Model of Multi-turn Dialogue in Emotional Chatbot

Chien-Hao Kao
Institute of Medical Science and Technology
National Sun Yat-sen University
Kaohsiung, Taiwan
kao910019@gmail.com

Chih-Chieh Chen
Institute of Medical Science and Technology
National Sun Yat-sen University
Kaohsiung, Taiwan
chieh@imst.nsysu.edu.tw

Yu-Tza Tsai
Department of Mechanical and Electro-Mechanical Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
jamestsaispc@gmail.com

***Abstract*—The intent recognition and natural language understanding of multi-turn dialogue is key for the commercialization of chatbots. Chatbots are mainly used for the processing of specific tasks, and can introduce products to customers or solve related problems, thus saving human resources. Text sentiment recognition enables a chatbot to know the user's emotional state and select the best response, which is important in medical care. In this study, we combined the multi-turn dialogue model and sentiment recognition model to develop a chatbot, that is designed for used in daily conversations rather than for specific tasks. Thus, the chatbot has the ability to provide the robot's emotions as feedback while talking with a user. Moreover, it can exhibit different emotional reactions based on the content of the user's conversation.**

Keywords-Chatbot; Multi-turn; Emotional category; Seq2Seq; SeqGAN;

I. INTRODUCTION

The commercial application of a conversational agent or chatbot is inputting and analyzing the intent of a user and providing appropriate responses, such as searching for a restaurant, making inquiries regarding the weather conditions, and answering frequently asked questions on specific platforms. Context judgment and intent transmission are critical for the conversational agent for conducting specific tasks. However, the conversational agent does not provide an adequate response in a daily chat, which lacks a particular intentions.

In recent years, many chatbots have applied the sequence-to-sequence (Seq2Seq) generative model [1] to generate replies for the daily conversations, thereby shifting the goal of the answer from a specific domain transaction to a general daily conversation. Although Seq2Seq is built on a translation model, the model can adapt to outputs or inputs of variable lengths and exhibit adequate performance in Q&A conversation. The purpose of the generative model is to make the attitude and feedback of human-computer interactions more humanized and less dull. Note that [2] applied the Seq2Seq architecture to the generative adversarial network (GAN), and introduced the method of reinforcement learning to calculate the discrete loss to increase the similarity of a translated sentence to human response. Moreover, [3] and [4] used sequence generative adversarial network (SeqGAN) as a chatbot to generate response, and [4] sloved the problem of long training-time of SeqGAN.

For the sentiment recognition of a text, [5], [6] categorized the input sentences in terms of emotions categorize, and followed the different needs from the database or from the generative model to give the user some advice to solve simple medical care problems. Some studies have used multimodal models to predict the emotions of user entering sentences, [7] designed an internal emotional memory to produce different emotional effects on decoded sentences, and [8] designed a sound-based approach for the human-computer interface. Input sound chat with an avatar can be used to observe the body movements of the chatbot avatar, thus highlighting the importance of natural interactions and physical movements in a conversation.

The main purpose of this study is to use a generative model chatbot that changes emotions on the basis of the user's conversation in daily chat and responds to the user with the transformed emotions. This will provide chatbots with same emotional changes ability as humans and more sensitive to the context of conversations, and exhibit improved machine interaction performance.

II. RELATED WORK

A. EmotionLines

This study used the EmotionLines [9] corpus training set, which contains multi-turn and multi-party conversation corpus with emotional tags. There are eight types of emotion tags: neutral, joy, surprise, disgust, fear, sadness, anger, and non-neutral.

B. SeqGAN

In this study, the SeqGAN [2] model is used as the basic model and extended. SeqGAN uses the Seq2Seq [1] model and combines the concept of the GAN [10], using Seq2Seq as a generator and training a classifier discriminator to identify true and false samples. The purpose for this training is to enable the discriminator to score the sentence generated by the generator. When the score is high, the discriminator thinks that the generated sentence is more similar to the sentence spoken by the real human. For using the GAN in the text generation, the policy gradient reward mechanism must be used by reinforcement learning because the dictionary data are discrete. This study uses a previous study [4] as reference to score every time step of the decoding, thus reducing the amount of training time compared with that required in [3] by the using Monte Carlo Search.

C. Response Judgment

Note that [11] used the self-attention and cross-attention mechanism to construct a chatbot model for a multi-turn dialogue. However, conversations are often not in the form of questions and answers. It is necessary for the robot to judge whether a response must be generated in the present moment. It is necessary to choose to remain silent in a multi-party conversation. This study uses a latent encoder to deliver multi-turn dialogues of a conversation's content and analyze whether the chatbot is generating a reply at the correct time.

