Multi-Task Learning for Emotion Recognition in Conversation with Emotion Shift

Juntao Wang
Kyushu University, Japan
wjt151956@gmail.com

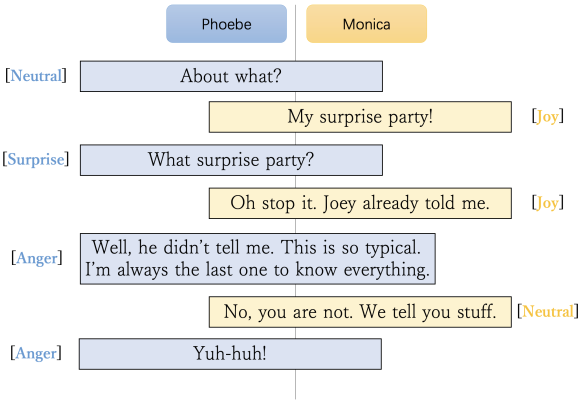
Tsunenori Mine
Kyushu University, Japan
mine@ait.kyushu-u.ac.jp

Abstract

As human-computer interaction continues to evolve, the importance of emotion recognition in conversation is becoming more apparent. For applications like chatbots to provide more human-like responses, it is essential for machines to understand the emotions embedded in conversations. Although most of the recent research has focused on extracting contextual information from conversations, the subtleties of emotion shifts (ES) within local conversations are often overlooked. However, acquiring knowledge about ES between interactions can reduce the rate of emotion recognition errors in dialogues with fluctuating emotions. As a solution for ES detection, we define ES between the same speaker as well as between different speakers using Emotion Recognition in Conversation (ERC) datasets. We propose a novel multi-task learning model, called Mtl-ERC-ES, which identifies three tasks simultaneously: Emotion Recognition in Conversation (ERC), Emotion Shift (ES), and Sentiment Classification (SC). Our approach provides high-quality performance on the ERC task, consistently ranking among the top performers across multiple datasets. Our approach also demonstrates the effectiveness of custom ES tasks.

1 Introduction

Emotion Recognition in Conversation (ERC) is of significant importance in the field of human-computer interaction systems, with applications in areas such as public opinion mining and medical consultation. The objective of ERC is to identify the emotions encapsulated within each utterance in a dialogue. This task is complex due to the context-dependent nature of emotional expression. A wave of ERC research has focused on effectively modeling conversational context, a critical factor that influences the emotional undertones of a dialogue.



[Neutral] About what?

My surprise party! [Joy]

[Surprise] What surprise party?

Oh stop it. Joey already told me. [Joy]

[Anger] Well, he didn't tell me. This is so typical. I'm always the last one to know everything.

No, you are not. We tell you stuff. [Neutral]

[Anger] Yuh-huh!

Figure 1: An example of emotion shift in conversations.

Within conversational settings, emotional inertia often leads speakers to maintain consistent emotional states. However, external stimuli can precipitate abrupt changes in these emotional states. Figure 1 illustrates a dialogue where the speaker's mood fluctuates in response to an external trigger. In this scenario, we can observe multiple points of emotional change. First, the conversation begins with Phoebe showing a neutral emotion. When Monica mentions her surprise party, Phoebe expresses her ignorance about the surprise party, her emotion changes to surprise, but still positive, because the surprise party is a thing that makes people happy. However, when Monica says that Joey told her, Phoebe's mood begins to change, from positive to angry and negative, due to her resentment at being excluded from the information circle. This example illustrates how the unpredictability created by dynamic dialogue affects the flow of emotions in a context.

The study by Poria et al. (2019b) shows that existing state-of-the-art methods tend to continuously replicate the same emotion for a particular party. This pattern is consistent with the understanding that, due to emotional inertia, a speaker's emotion is unlikely to undergo sudden shifts. Consequently, these methods often fall short when emotion shifts (ES) occur. This limitation highlights the fact that many existing techniques lack the ability to effectively detect ES between utterances, while also overlooking the dynamic relationship between emotion labels in a session. The study by Gao et al. (2022) proposes the use of multi-task learning to detect ES and emotion recognition simultaneously. However, their method only defines whether the ES occurs between utterances, so the model cannot fully capture the details of the ES. For example, a model may not be able to distinguish between transitions from happiness to sadness and transitions from sadness to happiness, even though the two transitions have very different meanings and consequences in real life. Therefore, to better understand and predict emotional transfer, the model must be able to capture a more specific direction and magnitude of ES.

