IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 12, 2021, accepted November 7, 2021, date of publication November 15, 2021,
date of current version November 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3128277

Utilizing External Knowledge to Enhance Semantics in Emotion Detection in Conversation

FUJI REN, (Senior Member, IEEE), AND TIANHAO SHE
Faculty of Engineering, Tokushima University, Tokushima 770-8506, Japan

Corresponding authors: Fuji Ren (ren@is.tokushima-u.ac.jp) and Tianhao She (zthzth2080@yahoo.co.jp)

This work was supported in part by the Research Clusters Program of Tokushima University under Grant 2003002.

ABSTRACT Enabling machines to emotion recognition in conversation is challenging, mainly because the information in human dialogue innately conveys emotions by long-term experience, abundant knowledge, context, and the intricate patterns between the affective states. We address the task of emotion recognition in conversations using external knowledge to enhance semantics. We propose KES model, a new framework that incorporates different elements of external knowledge and conversational semantic role labeling, where build upon them to learn interactions between interlocutors participating in a conversation. We design a self-attention layer specialized for enhanced semantic text features with external commonsense knowledge. Then, two different networks composed of LSTM are responsible for tracking individual internal state and context external state. In addition, the proposed model has experimented on three datasets in emotion detection in conversation. The experimental results show that our model outperforms the state-of-the-art approaches on most of the tested datasets.

INDEX TERMS Affective computing, text emotion classification, emotion recognition in conversation.

I. INTRODUCTION

In recent years, due to the development of deep learning, enhanced learning, and the construction of a large dataset for dialogue, the research on emotion recognition in conversation (ERC) has received attention. ERC is emerging research, which aims to detect the emotion expressed in discourse in the dialogue between two or more interlocutors. The task is important to research in several areas, such as affective dialogue systems [1]–[3], healthcare [4], [5] recommendation system [6], and so on.

For a long time, whether there is an emotion or not is one of the most essential to distinguish between human and machine. In other words, whether the machine has emotion is also one of the key factors for the degree of humanization of the machine [7].

As a comprehensive technology, emotional computing is a key step of artificial intelligence's emotionalization, including emotion recognition, expression, and decision [8], [9]. ''Recognition'' is to let the machine accurately recognize human emotions and eliminate uncertainty and ambiguity; ''Expression'' means that artificial intelligence expresses emotions with suitable information carriers, such as language, sound, posture, and expression; ''Decision-making'' mainly studies how to use emotional mechanism to make better decisions [10].

Recognition and expression are two key technical links in emotional computing [11]. Emotion recognition can extract the features of emotion signals and get the emotion feature data that can represent human emotions to the maximum extent. Based on this model, the mapping relationship between the external representation data of emotion and the internal emotional state is found out, and then the current internal emotion types of human beings are identified, including voice emotion recognition, facial expression recognition, and physiological signal emotion recognition.

Different from other emotion recognition tasks, conversational emotion recognition is not only for sentences/utterances, but also depends on the context and the state of participants for modeling [12]. Natural dialogues are complex, and they are governed by many variables, which depend on the time sequence of discourse and affect the emotional dynamics of participants. These variables include topic, argumentation logic, intention, interlocutors' personality, and so on [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

Recent works on ERC use recurrent neural networks (RNNs) to model the utterance [13], [14], which relies on spreading context and order information to utterance. RNNs, such as long short-term memory (LSTM) [15] and gated recurrent unit (GRU) [16], are used to simulate the dependence between utterances, so as to perceive the context at speaker level or situation level. In theory, such a network should be able to transmit long-term contextual information, but in practice, it is not always the case. To mitigate this issue, graph-based methods are introduced [17], which consider distant utterances and sequence information by time coding utterances. However, graph-based methods tend to use information from recent utterances that are relatively limited to update the status of query utterances, which makes it difficult for them to obtain satisfactory performance.

Although the method of the above genres has made well progress, we argue that current techniques which only focus on context modeling at the present limit the ability of language representation. The main limitation of these methods is that they only consider simple contextual features as representation and training objectives, and seldom consider well-defined contextual semantic clues. The common knowledge of conversation plays an indelible role in inferring the potential variables of a conversation. Even though a well-trained language model can express context semantics more or less implicitly, the introduction of a common sense-oriented framework can further enhance this point [18].

Most studies have found that deep learning Frameworks might not really understand the semantics of the natural language [19] and vulnerably suffer from adversarial attacks [20]. Deep learning models often ignore important words and choose relatively safe and unimportant ones. Briefly, semantic role labeling (SRL) [21] aims to restore the predicate-argument structure of sentences, and fundamentally discover change from who did what to whom, when and why did what as the central meaning of sentences, which naturally matches the task goal target ERC. In the conversation task, the conversation content of the participants usually involves various predicate-argument structures. A predicate is a statement or explanation of the subject, pointing out ''what to do'' or ''how to do,'' which represents the core of an event, and the question formed with who, what, how, when, and why can be conveniently formalized into the predicate-argument relationship in terms of contextual semantics. Motivated by these, this paper attempts to integrate extra SRL knowledge into the pre-trained model, and model the context and the emotional state of the participants to understand the conversation context.