978-1-7281-4666-9/19/$31.00 ©2019 IEEE

D. Emotion Prediction and Transfer

Training sentiment recognition can be used for future chatting applications and user data analysis [5], [6] by training the emotion classifier to decide whether to find answers in the database or to generate daily conversations by using a generative model. [12] and [13] specified emotional tags are used to generate sentences that contain emotions. Although there are some practical examples in identification and generation, there is no development pertaining to chatbot control and how it changes its emotions. In this study, the context of the multi-turn dialogue model can automatically change the emotion of the robot itself on the basis of the sentence input and emotional state of the current chatbot, which can affect the response generated by the chatbot and can make the reply of the current chatbot more emotional than that of the previous chatbot.

III. THE MODEL

Fig. 1 displays the model designed in this study, which contains the input of social software. A sound input by a user is converted to text information by voice recognition, or a sentence is directly input by typing. natural language processing filters out unknown characters and other special symbols and enters the sentence. The program network generates a response and then outputs the text to the social software. In the dialogue, the listening mode can be temporarily switched on by calling the name of the chatbot, and each sentence can be read in a short time to facilitate chatting.

The generative model presented in Fig. 2 uses the Seq2Seq architecture. To accommodate multi-turn of dialogue and emotional tag input, a latent encoder and emotion encoder were added. The input and output increase the responsibility of passing the last latent state and the emotion state. These states are stored in a database, and the stored states are used in the next round. The transfer state is then used for prediction and generation, but the initial state is only used at the beginning of the topic.

Fig. 3 displays the latent encoder in the recurrent neural network model. We used gated recurrent units (GRU) [14] to

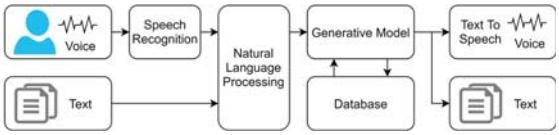

Figure 1. External structure of users communicating with chatbots by text input or voice input.

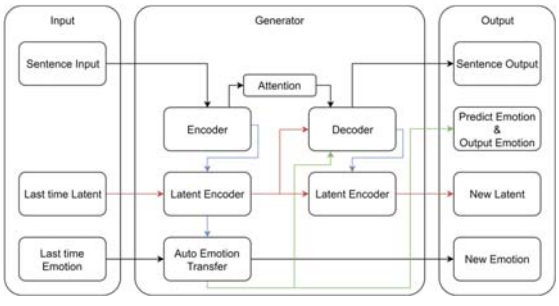

Figure 2. Generative model with different colors for distinguishing between overlapping arrows.

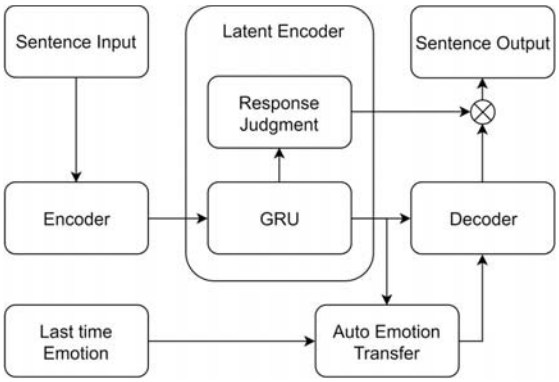

Figure 3. Internal structure of the latent encoder, including the resonse judgment classifier to determine whether to output the status for generating a response.

improve the efficacy and speed of the model. Only a time step output by the latent encoder when an input is provided. A classifier response judgment inside judges whether it is the correct time for answering. Classifier output control the output of the decoder and auto emotion transfer.

Fig. 4 illustrates the auto emotion transfer, which used GRU structure. The current mood is obtained by using the output of the latent encoder. There are two emotion classifiers—: 1) the one that is used to predict emotions related to the input sentences of the user and 2) the one that enables the chatbot to respond to emotions on the basis of the input sentence. By referring to the methods presented in [15] and [13], the output emotion was decoded as a start token for the decoder to generate a reply.

The training phase was divided into three steps. First, pre-train the generator. The state is passing between each turn, thus the classifier and decoder must be trained together. The response judgment can grasp the timing of an answer. The auto emotion transfer conduct an emotion conversion, and output an emotional tag.

Second, train the discriminator. Unlike SeqGAN, which is based on a multi-turn training set, our model must regroup generated sentence first, as shown in Fig. 5. Then, the <NONE> input and output data must be skipped. Finally, grouped data are used to train discriminator.

Finally, the generator is trained using the GAN to make the response more realistic. The regroup multi-turn sentence is used to train the generator, and the discriminator is used to score the sentence generated by the generator, as presented in

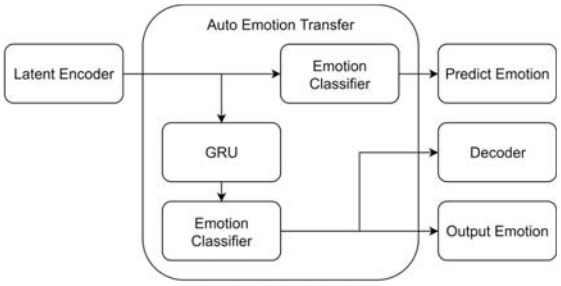

Figure 4. Internal structure of the auto emotion transfer. The emotion classifier contains predict and output emotions and then outputs the emotional tags to the decoder to obtain the decode response.

