Gesture Classification Based on System's Emotion and Strategy for Korean Dialogue Systems

Yeongbeom Lim, Jin Yea Jang, San Kim, Saim Shin and Minyoung Jung*
Artificial Intelligence Research Center
Korea Electronics Technology Institute
Seongnam-si, Korea
{warf34, jinyea.jang, kimsan0622, sishin, minyoung.jung}@keti.re.kr

***Abstract*—During the conversation, the speaker's gestures can give more information to the listener and make it easier to understand the speaker's utterance. Although appropriate system gestures based on the system strategy and emotion are crucial in a conversation between a person and a dialogue system, no Korean conversation corpora have system utterances annotated with both system strategies and emotions. In this paper, we pseudo-annotate system gestures on system utterances in the extended version of the Korean empathetic conversation corpus. Based on the pseudo-annotation, we introduce a system gesture list classification model on the basis of Korean-English T5 (KE-T5). We achieve an accuracy of 90.8% on the gesture list classification model.**

***Index Terms*—gesture classification, emotion classification, Korean dialogues, multimodal dialogues, empathetic dialogues**

I. Introduction

Research on human-computer interaction has been studied for the several decades as human-computer interfaces have evolved. With the advancement of technology and society, people want computers to understand human intelligence and to perform tasks at the same level as human intelligence [1]. The studies are conducted not only in verbal communication methods such as speech or text but also in non-verbal communication methods with faces, hands, or bodies.

[2] argue that agents must express emotions intelligently, not just respond based on user emotions, and propose an agent architecture that can distinguish between felt and expressed emotions, take into account the socio-cultural context, and express them through facial expressions. Table I shows the examples that user utterances are similar and their corresponding system strategies are different. Table II draws examples of different emotions between a user and a system. Both tables explain the reason why system gestures should be considered with system strategies and emotions. Nevertheless, no Korean multimodal dialogue systems [3] have regard to system strategies and emotions for expressing appropriate system gestures.

In this paper, we first pseudo-annotate system gestures on the extended version of the Korean empathetic conversation corpus. With the pseudo-annotated system gestures, we introduce a system gesture list classification model based on Korean-English T5 (KE-T5) [4], a language model pre-trained with Korean and English corpus. We also evaluate the performance of the gesture list classification model.

II. Related Work

Human-computer interaction comprises a variety of communication elements. In this work, we focus on emotions, dialog acts, and gestures.

A. Emotion

Emotion classification is one of the most widely used tasks in natural language processing. The representative emotion datasets are Meld [5] based on the emotion classification definition of [6] and IEMOCAP [7] classifying emotions into nine different categories. For Korean dialogues, [8] classified multi-labeled emotions based on KE-T5 [4].

B. Dialogue Act

Dialogue acts refer to the speaker's intention to convey the speaker's words to the listener. For the conversation between humans and dialogue systems, system strategies and user intentions can be considered dialogue acts. In Korean dialogues, research on user intent classification [9] and system strategy classification [10], [11] has been performed. And the Korean empathetic conversation corpus [12] reclassified dialogue acts into 10 empathetic utterance strategies (7 emotions and 3 response strategies), taking into account empathy levels.

C. Gesture

Gestures are non-verbal communication and are sometimes used in conjunction with verbal communication to reinforce the message being conveyed. [13] finds that there is a greater

TABLE I
Different strategy examples against similar user utterances

No.	Utterance	Emotion	Strategy
1	User: 남자친구한테 가방 선물을 받았어! (My boyfriend gave me a bag as a present!)	Happiness	
	System: 우와, 정말요? (Wow, really?)		Surprise
2	User: 남자친구가 비싼 가방을 사줬어! (My boyfriend bought me an expensive bag!)	Happiness	
	System: 그래요? (Really?)		Back-Channel

TABLE II
Different emotion examples between a user and a system

Utterance	Emotion	Strategy
User: 지들은 평생 병 안 걸리고 살 것 같나보지? (Do they think they will never get sick and live forever?)	Anger	
System: 화가 많이 나시나요? (Are you very angry?)		Clarification