In this paper, we propose a KES for emotion recognition in conversation. By introducing SRL information and clear context semantic clues into the pre-training language model, the algorithm enriches the sentence context semantics in multiple predicate-specific argument sequences. The proposed model incorporates the concept of external knowledge and SRL information, at the same time, it learns representation in a fine-grained manner on plain context representation and explicit semantics in order to achieve deeper meaning representation. Through an individual internal state encoder, our model tracks and predicts the speaker's continuous emotional self-dependence. The information reflecting the contextual state of context and the speaker's influence and dependence on others is encoded and processed by the global state, so that the proposed model can understand the contextual information and the emotional transfer among the participants in the conversation. We conduct extensive experiments on three different conversation corpus and comparisons with several baseline models. The results show that the proposed method achieves comparable performance with the state-of-the-art models. To sum up, the main contributions in this research are summarized as below:

- We introduce SRL information to enrich the semantic structure in conversation, and obtain commonsense knowledge from external knowledge graph to promote emotion detection in conversation.
- We design the attention mechanism to integrate external commonsense knowledge and conversation level SRL information, and utilize a transformer structure to replace the recurrent attention neural networks commonly used for emotion detection in conversation.
- We conduct extensive experiments, which prove that SRL information and commonsense knowledge are beneficial to the performance of emotion detection. The proposed model KES is superior to the state-of-the-art models in most test datasets.

The rest of the paper is organized as follows: Section II discusses related works. Section III shows an overview of the proposed method. Section IV introduces the relevant situation of the experiment. Section V provides experimental results and analysis of Section IV. Section VI presents our conclusions and future work.

II. RELATED WORK

A. CATEGORIZATION OF EMOTIONS

Emotion recognition has been an active research field for many years and has been explored in interdisciplinary fields such as machine learning, signal processing, social and cognitive psychology, etc [22].

The discrete emotion model uses adjective labels to express emotion. Simply divides discrete emotions into two basic emotions: pain and happiness [23]; Furthermore, Ekman [24] concludes six basic emotions: anger, disgust, fear, happiness, sadness and surprise. Although there are more than 10 emotion description models used within the field, Ekman's six basic emotions and seven basic emotions added with neutrality are the most used ones [25]. The discrete emotion model is simple, intuitive and widely used, but its description accuracy is not high, its continuity is not good, and the emotions that can be represented by the model are limited. Scholars established dimensional categorization models to overcome these deficiencies instead.

Most dimensional categorization models [26], [27] adopt two dimensions: valence and arousal. Valence represents the degree of emotion, and arousal represents the intensity of emotion. The commonly used ERC dataset IEMOCAP [28] is compatible with classification and dimension models. However, some later ERC datasets like DailyDialogue [29] can only be used for classification models. Most datasets use simple classifications, similar to variations of Ekman. Each emotional utterance in the EmoContext dataset is labeled as one of the following emotions: sadness, happiness and anger. Annotators of the EmoContext dataset are more consistent compare because its emotion classification is very simple. However, the short context length and simple sentiment classification make the ERC task less challenging on this dataset.

B. EMOTION RECOGNITION IN CONVERSATION

ERC is a research topic under the spotlight in the field of natural language processing in recent years, which has potential significance in many fields, such as healthcare, opinion mining, education, recommendation system, and so on. Along with the open-source of numerous conversation datasets that include acoustic, textual, and visual features, many deep learning methods have been implemented in conversational emotion recognition tasks.

Numerous efforts have been devoted to the modeling of conversation context. Basically, they can be divided into two categories: graph-based methods and recurrence-based methods [48]. For graph-based methods, they collect the information of surrounding utterances in a specific window at the same time, while ignoring distant utterances and sequential information. For recurrence-based methods, they consider distant utterances and continuous information by time coding the utterances. However, they tend to use only the relatively limited information from recent utterances to update the status of query utterances, which makes it difficult for them to obtain satisfactory performance. DialogueRNN [14] proposes to model dynamic emotion through GRU, and capture context information by using an attention mechanism. DialogueGCN [17] applies graphs to model conversations and regards both speakers and utterances as graph nodes, thus solving the problem of context propagation in DialogueRNN. HiGRU [49] contains two GRUs, one for the utterance encoder and the other for the conversation encoder. Zhong *et al.* [45] propose KnowledgeEnriched Transformer (KET), which learns structured conversation representation through layered self-attention and external common knowledge. DialogXL [46] improves XLNet [47] enhanced memory to store longer historical context and dialog-aware self-attention to deal with the multi-party structures. COSMIC [18] proposes a neural network structure framework that introduces external knowledge which introduces external knowledge to improve performance by establishing a huge knowledge base. Different from the existing methods, we propose a model to enhance the knowledge of external commonsense built on a Transformer Encoder-Decoder structure for conversation emotion detection.

C. CONVERSATIONAL SEMANTIC ROLE LABELING

Semantic role labeling systems aim to recover the predicate-argument structures of sentences – basically to determine ''who did what to whom,'' ''when,'' and ''where.'' Traditional SRL often failed to analyze conversations since only a single utterance can be analyzed by traditional SRL whereas ellipsis and anaphora occur in conversation as well. Conversation semantic role labeling (CSRL) [30] directly models the predicate-argument structure of the whole conversation instead of a single utterance. Most of the discarded or referenced components in the latest conversation can actually be found in the conversation history. CSRL allows arguments to be indifferent utterances as the predicate, while SRL can only work on every single utterance. Compared with the standard SRL, which needs utterance rewriting or co-referential parsing as the preprocessing step of analyzing dialogues, CSRL can directly deal with conversation and avoid error propagation.