To better equip the model to comprehend the ES between utterances, we propose the Mtl-ERC-ES network, a Multi-task Learning method for Emotion Recognition in Conversation with Emotion Shift. This approach allows the model to understand and process three tasks simultaneously. The main task is ERC. The first auxiliary task of ERC, called Sentiment Classification (SC), involves the model learning about emotional polarity, which provides a broader representation of emotions. This understanding enhances the model's ability to distinguish between the magnitude of positive and negative emotions, and thus, to identify whether there have been major or minor shifts in the emotional state within a dialogue. The second auxiliary task specifically addresses Emotion Shift, which has not been well defined and explored in existing research. As such, we propose a novel method that defines and deals with ES based on the sentiment polarity observed in conversations. We conducted our experiments on four ERC datasets. Experimental results show that our model consistently demonstrates exceptional performance. Ablation studies were carried out on both auxiliary tasks, which confirmed the effectiveness of the multi-task learning approach that integrates SC and ES. The main contributions of this study are as follows:

- We develop a detailed definition of emotion shift based on the ERC datasets.
- We propose the Mtl-ERC-ES network, which deals with ERC and related tasks simultaneously: SC and ES using a multi-task learning framework.
- We conduct extensive experiments on four benchmark ERC datasets. The experimental results not only demonstrate the superiority of our proposed method over baseline models but also show its competitiveness with current SOTA models.

2 Related work

2.1 Emotion Recognition in Conversation

In recent years, the field of Emotion Recognition in Conversation (ERC) has attracted increasing attention from researchers due to advances in deep learning technology and the growth of available ERC datasets. Most of these studies focus primarily on modeling static contextual information, which includes information about the speaker and utterances relevant to the target statement. DialogueRNN (Majumder et al., 2019) utilizes recurrent neural networks to track the state of each participant in a conversation and employs this information for emotion recognition. COSMIC (Ghosal et al., 2020) learns the interaction between the interlocutors involved in the dialogue by introducing external knowledge. DAG-ERC (Shen et al., 2021) uses a directed acyclic graph network to learn long-distance context information. DialogueCRN (Hu et al., 2021) explores the perceptual reasoning ability of the model for context based on emotion theory. CoMPM (Lee and Lee, 2022) improves performance by constructing an additional model to extract context information and incorporate it into a pre-trained language model.

However, most of these studies focus on modeling the speaker state in the context and related conversations within a specified range, but omit the dynamic flow of emotion in the local pairwise dialogue. Recently, some studies have explored the dynamic transfer relationship of emotions in conversations. EmotionFlow (Song et al., 2022) captures the influence of emotional propagation during conversations through an additional CRF layer. HCL-ERC (Yang et al., 2022) uses the curriculum learning method to develop a strategy for providing training examples at the dialogue level according to the frequency of ES. DialogueEIN (Liu et al., 2022) designs an emotion interaction network to model the intra-speaker, inter-speaker, and global and local emotional interactions.

2.2 Multi-Task Learning

Multi-Task Learning (MTL) trains machine learning models from multiple related tasks simultaneously, with the aim of utilizing the valuable information contained in multiple tasks to improve the generalization performance. In the field of ERC,

Table 1: The mapping relationship between emotion and sentiment in MELD dataset and IEMOCAP dataset.

MELD		IEMOCAP	
Y_{emo}	Y_{senti}	Y_{emo}	Y_{senti}
neutral	neutral(1)	happy	positive(2)
surprise	neutral(1)	sad	negative(0)
fear	negative(0)	neutral	neutral(1)
sadness	negative(0)	angry	negative(0)
joy	positive(2)	excited	positive(2)
disgust	negative(0)	frustrated	negative(0)
anger	negative(0)		

some of the previous studies have applied MTL. In the study by Li et al. (2020), speaker identification is used as an auxiliary task to enhance contextual information in dialogue. ERC-ESD (Gao et al., 2022) enhances the performance of ERC by defining and detecting ES as a sub-task, determining whether such a shift occurs within the same speaker. In an MTL framework, the model's performance increases notably when the tasks are closely related. For example, Chauhan et al. (2020) conducts a study assuming that sarcasm relates to emotion and sentiment closely, and they propose a method using MTL, leveraging two auxiliary tasks - emotion and sentiment analysis, to enhance sarcasm detection performance significantly. Within the scope of the ERC task, sentiment polarity represents a coarse-grained classification of emotion recognition, while ES indicates the dynamic changes in emotional expression throughout a conversation. These three elements - SC, ES, and the ERC task itself - have a close relationship. So it is reasonable to believe that ES recognition and SC can be used as auxiliary tasks to enhance the performance of ERC.