* Corresponding author.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

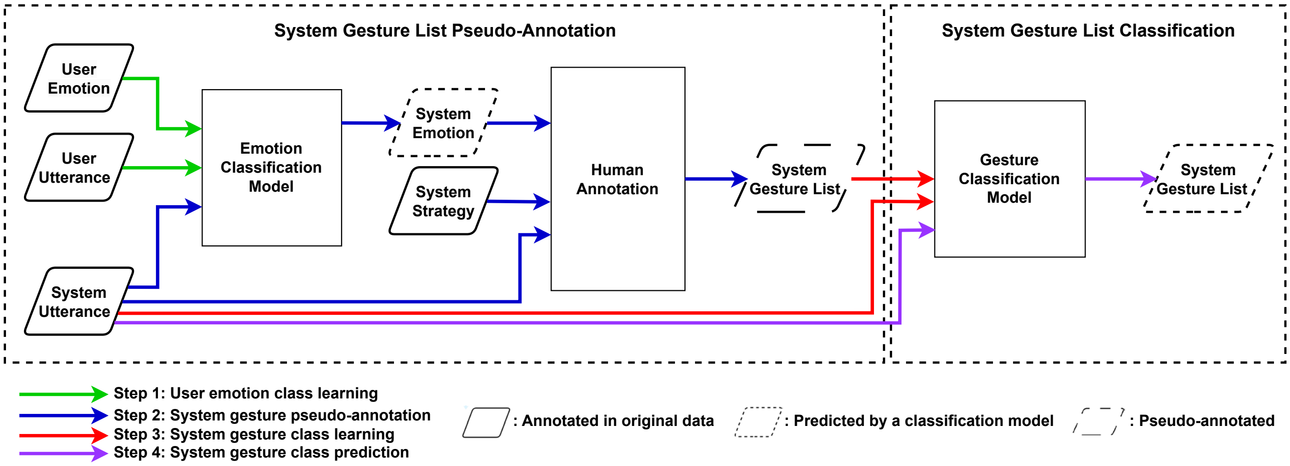
System Gesture List Pseudo-Annotation
System Gesture List Classification
User Emotion
User Utterance
System Utterance
Emotion Classification Model
System Emotion
System Strategy
Human Annotation
System Gesture List
Gesture Classification Model
System Gesture List
Step 1: User emotion class learning
Step 2: System gesture pseudo-annotation
Step 3: System gesture class learning
Step 4: System gesture class prediction
: Annotated in original data
: Predicted by a classification model
: Pseudo-annotated

Fig. 1. Overall architecture of system gesture list pseudo-annotation and classification

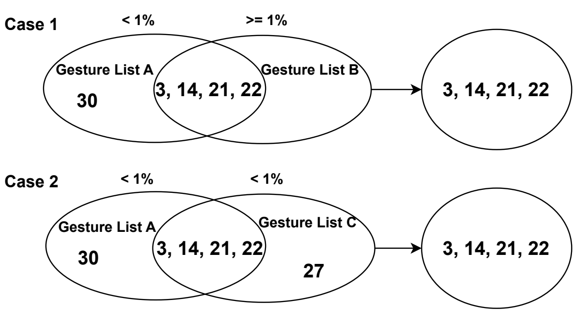

Fig. 2. Pseudo-annotated gesture merging process

learning impact when the avatar uses hand gestures in a math learning animation. The emotion-based Korean multimodal empathetic dialogue system [3] utilized an avatar-based gesture classification module that randomly selects a gesture from the seven different general-purpose avatar gestures.