D. KNOWLEDGE ENHANCED LANGUAGE

Researchers have paid more attention to enhancing natural language models with knowledge graphs these days, since knowledge graph has gained a lot of systematic knowledge. Liu *et al.* [31] combined the knowledge triple in the knowledge graph with the original text and then modeled it with BERT to get more hidden information. Lin *et al.* [32] proposed a knowledge-aware graph network model based on a graph convolution network, which has a path-based attention mechanism. Zhang *et al.* [33] combined entity information with BERT to enhance language representation, which can utilize vocabulary, syntax and knowledge information at the same time. More recently, Bosselut *et al.* [34] proposed COMmonsEnse Transformers (COMET), which learned to generate commonsense descriptions in natural language by fine-tuning the pre-trained language model in ATOMIC knowledge base. Compared with the extraction method, the fine-tuned language model in the knowledge base has unique advantages of generating knowledge for invisible events, These advantages are very important for tasks that need to combine common knowledge in conversation. Ghosal *et al.* [18] proposed COmmonSense knowledge for eMotion Identification in Conversations (COSMIC) model based on DialogueRNN structure, which uses external commonsense knowledge generated by COMET and obtains advanced results in ERC. Compared with the extraction method, the fine-tuned language model in the knowledge base has the unique advantage of generating knowledge for invisible events, which is very important for tasks that need to combine common knowledge such as emotion detection in conversation.

The ERC task in our proposed method also extracts external commonsense knowledge information based on a knowledge graph. Different from the COSMIC model, we use

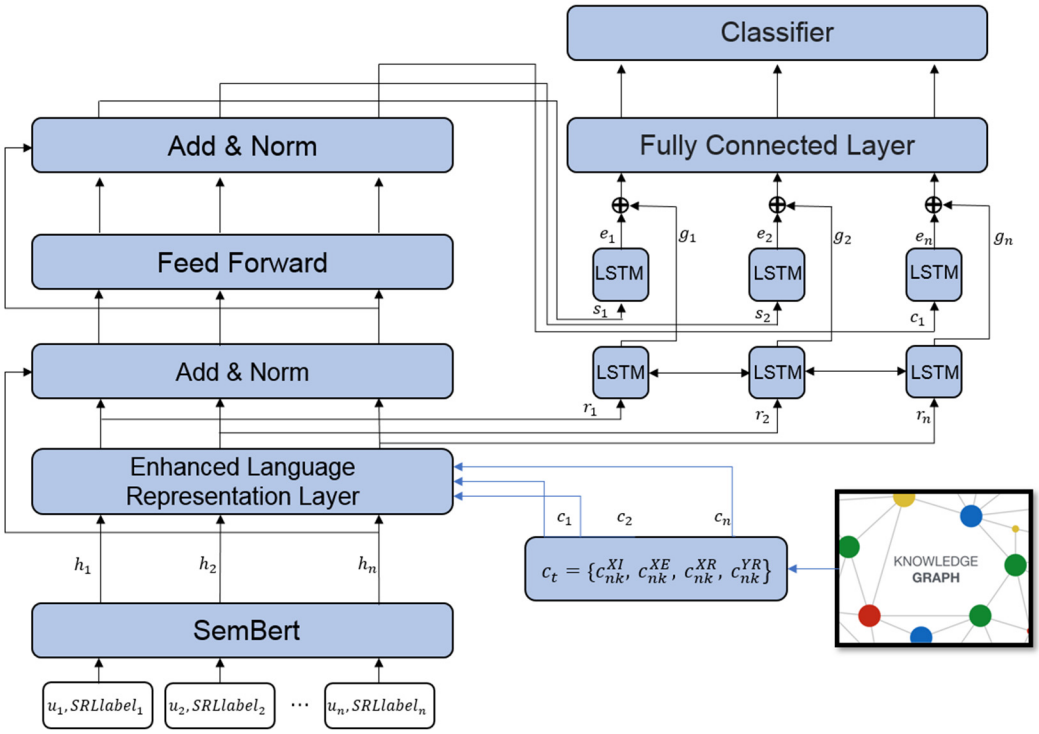

FIGURE 1. Overview of KES. Enhanced Language Representation Layer takes the textual feature u_t of the tth utterance and its SRL labels in conversation as input and generates code the enhanced semantic feature h_t, which is input to the transformer structure network. We design the first layer self-attention mechanism of the transformer structure to integrate the candidate knowledge, and h_t and external commonsense knowledge c_t as the input of this attention mechanism to generate r_t. The transformer structure network generates encoding s_t and inputs it into the individual internal encoder LSTM to generate e_t. Also, r_t is fed into the individual context encoder BiLSTM to outputs another encoding g_t. *Each e_t and g_t to obtain the final prediction for each utterance in the conversation.*

BERT to obtain external knowledge instead of modeling the context in the conversation based on the traditional RNN network structure. The framework structure of the proposed model is reported in Section III.

III. METHODOLOGY

A. TASK DEFINITION

ERC task aims to recognize the emotion of each utterance from several predefined emotions within/among the provided conversation records and participant information of each utterance in a conversation emotion recognition task. In a conversation between two people, each utterance is marked by latent emotion. Formally, let there be M participant/parties $\{p_1, p_2, \ldots, p_M\}$ in a conversation, which is defined as a sequence of utterances $\{u_t = u_1, u_2, \ldots, u_N\}$, where N is the number of utterances. The task is to predict the emotion labels (happy, sad, frustrated, excited, angry, and neutral) of the constituent utterances u_t, where utterance u_t is uttered by participant $p_{s(u_t)}$, where s representing the mapping between utterance and index of its corresponding participant. In related research, the traditional method is to first produce context-independent representations by pre-trained language model and then perform context modeling to classify each of the constituting utterances into its appropriate emotion category.