3 Proposed Approach

3.1 Preliminary

Emotion Recognition in Conversation (ERC)

In ERC, a dialogue is defined as a sequence of N utterances $U = \{u_1, u_2, \dots, u_N\}$, where there are M ($M \geq 2$) speakers $\{p_1, p_2, \dots, p_M\}$, each utterance is spoken by a speaker. For each $\lambda \in [1, M]$, we define U_λ as the set of utterances spoken by speaker p_λ, i.e., $U_\lambda = \{u_j \mid u_j \in U$ is uttered by $p_\lambda\}$. The task of ERC is to predict the emotion $e_t \in Y_{emo}$ of u_t. Y_{emo} is the set of predefined labels from the ERC datasets.

Emotion Shift (ES)

We define ES based on the ERC dataset. For this, we need to ensure that the number of samples for each category of the ES task defined based on the ERC dataset is sufficient. Limited by the amount of data, we consider ES as a fixed category classification task. The advantage of treating it as a classification task is that it is easy to analyze the model's ability to recognize ES in each category based on experimental results. Inspired by the labeling of sentiment (positive, negative, and neutral) of each utterance in the MELD dataset (Poria et al., 2019a), we mark the polarity of the emotion labels of the ERC datasets according to whether the emotion is positive or not and divide them into three types of sentiment labels: positive, neutral, and negative. Taking the MELD dataset and the IEMOCAP dataset as examples, the mapping from emotion to sentiment is shown in Table 1. We rank the sentiment labels from negative(0) to neutral(1) to positive(2), so that ES can be defined as the difference in sentiment labels between the current utterance and the previous one. For instance, if the current utterance is positive (2) and the previous utterance is neutral (1), the difference of 1 indicates a positive ES. Alternatively, if the current utterance is positive (2) and the previous utterance is negative (0), the difference of 2 characterizes a strongly positive ES. This concept is consistent with intuition: a sudden transition from a negative to a positive mood should indicate a strongly positive shift. At the same time, we recognize that the speaker's internal emotional evolution is different from the emotional flux of the overall conversation. Consequently, we first define the ES of adjacent utterances within the dialogue, followed by the speaker's own ES independently. These are defined as follows:

$$Y_{shift}^{global}(u_i) = Y_{senti}(u_i) - Y_{senti}(u_{i-1}), u_i \in U, \quad (1)$$

$$Y_{shift}^{speaker}(u_j) = Y_{senti}(u_j) - Y_{senti}(u_{j-1}), u_j \in U_\lambda, \quad (2)$$

where *global* represents the overall utterance level, specifically, the adjacent utterances in the dialogue, and *speaker* indicates the speaker's utterance level, namely, the adjacent utterances uttered by the speaker themselves. Therefore, for each utterance in U, we assign two additional labels: the speaker's own ES label and the adjacent ES label. The classification for these ES labels is as follows: Y_{shift} = {*Strongly negative shift, Negative shift, No shift, Positive shift, Strongly positive shift*}. Table 2 shows a part of the category distribution of the ES label we built based on the ERC datasets.