III. System Gesture List Pseudo-Annotation

We introduce our system gesture list pseudo-annotation and classification approaches for appropriate system gesture classification in Korean multimodal empathetic dialogue systems. As shown in Fig. 1, Steps 1 and 2 utilize an emotion classification model to pseudo-annotate the gestures on system utterances, and Steps 3 and 4 employ a gesture classification model to learn and predict system gestures annotated in Step 2. In this section, we describe our system gesture list pseudo-annotation (Steps 1 and 2). Through the pseudo-annotation, 44,137 system utterances from the extended version of the Korean empathetic conversation corpus [12] are pseudo-annotated with the corresponding system gesture list.

The system and user utterances from the extended version of the Korean empathetic conversation corpus [12] are originally annotated with one of 16 empathetic system strategies (approval, disapproval, back-channel, clarification, echoic response, encouragement, evaluation, facilitation, greeting, opinion, suggestion, surprise, why, what, how, and persona) and one of 7 user emotions (neutral, surprise, happy, sadness, disgust, anger, and fear/anxiety) in each. Because system gestures are related to system emotions and strategies, we propose to predict emotions on the system utterances for system gesture classification.

TABLE III
30 System Gesture Candidates

ID	Gesture Type	ID	Gesture Type
1	Point to left with a hand	16	Thumb up with hands
2	Point to right with a hand	17	Clap hands
3	Point to front with a hand	18	Nod a head
4	Block front with a hand	19	Raise a fist
5	Shake a head	20	Raise fists
6	Cross with hands	21	Put a hand on chest
7	Tilt a head with a finger	22	Put hands on chest
8	Tilt a head with fingers	23	Hold hands in front of chest
9	Cross arms	24	Put a hand in front of mouth
10	Shrug shoulders	25	Point to the front with a finger
11	Bow a head	26	Raise a finger up
12	Wave a hand	27	Clap a palm with a fist
13	Wave hands	28	Scratch a head
14	Point to front with hands	29	Put hands together
15	Thumb up with a hand	30	Shake a head with hands

In consideration of tremendous human annotation effort, we utilize the emotion classification model [8] by re-training the model with user utterances and emotions in the extended version of the Korean empathetic conversation corpus [12], to perform emotion pseudo-annotation on system utterances. System utterances with pseudo-annotated emotions and originally human-annotated strategies can be categorized into 112 [system emotion]-[system strategy] pairs, combining 7 emotions and 16 strategies. After extracting a maximum of 50 random utterances for each emotion-strategy pair, a human annotator performs multi-annotation among 30 system gesture candidates in Table III, based on the system utterance, emotion, and strategy information.

After the multi-annotation process, a group of multi-annotated gestures is called a gesture list, and we merge some gesture lists whose number of data is relatively small, to reduce the number of gesture list candidates. Case 1 in Fig. 2 shows the process of merging a gesture list ($<$ 1% of the total data) and the other gesture list ($\geq$ 1% of the total data). For example, the gesture list A ($<$ 1% of the total data) is compared to all other gesture lists. The gesture list A and the other gesture list B are merged in such a way that the difference set (A-B) is minimized. Case 2 in Fig. 2 draws the process of merging two

TABLE IV
21 GESTURE LISTS

List ID	Gesture IDs
1	15, 16, 17, 18, 23, 27, 29
2	18, 23, 29
3	4, 5, 6, 23, 28, 29
4	18, 23, 26, 27, 29
5	18, 19, 20, 23, 29
6	12, 13, 15, 16, 18, 23, 26, 27, 29
7	11, 12, 13, 23, 29
8	10, 23, 24, 29
9	12, 13, 18, 23, 29
10	7, 10, 18, 23, 29
11	18, 23, 28, 29
12	18, 23, 27, 29
13	18, 23, 24, 27, 28, 29
14	7, 10, 18, 23, 26, 27, 29
15	15, 16, 18, 19, 20, 23, 26, 27, 29
16	12, 13, 18, 23, 26, 29
17	12, 13, 15, 16, 17, 23, 26, 29
18	18, 23, 24, 27, 29
19	18, 23, 24, 29
20	24, 27
21	3, 14, 21, 22

TABLE V
ACCURACY OF EMOTION AND GESTURE LIST CLASSIFICATION MODELS

	Emotion Classification	Gesture List Classification
Accuracy	67.8%	90.8%

gesture lists ($< 1\%$ of the total data). Consequently, all system utterances with the same emotion and strategy are annotated with the corresponding human-annotated gesture list among 21 system gesture lists in Table IV.