B. UTTERANCE FEATURE EXTRACTION

He *et al.* [21] presented a deep highway BiLSTM architecture with constrained decoding, which is simple and effective. In the practice of data preprocessing, each utterance is annotated into several semantic sequences by the pre-trained semantic annotator.

The original utterance sequence and semantic role tag sequences are expressed as embedding vectors to feed a pre-trained BERT. The input utterance is a sequence of words of length n, which is first labeled as a word fragment. Then, the transformer captures the context information of each token through self-attention and generates a sequence of context embeddings. For m SRL label sequences associated with each predicate, we have $\{t_1, t_2, \ldots, t_m\}$ where t_i contains n labels denoted as $\{\text{label}_1^i, \text{label}_2^i \ldots, \text{label}_n^i\}$. We employ SemBert [35] to extract semantically enhanced language representation by SRL semantic sequences, and fine-tune SemBert large-scale emotional label prediction model from the transcript of the utterances. It can be trained jointly with KES, so its gradient will be updated in the whole building training process. It can also be trained as an individual task of discourse classification with emotional labels. However, related experiments show that although the traditional SRL system (even with the help of common reference parsing or rewriting) does not perform well in analyzing conversation,

modeling the conversation history and participants is of great help to the performance, which indicates that adapting SRL to conversations is very promising for general conversation understanding. Therefore, we introduce the concept of external knowledge to try to further extract the missing features in the conversation.

C. KNOWLEDGE FEATURE EXTRACTION

In this work, we employ an external knowledge graph ATOMIC [36] to extract the knowledge sources. ATOMIC is a collection of if-then common knowledge that describes the daily reasoning of an organization through text. It is composed of nine different types of if-then relationships to distinguish between agents and themes, causes and effects, voluntary and non-voluntary events, and actions and mental states. Due to the expressiveness of events and the improved relationship type, ATOMIC is used in the If-Then reasoning task to achieve the result competing with human evaluation. Items with weights less than the threshold or containing words that are not in the selected vocabulary will be removed from the knowledge graphs. Items are triples with the form {subject, relation, object}. Given an event in which the speaker participates, the 9 relation types are inferred as follows: intent of speaker, need of speaker, attribute of speaker, effect on speaker, wanted by speaker, reaction of speaker, effect on others, wanted by others and reaction of others. For example, given an event or topic phrase: ''PersonX puts PersonY in touch.,'' from ATOMIC's inference of relation phrases, PersonX's intention and reaction of others would be ''PersonX want to keep the relationship'' and ''others want to express gratitude,'' respectively. There are a total of 9 relation types, of which four are used: the intent of speaker (denoted as XI), effect on speaker (denoted as XE), the reaction of speaker (denoted as XR), and reaction of others (YR).

Given an utterance u_t, we can compare it with each node in the knowledge graph and retrieve the most similar one. Each utterance u_t is annotated with a part-of-speech (POS) tag by NLTK [37]. Usually, nouns, adjectives, and verbs with parts of speech contain more information than other tokens. Therefore, the items related to them are searched preferentially in the knowledge map. In all the chosen items, we extract the top K events, and obtain their intentions and reactions. We employ BERT [38] calculation to capture the causes between two sequences, and the last hidden state is taken as the output, which is denoted as $c_t = \{c_{nk}^{XI}, c_{nk}^{XE}, c_{nk}^{XR}, c_{nk}^{YR}\}$, $k = 1, 2 \ldots, K$.

D. MODEL

It is crucial to consider contextual information when classifying discourse in a sequence, since other discourses in the sequence have a great influence on the emotion of the current discourse. In other words, the dependence between speakers is important for the emotional dynamics in a conversation. For example, the current speaker's emotions can be changed by the other's words, and it is crucial to consider context information for simulating the emotional dynamics in a conversation.

The conversation is a sequence of coherent and orderly discourses. For neural networks, the capture of long-range context information is a weakness. We adopt Transformer [39] a structure composed of self-attention and feed forward neural network, instead of the traditional RNN model, aiming at capturing remote context information.

1) KNOWLEDGE ENHANCED LANGUAGE REPRESENTATION LAYER

With the knowledge source extracted, the commonsense features from the knowledge graph are obtained. We design an attention mechanism to integrate the candidate knowledge in Transformer. We modified the structure of the transformer so that it separately encodes the internal state information of the individual in the conversation and the contextual global state information. The Attention mechanism in the encoder is used to fuse different common sense and knowledge information, and integrate and filter effective information into the connected layer. A conversation consists of N utterances $\{u_1, u_2, \ldots, u_n\}$, in which M distinct participants $\{p_1, p_2, \ldots, p_M\}$ take part. For every $t \in \{1, 2 \ldots, \mathrm{N}\}$, we apply the enhanced semantic features h_t obtained by fusing Bert feature vector and SRL information to generate new knowledge representation. The attention mechanism is used to refine the result of each knowledge source and aggregate the representation of each feature as follows:

$$v_t = \tanh\left([c_t, h_t]\,\mathbf{W}_\alpha\right), \tag{1}$$

$$\alpha_t = \frac{\exp(v_t v_t^T)}{\sum_{k=1}^{K} \exp(v_t v_t^T)} \tag{2}$$

$$r_t = \sum_{k=1}^{K} \alpha_t c_t. \tag{3}$$

We further integrate the knowledge feature through the self-attention mechanism and the final event representation is denoted r_t.