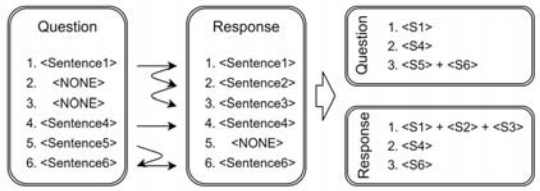

Figure 5. Regrouping of the sentence combinations without consideration of the non-response parts. Each sentence is separated by the end of the signal tag. The discriminator accurately judges the multi-turn of the dialogue..

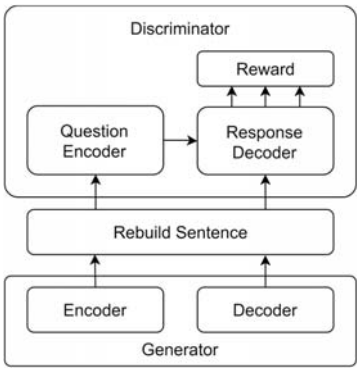

Figure 6. Adversarial training structure. The difference is that a stentence must be regrouped, and the response decoder subsequently generates a reward at every time step.

Fig. 6. Subsequently, the used policy gradient of reinforcement learning is used to update the model parameters, for solving the problem pertaining to text discretization in the vocabulary.

IV. EXPERIMENTS

A. Dataset

Table I lists the number of emotional tags included in the training, development, and testing datasets in EmotionLines. EmotionPush is a multi-turn corpus for two people, and Friends is a multi-party and multi-turn corpus. Some of the emotional tags are few. The emotion transfer response begins after about approximately ten training sessions, because the data in each group basically have a neutral mood; thus, we did not use the upsampling method to enlarge the training dataset.

TABLE I. TOTAL EMOTION AMOUNT

Categories	*EmotionPush*		*Friends*	
	Train	***Test***	***Train***	***Test***
neutral	7148	1882	4752	1287
joy	1482	458	1283	304
sadness	389	87	351	85
fear	36	2	190	36
anger	94	37	524	163
surprise	435	93	1221	286
disgust	85	15	244	68
non-nerural	1064	223	2017	541

B. Response Judgment

Table II reveals that the accuracy of responses to replies is high. This indicates that after training, the network could determine whether the speaker had finished speaking; consequently, the chatbot would wait for the user to finish speaking before replying.

C. Emotion Prediction and Transfer

Table III presents the emotion prediction process with the highest natural emotion accuracy. There were fewer tags for anger, disgust, sadness, and fear and accuracy of identification of these emotions in the training set was greater than 0.8, but the test accuracy was relatively low, indicating that the correlation between the two datasets with fewer emotion tags is low.

Because the response judgment affects the emotion transfer when pre-training is performed, emotion transfer begins to be effectively trained after the response judgment accuracy is improved. The response starts later, and Table III reveals that the emotion transfer accuracy is much lower than the emotion prediction accuracy. Except for the emotional tags with weak correlations, the output emotions could not achieve higher accuracy. This result may indicate that the emotional tags of the dialogue model are different from actual potential human emotions and that the actual human emotions cannot be interpreted with few tags.

We used perplexity as the evaluation criterion for the generator that generates good or bad sentences. Table IV lists the perplexity of the two training sets. The Friends training

TABLE II. RESPONSE JUDGMENT ACCURACY

Categories	*Response Judgment*	
	EmotionPush	***Friends***
Accuracy	0.787	0.859

TABLE III. EMOTION ACCURACY

Categories	*Emotion Accuracy*			
	EmotionPush		***Friends***	
	Predict	***Transfer***	***Predict***	***Transfer***
neutral	0.845	0.792	0.722	0.594
joy	0.594	0.159	0.436	0.072
sadness	0.379	0.011	0.175	0.0
fear	0.0	0.0	0.012	0.0
anger	0.027	0.0	0.322	0.006
surprise	0.323	0.032	0.418	0.025
disgust	0.067	0.0	0.068	0.0
non-nerural	0.141	0.082	0.222	0.453
total	0.698	0.565	0.497	0.380

TABLE IV. PERPLEXITY AND ACCURACY

Categories	*EmotionPush*	*Friends*
Perplexity	332.7	127.3

set had a lower perplexity. Table V reveals the difference between the target and the generated response in multi-turn dialogue. The chatbot can grasp the timing of an answer, show Response Judgment is training well. For some stuttering texts that need to improve the NLP.