IV. System Gesture List Classification

We discuss our system gesture list classification (Steps 3 and 4) in Fig. 1. Our KE-T5 [4] based gesture list classification model is designed to learn pair data of the system utterance and the gesture list, achieved from the system gesture list pseudo-annotation. KE-T5 is a Korean pre-training model learned from Korean and English corpora, based on the architecture of Google's T5 model [14]. In Step 3, the gesture classification model learns the system utterances and the gesture lists generated from Step 2. Given the system utterance, Step 4 shows that the trained gesture classification model predicts its system gesture list.

V. Experiments

We conduct experiments for both the emotion and gesture list classification models introduced in Fig. 1. Both models were trained with a learning rate of 0.001 and a dropout ratio of 0.1, and the batch size is set to 64. The total number of system utterances is 44,137 and is divided into a ratio of 6:2:2 for learning, validation, and testing respectively.

As shown in Table. V, the emotion classification model utilized for system gesture list pseudo-annotation exhibits an accuracy of 67.8%. And the gesture list classification model achieves an accuracy of 90.8%. In contrast to the accuracy of the emotion classification model, the gesture list classification model shows a comparatively high accuracy. This observation may be due to the relatively limited pattern of system utterances and the influence of human annotation.

VI. Conclusions

In this paper, we pseudo-annotate gestures on system utterances in the extended version of the Korean empathetic conversation corpus by considering system utterances, emotions, and strategies. We re-train the KE-T5-based emotion classification model whose accuracy is 67.8%, and introduce the KE-T5-based system gesture list classification model whose accuracy is 90.8%.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions).

References

[1] A. L. Guzman and S. C. Lewis, "Artificial intelligence and communication: A human–machine communication research agenda," New media & society, vol. 22, no. 1, pp. 70–86, 2020.

[2] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, "Intelligent expressions of emotions," in Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1. Springer, 2005, pp. 707–714.

[3] M. Jung, Y. Lim, S. Kim, J. Y. Jang, S. Shin, and K.-H. Lee, "An emotion-based Korean multimodal empathetic dialogue system," in Proceedings of the Second Workshop on When Creative AI Meets Conversational AI. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 16–22.

[4] S. Kim, J. Y. Jang, M. Jung, and S. Shin, "A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems," in Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 352–365.

[5] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," arXiv preprint arXiv:1810.02508, 2018.

[6] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3-4, pp. 169–200, 1992.

[7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, pp. 335–359, 2008.

[8] Y. Lim, S. Kim, J. Y. Jang, S. Shin, and M. Jung, "Ke-t5-based text emotion classification in korean conversations," in HCLT, Oct. 2021, pp. 496–497.

[9] M. Jung, J. Kim, J. Y. Jang, H. Jung, and S. Shin, "A performance comparison among different amounts of context on deep learning based intent classification models," in 2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020, pp. 1336–1338.

[10] J. Y. Jang, J. Kim, M. Jung, H. Jung, and S. Shin, "Utilizing multi-modal emotion information in dialogue strategy classification," in 2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020, pp. 943–945.

[11] J. Y. Jang, M. Jung, H. Jung, and S. Shin, "Dialogue strategy prediction using multi-modal information for empathic dialogue generation," in ICONI, Dec. 2020.

[12] J. H. Yang, J. Y. Jang, M. Jung, and S. Shin, "Establishing a corpus for an ai-based empathic response system," in ICONI, Dec. 2020.

[13] S. W. Cook, H. S. Friedman, K. A. Duggan, J. Cui, and V. Popescu, "Hand gesture and mathematics learning: lessons from an avatar," Cognitive science, vol. 41, no. 2, pp. 518–535, 2017.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020.