The context encoder takes enhanced textual features and knowledge features of utterances as input and applies multi-head self-attention attention operation to it by a feed-forward layer which is completely concatenated point by point, so as to generate the contextualized cultural vectors of utterances:

$$s_t = FFN(r_t). \tag{4}$$

2) CONTEXT STATE

It is essential to consider the contextual information when classifying the emotions of a conversation discourse in a sequence since other discourse information in the sequence has a great influence on the emotions of the current discourse, and the contextual state stores and transmits the information of the whole utterance-level along the sequence of the conversation flow. In other words, the current speaker's mood will be forced to change by the other's words. This fact reflects the dependence between speakers, which is closely related to the tendency of speakers to imitate each other in

conversation [32] and is important for simulating the emotional dynamics in a conversation. Our work is not only to extract the emotional influence information between speakers from the knowledge graph but also to encode the context state of the current discourse to obtain the dynamic emotional changes [50]. In view of the sequence of conversation, we use BiLSTM to capture the contextualized cultural vectors. The upper and lower cultural vectors generated by the Transformer structure are fed to the BiLSTM layer, and the BiLSTM layer fuses remote sequential contextual information to generate context coding.

Finally, the contextualized feature representation is input into BiLSTM, and context feature is obtained:

$$g_t = \text{BiLSTM}_t\left(r_t\right). \tag{5}$$

3) INDIVIDUAL INTERNAL STATE

The individual internal state tracks each utterance in the conversation, which reflects the speaker's emotional influence on himself in the conversation. The individual internal state of participants depends on their feelings and the effects they feel from other participants. Participants may not always express their feelings or opinions clearly through external positions or reactions. This state can also be considered to include aspects that participants actively try not to express or features that are considered common sense and do not need explicit communication. Under the influence of emotional inertia, every speaker in the conversation tends to keep a stable emotional state until the other person causes changes.

We model the individual internal state of the participants using LSTM, which is the internal encoder to output all speaker states for timestep t. It exploits the currently integrated knowledge discourse representation to update the state of the corresponding speaker:

$$e_t = \text{LSTM}_t\left(s_t\right). \tag{6}$$

In time-step t, the output of each LSTM corresponds to the speaker and is updated by the knowledge discourse representation r_t of the current utterance u_t.

4) EMOTION CLASSIFICATION

Finally, we connect the global feature vectors generated by the context state and the internal feature vectors generated by the individual internal state, calculate the probabilities of six emotion-class and select the most possible emotion class.

$$P_t = \text{softmax}\left(W_{smax}\left(g_t \oplus e_t\right) + b_{smax}\right), \tag{7}$$

$$\hat{y}_t = \underset{i}{\text{argmax}}\, P_t\left[i\right]. \tag{8}$$

IV. EXPERIMENTAL SETUP

For ease of comparison with state-of-the-art methods, we evaluate our model on three benchmark datasets: IEMOCAP [28], DailyDialogue [29], and MELD [40], and mention their properties. Further, our report summarizes the experimental results of conversational emotion recognition from the text information of all three benchmark data sets.

A. DATASETS

Information about the datasets is shown in Table 1.

IEMOCAP is a multimodal ERC dataset, which contains videos of two-way conversations of ten unique speakers. The trainset conversations come from the first eight speakers, whereas the testset conversations are from the last two. Each video in IEMOCAP contains a single dyadic conversation from the performance based on a script by two actors. Each discourse has 2476 annotations, with one of the following six emotions: happiness, sadness, neutrality, anger, excitement and depression.

DailyDialogue is a human-written dyadic conversation dataset from daily communications. DailyDialogue takes the Ekman's six emotion types [24] as the annotation protocol and reflects daily communication way and covers various topics about human daily life. The emotion can belong to one of the following seven labels: anger, disgust, fear, joy, neutral, sadness, and surprise. The dataset contains more than 83% neutral emotion labels, which were excluded during the evaluation of Micro F1.

MELD is a multi-modal ERC dataset extended from Emotionlines dataset [41]. MELD is constructed from the script of the urban life TV series Friends, which contains more than 1400 dialogues and 13000 words of contains textual, acoustic, and visual information. Each utterance has seven emotional labels, including neutrality, happiness, surprise, sadness, anger, disgust and fear.

TABLE 1. The statistics of three datasets.

Dataset	# Conversations			# Uterrances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
DailyDialogue	11118	1000	1000	87170	8069	7740
MELD	1038	114	280	9989	1109	2610

B. BASELINES

To comprehensively evaluate the proposed model KES, we use the following methods to compare its performance:

1) CNN [42]

CNN is a convolutional neural network model, which is trained on the basis of pre-trained GloVe [43]. It is the only baseline model without modeling contextual information.

2) CNN + cLSTM [44]

CNN is used to extract Textual features, and LSTM is used to model context information based on CNN.

3) DialogueRNN [14]

An RNN-based method uses the speaker, context, and emotion information from adjacent utterances to model the emotion of utterance in conversation.

In this model, CNN is used to extract text features, and independent GRU networks are used to model speaker state and contextual information respectively.

TABLE 2. Coverall performance on the three datasets. best performances are highlighted in bold.