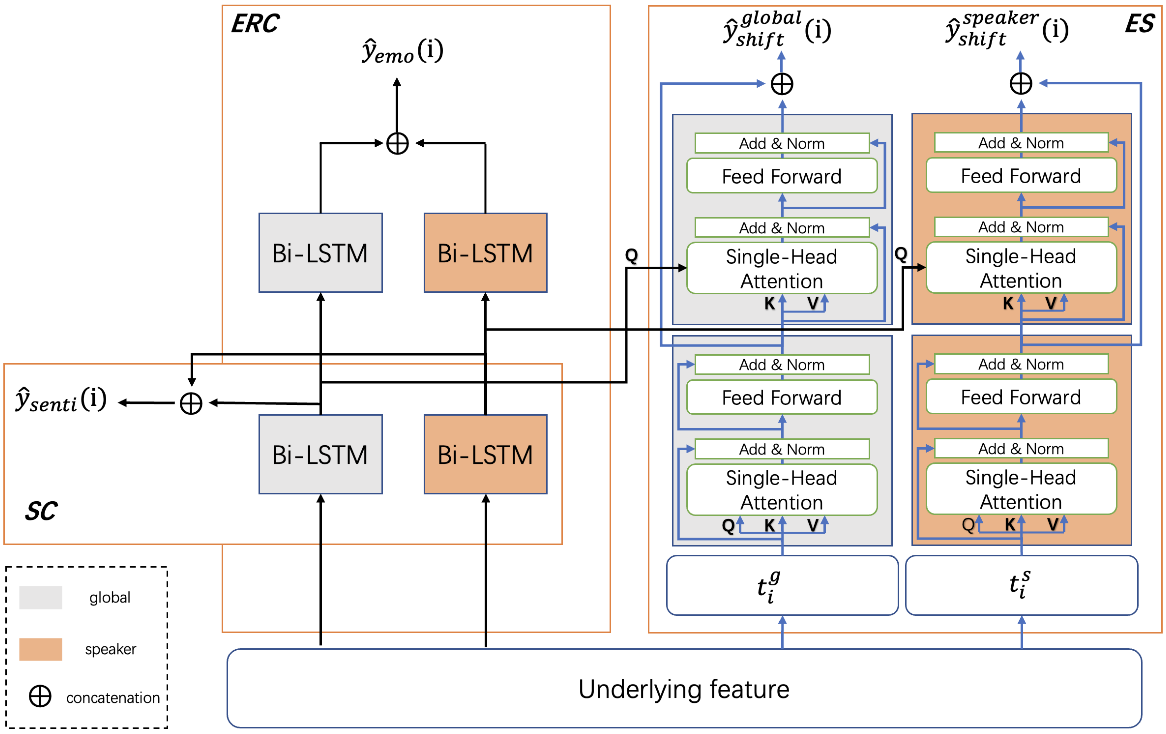


Figure 2: Framework illustration of the proposed model.

Table 2: Label distribution of Emotion Shift (ES) in the test set of the ERC datasets.

Label	**IEMOCAP**		**EmoryNLP**	
	global	speaker	global	speaker
Strongly Negative Shift	48	79	162	310
Negative Shift	221	96	167	132
No Shift	1125	1337	413	348
Positive Shift	209	92	161	127
Strongly Positive Shift	20	19	81	67
Label	**MELD**		**DailyDialog**	
	global	speaker	global	speaker
Strongly Negative Shift	325	780	1007	2004
Negative Shift	467	350	483	327
No Shift	1298	1092	5574	4849
Positive Shift	473	353	665	556
Strongly Positive Shift	47	35	11	4

3.2 Proposed Model

Figure 2 provides an overview of our multi-task learning model. Our model consists of two modules. The first module handles the SC and ERC tasks, classifying sentiment labels at the lower layers of the model and emotion labels at the top layer. It captures both coarse-grained and fine-grained emotional information in utterances. The second module handles the ES task we defined, which is architecturally similar to the Transformer encoder. It includes a custom attention mechanism for querying sentiment information, a single-head self-attention mechanism, and the Feed-Forward Network (FFN). It allows the model to learn sentiment transfer information between utterances. The model feeds a continuous utterance session $\{u_1, \dots, u_t, \dots, u_N\}$, and finally outputs the predicted value of each task corresponding to each mini-batch. We anticipate that through multi-task learning, the model can effectively learn the information about the ES between utterances, thereby enhancing its ability to tackle the challenge of the ERC.

3.2.1 ERC and SC tasks

In order to correspond to the speaker's own unit and the global unit in the ES task, we adopt the context information representation module from DialogueCRN (Hu et al., 2021). We utilize two bidirectional LSTM networks (Hochreiter and Schmidhuber, 1997) to capture speaker-level and global-level context information, respectively. The input $u_i \in U$ has a dimension of d, and each direction of the LSTM has a hidden size of $d/2$, resulting in a total output dimension of d for both directions.

