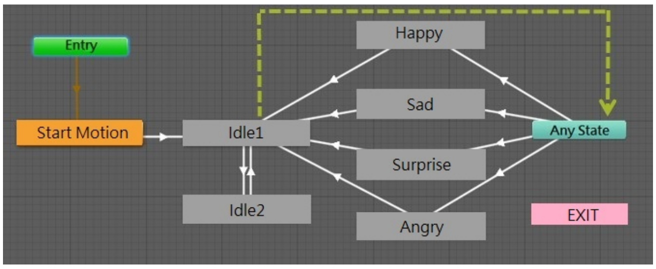

Fig. 5. Emotion & Motion Control Layer Structure

C. Music and sound effects Positioning

The sound part is divided into action sound and Background Music (BGM). Through script writing, the audio file is read as a variable, and the threshold control is used to achieve the sound playback effect synchronized with the animation.

Authorized licensed use limited to: National Central University. Downloaded on July 15,2020 at 07:29:04 UTC from IEEE Xplore. Restrictions apply.

D. Background image dynamic loading

Different cameras are used to catch background images and the buttons are used to trigger camera switching. The design of the dynamic loading is to save memory space.

V. CHARACTERISTIC

A. Highly flexible: Users are free to choose characters, scenes, and sound effects.

B. Neural network modules: Various characters can be built by using different data set.

C. Use VRM files: Extension of the module on VR, AR hardware is available.

D. With the Client-Server architecture: Client end can be deployed anywhere.

E. Without expensive VR equipment, you can still achieve 3D projection with a simple transparent plastic case (Fig6.) 。

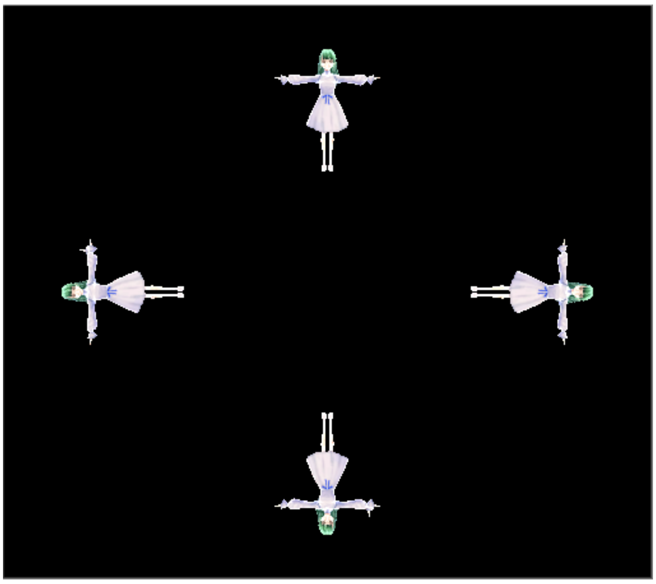

Fig. 6. 3D projection diagram

VI. FUTUREWORK

In this work, our Midoriko can only respond to sentences input by the user, however, in most chatty scenarios, the chat will be around a specific topic and our system should be able to remember what the user has said in the past to make the whole conversation process more reasonable. Therefore, simply responding to the user's sentence doesn't complete our goal. We believe that chat bots should have the ability to ask questions from questions and answers to continue the subject of a talk. And may even have the ability to open the topic. We also noticed that being able to handle some simple requirements can also increase the user's adhesion to Midoriko, making her presence more like a family.

We will improve our Midoriko Chatbot step by step, and we believe that it will be all the best after strengthening the semantic understanding and 3D Module. Here are some work priorities in the future.

A. Improve semantic understanding performance

The size of our LSTM Model is currently limited by the computing power of the server. In order to ensure that the response time of each request is not too long, we also hope that each response processing time does not exceed 2 seconds when increasing the complexity of the model. At the same time, the source of our training set is too small, and the diversity is not enough, which makes our Model's understanding of the text not accurate enough, so often can not make a correct response. This is also reflected in our sentiment analysis neural network. Sometimes the judgment of emotions misleads the 3D Module to make incorrect actions, so strengthening semantic understanding is our important task.

B. Improve chat performance

A purposeless chat may be a good way to relax, but for someone who wants to focus on a topic may not be happy. Anyway, we hope that Midoriko can be more humanized. He must at least have the ability to grasp the sentence and be able to ask questions actively. At the same time, he must also increase the ability to remember the content of conversations in the dialogue in order to give the user a good experience. A further goal is to make Midoriko can analyze the information in the current society for chatting.

C. Customized voice with the character model

Currently our character pronunciation model uses the Google TextToSpeech API, so the sound is determined by the API. We noticed that the pronunciation sound is not very well matched with our Midoriko module. In the subsequent improvements, a speech model will be trained to improve the problem, making the 3D module and the pronunciation sound more matching.

D. Task-Oriented Function

After the basic functionality of the chat is in place, we will try to add some special features to our model, just like the artificial intelligence assistants on the market, giving him the ability to perform some simple tasks for the user, if Midoriko can Setting an alarm for the user, I think it must be a great experience. Although these features are very good at other artificial intelligence assistants, those artificial intelligence assistants are currently only available on certain platforms, so we want to extend this service to the world with Midoriko's cross-platform features.

VII. RESULT

This work proposed a LSTM-based emotional 3D module Chat robot, Midoriko, which features on its excellent interactive interface and the ductility. Separated modules can be replaced by user-defined modules to fulfill different requirements. Although this neural network model requires a large computational consumption, it can be deployed on any powerful computer as a server end so that the client end can be launched on devices by users. With the flexibility of choosing the characters and scenes [6], we believe that this module could bring users a brand new experience of interacting. “Midoriko” could be the best friend to the user. In the future, we will try to let "Midoriko" run on VR devices.

Authorized licensed use limited to: National Central University. Downloaded on July 15,2020 at 07:29:04 UTC from IEEE Xplore. Restrictions apply.

VIII. DEMO LINK

DEMO Video: https://youtu.be/a4rOxzL_fAs

ACKNOWLEDGMENT

This work was supported in part by Ministry of Science and Technology under grant no. MOST 108-2634-F-008-004.

REFERENCES

[1] H. Chaturvedi, N. D. Newsome, and S. V. Babu, “An Evaluation of Virtual Human Appearance Fidelity on User’s Positive and Negative Affect in Human-Virtual Human Interaction”, *IEEE Virtual Reality (VR)*, Aug. 2015

[2] S. Hochreiter and J. Schmidhuber. “Long short-term memory,” *Neural Computation,* 9(8):1735–1780, 1997.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Representations by Back-Propagating Errors”, *Nature,* vol. 323, pp. 533-536, 1986.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space.” *ICLR Workshop*, 2013.

[5] T. Bergmanis and S. Goldwater, “Context sensitive neural lemmatization with lematus,” *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long Papers), volume 1, pages 1391– 1400

[6] Y. Kuang and X. M. Bai, “The Research of Virtual Reality Scene Modeling Based on Unity 3D,” *2018 13th International Conference on Computer Science & Education (ICCSE)*, Sep. 2018.

Authorized licensed use limited to: National Central University. Downloaded on July 15,2020 at 07:29:04 UTC from IEEE Xplore. Restrictions apply.