Model	IEMOCAP							DailyDialogue	MELD
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-Avg F1	Macro F1	W-Avg F1
CNN	35.34	53.66	51.61	62.17	50.66	55.56	51.28	36.87	55.02
CNN+cLSTM	33.90	69.76	48.40	57.55	62.37	57.64	56.04	51.84	56.87
DialogueRNN	37.94	78.08	58.95	**64.86**	68.11	58.85	62.26	41.80	57.03
DialogueGCN	42.75	84.54	63.54	64.19	63.08	**66.99**	64.18	49.95	58.10
KET	-	-	-	-	-	-	59.56	-	58.18
COSMIC	-	-	-	-	-	-	65.28	51.05	65.21
KES	**47.74**	**84.63**	**64.25**	62.48	**73.26**	63.48	**66.32**	**52.28**	**66.46**

4) DialogueGCN [17]

DialogueGCN creates a graph based on the interaction to take into account the conversation structure between the participants of speakers. A Graph Convolutional Network (GCN) is employed to encode the speakers. Contextual features and speaker-level features are connected, and attention mechanism based on similarity is used to obtain the final classified discourse representation.

5) KET [45]

KET is the first model which integrates the external commonsense knowledge and emotion information in emotion lexicon into conversation text feature. The model uses the Transformer decoder to predict the emotional label of the target utterance.

6) COSMIC [18]

In this model, COMET [34] is used to retrieve commonsense knowledge of event eccentricity from ATOMIC [36]. Based on DialogueRNN structure, this model is applied to ERC tasks with common sense knowledge and has achieved advanced results.

V. RESULTS AND ANALYSIS

A. COMPARISON WITH BASELINES

We compare the performance of the proposed model KES with the baseline method, and the experimental results are shown in Table 2. The overall results of all the compared methods on the three datasets are reported. We can note from the results that the proposed KES model competitive performances across the three datasets and reaches a new state-of-the-art on the IEMOCAP, DailyDialogue, and MELD datasets.

1) IEMOCAP

IEMOCAP datasets contain binary conversations with natural and coherent discourses. Since the six emotional tags in IEMOCAP are unbalanced, the F1 score of a single label is also reported. The new and most advanced F1 score of KES model is IEMOCAP with 66.32. We observe that the proposed model is around 4% better than DialogueRNN, 2% better than DialogueGCN. As for the models which also adopt the method of quoting the concept of external commonsense knowledge, the proposed model is around 7% better than KET, 1% better than COSMIC. For the model based on CNN and LSTM, there is a performance gap of more than 10%. One of the main reasons of this large performance gap is that some models, such as CNN, CNN + cLSTM and KET, ignore the speaker-level information modeling, which makes the models treat different speakers equally, resulting in a certain loss of performance.

TABLE 3. Ablation results on three datasets.

Method	IEMOCAP	DailyDialogue	MELD
KES	66.32	52.28	66.46
w/o SRL	64.86	51.03	65.78
w/o KG	61.54	50.24	63.96

Considering that the average utterance length in IEMOCAP is more than 50, and even the maximum session length is more than 100, the Transformer can capture remote dependencies better than RNNs-based context encoders. Moreover, the conversational semantic role labeling information clarifies the utterance structure and concentrates the conversation information. Besides, External knowledge also enriches semantic information, makes the conversation context more closely related, and expresses the emotional influence of the speaker on himself and other conversation participants.

2) DailyDialogue

In the DailyDialogue dataset, neutral emotion accounts for more than 80% of the test dataset. Because of the unbalanced data distribution, we use the macro F1 score excluding the neutral class as the evaluation index. DailyDialogue dataset contains many short utterances with an average length of 8. In this case, using a speaker encoder to model speaker-level information can release more capabilities to improve performance. According to Xu *et al.* [30], conversational semantic role labeling information is more sensitive to the information of single sentence dialogue, and it is easier to find the central semantic information fundamentally.

TABLE 4. Case study from the IEMOCAP dataset.

Utterances	Commonsense knowledge	Emotion
A: I'm just so tired all the time.	XI: rest, XE: gets hurt, XR: exhausted, YR: -	Sad
B: Have you been trying to get a job?	XI: to be supportive, XE: gets yelled at, XR: proud, YR: becomes annoyed with B	Neutral
A: I've been looking for like eight months.	XI: -, XE: gives up, XR: worried, YR: regret	Frustrated
B: It's really hard to find a job.	XI: to be better, XE: accomplish, XR: worried, YR: regret	Frustrated
A: I'm tired of the same excuses.	XI: to escape from responsibility, XE: lies, XR: stupid, YR: disappointed	Frustrated
B: Well, okay	XI: -, XE: okay, XR: relieved, YR: -	Neutral

3) MELD

MELD dataset consists of multiparty conversations, and We follow the same metrics used on the IEMOCAP dataset. Utterances in MELD are much shorter compares to other datasets, and rarely contain emotion-specific expressions, which means that emotion modeling is highly dependent on context. Many dialog scenes contain conversations of more than five speakers, but the average conversation length is only 10 and the minimum length is only 1, which means that emotion modeling is highly dependent on context. Most participants in MELD conversation have only a few words. It is difficult to build a self-reliance model for short conversations. The advantages of transformer over RNNs in capturing the dependence between long-distance speakers are not obvious. Additionally, the discourse in MELD lacks specific emotional expression, which further increases the difficulty of emotional modeling. Nevertheless, the proposed model achieves better results than other baselines, because of the screening and fusion of commonsense knowledge, the utterances in the conversation show more emotional connection in the dialogue context.