V. Conclusion and Future Work

The reasons for the poor emotion recognition, including the data is imbalance, and the dataset is generated by the TV series in which the actors may express strong emotional ups and downs to express the tension of the story. We will improve this issue by adding tags to quantify the emotion. For continuous positive or negative emotions, give a higher value than usual, which can make the emotional transition appear smoother, rather than a sudden change.

Most of the training datasets of the generative model for current chatbots are question–answer chats, although the generative model differ from those for specific task, the answers are dull and vague in daily conversation. There are still many factors that affect the content of a conversation. We assumed that there is no standard answer in a chat, but the generative model chatbot uses Seq2Seq from the translation model as a generator. Therefore, in this study, changes have been made to generate multiple types of responses in the presence of many different factors. The emotion feedback by a chatbot is not specified by a human or rule-base but is automatically changed by learning, thus making the response more natural.

In the future, we hope to improve this model so that it can serve as a basis for integrating life-long learning and intent judgment corpus as well as applied in the contexts of long-term care and home medical care.

References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104-3112.

[2] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[3] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," arXiv preprint arXiv:1701.06547, 2017.

[4] Y.-L. Tuan and H.-Y. Lee, "Improving conditional sequence generative adversarial networks by stepwise evaluation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 788-798, 2019.

[5] D. Lee, K.-J. Oh, and H.-J. Choi, "The chatbot feels you-a counseling service using emotional response generation," in 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017: IEEE, pp. 437-440.

[6] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in 2017 18th IEEE International Conference on Mobile Data Management (MDM), 2017: IEEE, pp. 371-375.

[7] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[8] P. A. Angga, W. E. Fachri, A. Elevanita, and R. D. Agushinta, "Design of chatbot with 3D avatar, voice interface, and facial expression," in 2015 International Conference on Science in Information Technology (ICSITech), 2015: IEEE, pp. 326-330.

[9] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," arXiv preprint arXiv:1802.08379, 2018.

[10] I. Goodfellow et al., "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.

[11] X. Zhou et al., "Multi-turn response selection for chatbots with deep attention matching network," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1118-1127.

[12] R. E. Banchs, "On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data," in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017: IEEE, pp. 1364-1367.

[13] X. Sun, X. Chen, Z. Pei, and F. Ren, "Emotional Human Machine Conversation Generation Based on SeqGAN," in 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018: IEEE, pp. 1-6.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[15] C. Huang, O. Zaiane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 49-54.

TABLE V. MULTI-TURN AND MULTI-PARTY DIALOG RESPONSE

Source	1.	oh well , it 's not so bad .
	2.	yeah , most of the damage is pretty mostly contained in the bedrooms .
Target	1.	oh !
Generate	1.	what ?
Source	1.	my god !
	2.	hey buddy , do you think i can borrow your uniform this thursday ?
	3.	excuse me ?
	4.	joey ! he 's working ! you would look
	5.	oh , how bad is it ?
Target	1.	oh , it 's bad . it 's really bad . the only thing in there that is n't burned is an ass . which i do
Generate	1.	i wouldn come on , that 's sorry .
Source	1.	oh , bob , he was nothing compared to you . i had to bite my lip to keep from screaming your name .
Target	1.	well , that makes me feel so good .
Generate	1.	he is seven !
Source	1.	it was just so awkward and bumpy .
	2.	bumpy ?
Target	1.	well , maybe he had some kind of uh , new , cool style , that you are not familiar with .
	2.	and uh maybe you have to get used to it .
Generate	1.	so what me , that was not ready is ready now , my woman are he to play .
	2.	okay , and i can n't go you could .
Source	1.	hey , joey , could you pass the cheese ?
Target	1.	yeah .
	2.	listen uh , i 'd prefer it if you did n't call me joey .
	3.	since i do n't know anyone here , i thought it 'd be cool to try out a cool work nickname .
Generate	1.	but i worry monica to prank , i think i could get to do you something !
	2.	it's right man as .
Source	1.	hey , dragon ! here's your tips from monday and Tuesday .
Target	1.	There 's like-there 's like 300 bucks in this one !
Generate	1.	that ?
Source	1.	yeah , people get pretty generous around the holidays . and it never hurts to wear tight trousers .
Target	1.	<NONE>
Generate	1.	kissed this is , you do not the day part look-look we eat . . .
Source	1.	so you like the short hair better .
Target	1.	what ? yeti-i mean danny ?
Generate	1.	the name time that house it said in the same .
Source	1.	can i tell you a little secret ?
Target	1.	yeah !
Generate	1.	no .
Source	1.	i want to keep one .
Target	1.	ohh , i 'm gonna be on the news !
Generate	1.	you know ! you-re-you ? okay . i am check rachel and this is when the . i want to buy . so , if we be the street coffee how is the