B. THE ROLE OF EXTERNAL KNOWLEDGE

In Table 3, we also report the results of ablation studies by removal of various components from the proposed model KES.

It can be observed that the performance of KES continues to decline in all datasets. In IEMOCAP dataset, compared with all other datasets, we observed a more severe performance decline without using an external knowledge graph (KG), and the weighted macro F1 dropped by nearly 5%. This might be attributed to considering that the average conversation length in IEMOCAP is at least 50, it is difficult to grasp the core meaning that affects emotion in the absence of commonsense knowledge. It confirms the importance of using commonsense knowledge to identify conversation emotions. Comparing two different knowledge features extracted based on BERT, using ATOMIC knowledge graph or using conversational semantic role labeling, we observe mixed results and can prove the effectiveness of the self-attention-based fusion method.

C. CASE STUDY

We illustrate a case study on a conversation instance from the IEMOCAP dataset in table 4. We introduce four key commonsense knowledge relations: intent of speaker (denoted as XI), effect on speaker (denoted as XE), reaction of speaker (denoted as XR), and reaction of others (YR). The whole conversation transited from negative emotion to neutral utterance, but then the situation quickly turned to negative emotion and ended with neutral utterance. When there is a sudden mood change, it is difficult to find this scene by traditional methods. These models can cause confusion when analyzing similar emotions, such as Frustrated and sad. As can be seen from Table 4, due to the intervention of common knowledge, the model is easier to deal with sudden emotional transition and has better sensitivity to similar emotions. The commonsense knowledge model not only predicts the emotional type of the next utterance from the current emotional state of itself, but also predicts the emotional state of the listener. When the conversational utterance is neutral, the commonsense knowledge model predicts that the reaction of others is a negative emotion, which plays a huge role in determining the contextual emotional state of the conversation.

TABLE 5. Impact of relation types on KES.

Dataset	Relation Type		
	4 Relation Types	5 Relation Types	All
IEMOCAP	66.32	66.23	65.68
DailyDialogue	52.28	51.47	50.96
MELD	66.46	65.92	65.14

D. IMPACT OF COMMONSENSE RELATION TYPE

We investigate the impact of commonsense relation types on the performance of our proposed model KES. Considering that five of the nine relation types of ATOMIC are used in the COSMIC model, that is, the intent of speaker, effect on speaker, reaction of speaker, Effect of others and reaction of others. Intentions and effects on the speaker and others can be divided into psychological states, and their reactions are events. Intention is also a causal variable, and the rest is effect. There are other relation types, which determine the preconditions and post-conditions of a given event and describe how the subject is perceived by others. We expand the relation set to five relation types and all nine relation types, respectively. We calculate the F1 scores of KES with these two categories of relation types added step by step.

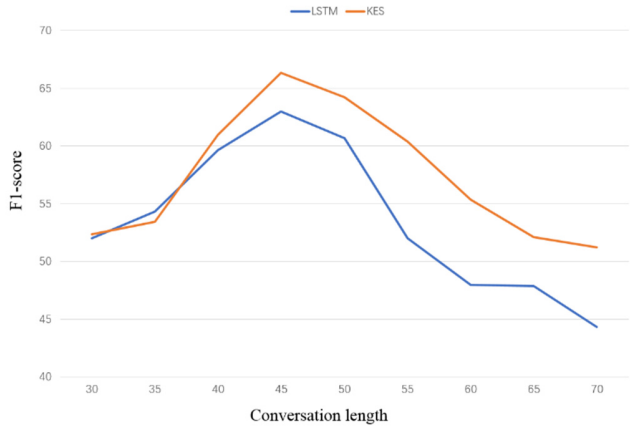

FIGURE 2. Performance of two models in different context lengths.

From Table 5 we can conclude that the inclusion of two extra relation types or all relation types degrades the F1 scores on almost all datasets. We find that too many types of If-Then commonsense relationships will not bring any benefits to the enhancement of knowledge. Although the extra event description enriches the commonsense information in a certain sense, for our model, in the integration and screening of knowledge, the model will miss some important information and focus on the commonsense information that is not critical to the emotional relation.

E. ANALYSIS OF CONTEXT LENGTH

The context encoder of KES model is used to process the context of the conversation. Conversation length will seriously affect the performance of the model. In order to compare and verify the performance of our model, we evaluated the influence of KES on conversations with different lengths. On the IEMOCAP dataset, conversations are grouped by length and fed into two models: our semantic enhanced global context encoder and the contrast model using only LSTM global context encoder.

The F1 scores of different lengths in the two models are shown in figure 2. It is clear that incorporating context into KES improves performance on all datasets. The two context encoders have similar effects on relatively short conversations. However, as the conversation length exceeds 36, KES has more obvious advantages, which proves the contribution of the enhanced semantic encoder based on Transformer to remote context information modeling.

VI. CONCLUSION

This paper proposes utilizing external knowledge to enhance semantics network architecture that incorporates conversational semantic role labeling Information and the commonsense knowledge feature from ATOMIC for emotion recognition in conversation. A knowledge enhanced language representation layer based on self-attention has been developed for fusion extraction. Based on the utterance representations rich in external knowledge, the contextual external state, and individual internal state are modeled to predict the emotional label of conversation. We have done a lot of experiments on three benchmark data sets. KES has made new state-of-the-art advanced achievements, which proves the effectiveness of the proposed model in external knowledge integration.

Future work will focus on integrating more diversified external knowledge. We also plan to incorporate multimodal information into KES and evaluate it on more natural conversation datasets.

REFERENCES

[1] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Positive emotion elicitation in chat-based dialogue systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 866–877, Apr. 2019.

[2] F. Ren, Y. Wang, and C. Quan, "TFSM-based dialogue management model framework for affective dialogue systems," *IEEJ Trans. Electr. Electron. Eng.*, vol. 10, no. 4, pp. 404–410, Jul. 2015.

[3] F. Ren, Y. Wang, and C. Quan, "A novel factored POMDP model for affective dialogue management," *J. Intell. Fuzzy Syst.*, vol. 31, no. 1, pp. 127–136, Jun. 2016.

[4] T. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 463–476, Dec. 2016.

[5] F. Ren, X. Kang, and C. Quan, "Examining accumulated emotional traits in suicide blogs with an emotion topic model," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 5, pp. 1384–1396, Sep. 2016.

[6] T.-Y. Kim, H. Ko, S.-H. Kim, and H.-D. Kim, "Modeling of recommendation system based on emotional information and collaborative filtering," *Sensors*, vol. 21, no. 6, p. 1997, Mar. 2021.

[7] P. Robinson and R. E. Kaliouby, "Computation of emotions in man and machines," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 364, no. 1535, pp. 3441–3447, Dec. 2009.

[8] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Trans. Affect. Comput.*, early access, Jan. 21, 2021, doi: 10.1109/TAFFC.2021.3053275.

[9] F. Ren, "From cloud computing to language engineering, affective computing and advanced intelligence," *Int. J. Adv. Intell.*, vol. 2, no. 1, pp. 1–14, 2010.

[10] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," in *Proc. Interspeech*, Sep. 2009, pp. 320–323.

[11] D. R. Heise, "Enculturating agents with expressive role behavior," in *Agent Culture: Designing Human-Agent Interaction in a Multicultural World*, R. Trappl and S. Payr, Eds. 2004, pp. 127–142.

[12] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 1, pp. 100943–100953, 2019.

[13] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. Morency, *Context-Dependent Sentiment Analysis in User-Generated Videos*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.

[14] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," 2018, *arXiv:1811.00405*.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

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

[17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*.

[18] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion identification in conversations," in *Proc. Findings Assoc. Comput. Linguistics, (EMNLP)*, 2020, pp. 2470–2481.

[19] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, "Did the model understand the question?" 2018, *arXiv:1805.05492*.

[20] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–11.
[21] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep semantic role labeling: What works and what's next," in *Proc. 55th Annual Meeting Assoc. Comput. Linguistics*, 2017, pp. 473–483.
[22] R. W. Picard, "Affective computing: From laughter to IEEE," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 11–17, Jan. 2010.
[23] A. Ortony and J. T. Turner, "What's basic about basic emotions," *Psychol. Rev.*, vol. 97, no. 3, pp. 315–331, 1990.
[24] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, and K. Scherer, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.
[25] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.
[26] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
[27] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychol.*, vol. 14, no. 4, pp. 261–292, 1996.
[28] 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," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.
[29] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.
[30] K. Xu, H. Wu, L. Song, H. Zhang, L. Song, and D. Yu, "Conversational semantic role labeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2465–2475, 2021.
[31] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-BERT: Enabling language representation with knowledge graph," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2019, pp. 2901–2908.
[32] B. Y. Lin, X. Chen, J. Chen, and X. Ren, "KagNet: Knowledge-aware graph networks for commonsense reasoning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 2822–2832.
[33] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451.
[34] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," 2019, *arXiv:1906.05317*.
[35] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9628–9635.
[36] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for ifthen reasoning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3027–3035.
[37] S. Bird, "NLTK: The natural language toolkit," in *Proc. COLING/ACL Interact. Presentation Sessions*, 2006, pp. 63–70.
[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[40] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*.
[41] S.-Y. Chen, C.-C. Hsu, C. C. Kuo, and L. W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," 2018, *arXiv:1802.08379*.
[42] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
[43] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[44] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 873–883.
[45] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 165–176.
[46] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," 2020, *arXiv:2012.08695*.
[47] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.
[48] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," 2021, *arXiv:2105.12907*.
[49] W. Jiao, H. Yang, I. King, and M. R. Lyu, "Hierarchical gated recurrent units for utterance-level emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (Long Short Papers)*, vol. 1, 2019, pp. 397–406.
[50] C. Navarretta, "Mirroring facial expressions and emotions in dyadic conversations," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 469–474.

FUJI REN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, Hokkaido University, Japan, in 1991. From 1991 to1994, he was a Chief Researcher with CSK. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor with the Faculty of Engineering, Tokushima University. His current research interests include natural language processing, artificial intelligence, affective computing, and emotional robot. He is a fellow of the Japan Federation of Engineering Societies, IEICE, and CAAI. He is also the Academician of the Engineering Academy of Japan and EU Academy of Sciences. He is the Editor-in-Chief of the *International Journal of Advanced Intelligence* and the Vice President of CAAI. He is the President of International Advanced Information Institute, Japan.

TIANHAO SHE received the bachelor's degree from the Osaka Institute of Technology, Japan, in 2017, and the master's degree from Advanced Technology and Science, Tokushima University, Japan, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include emotion analysis and natural language generation.