At the global level, we use a two-layer bidirectional LSTM network to capture contextual information between successive adjacent utterances within the dialogue. The context representation at the global level is computed as follows:

$$c^{g_l}_i, h^{g_l}_i = BiLSTM(u_i, h^{g_l}_{i-1}), \qquad (3)$$

where $h^g_i \in \mathbb{R}^d$ is the i-th hidden state of the global-level LSTM, $l \in \{1, 2\}$ is the number of layer.

$c_i^g \in \mathbb{R}^d$ is the global-level context representation.

At the speaker level, we also use a bidirectional LSTM network to capture contextual information between adjacent utterances by the same speaker. The speaker-level context representation is computed as follows:

$$c_i^{s_l}, h_{\lambda,j}^{s_l} = BiLSTM(u_i, h_{\lambda,j-1}^{s_l}), j \in [1, |U_\lambda|], \quad (4)$$

where U_λ refers to all utterances of the speaker p_λ. $h_{\lambda,j}^s \in \mathbb{R}^d$ is the j-th hidden state of speaker-level LSTM for the speaker p_λ, $l \in \{1,2\}$ is the number of layer. $c_i^s \in \mathbb{R}^d$ is the speaker-level context representation. Based on both speaker and global context representations, we define the final representation $o_i(emo)$ as their concatenation, which is subsequently used for the classification of the ERC task.

$$o_i(emo) = [c_i^{g_2}; c_i^{s_2}] \quad (5)$$

At the same time, based on the insights from the study by Chen et al. (2021), which suggests that when MTL is used to improve the performance of a primary task, the introduction of auxiliary tasks at different levels can be beneficial, our model implements a coarser-grained classification at a lower level. This strategy is intended to enhance the model's performance. We use the output $o_i(senti)$ of the first layer of the LSTM network for SC. It computes the SC as follows:

$$o_i(senti) = [c_i^{g_1}; c_i^{s_1}] \quad (6)$$

3.2.2 ES task

Since we define the ES label as the difference between two ERC labels, we define the input to the ES task as the difference between the ERC input features. The ES feature representation is calculated as follows:

$$t_i^g = u_i - u_{i-1}, u_i \in U, \quad (7)$$

$$t_j^s = u_j - u_{j-1}, u_j \in U_\lambda, \quad (8)$$

where $t_i^g \in \mathbb{R}^d$ represents the feature input of the global module and $t_j^s \in \mathbb{R}^d$ symbolizes the feature input of the speaker's own module.

We use two units similar to encoders in Transformer (Vaswani et al., 2017) to model the ES task. In the structure of the first unit, the Single-Head Attention mechanism is applied to learn the relationship between inputs and generate corresponding ES representations. In the second unit, considering the close correlation between the ES task and the SC task, we replace the Query vector in the self-attention mechanism with the output of the LSTM layer in the SC task, with the aim of having the model try to find information related to sentiment features in the representation learned in the first unit, and manifest this part of the information in the final output. Finally, we concatenate the output of the first unit and the second for classification. Based on the input of the adjacent utterances of the overall dialogue and the input of the adjacent utterances of the same speaker, we also construct the global-level and speaker-level models and compute the global and speaker-specific ES, respectively. The calculation process for the global-level output $o_i^g(shift)$ and speaker-level output $o_i^s(shift)$ is as follows:

$$x_i^{g_1} = TRME(Att_{single}(W_q t_i^g, W_k t_i^g, W_v t_i^g)), \quad (9)$$

$$x_i^{g_2} = TRME(Att_{single}(W_q c_i^{g_1}, W_k x_i^{g_1}, W_v x_i^{g_1})), \quad (10)$$

$$o_i^g(shift) = [x_i^{g_1}; x_i^{g_2}], \quad (11)$$

$$x_i^{s_1} = TRME(Att_{single}(W_q t_i^s, W_k t_i^s, W_v t_i^s)), \quad (12)$$

$$x_i^{s_2} = TRME(Att_{single}(W_q c_i^{s_1}, W_k x_i^{s_1}, W_v x_i^{s_1})), \quad (13)$$

$$o_i^s(shift) = [x_i^{s_1}; x_i^{s_2}], \quad (14)$$

where $TRME$ represents the encoder module in the Transformer model. Att_{single} represents the single-head self-attention mechanism, where the three parameters are Q, K, V in the attention mechanism. c is the output of the LSTM layer in the SC task. $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ are trainable parameters.

3.2.3 Training and Prediction for Multi-task Learning (MTL)

Classification

For each task, we process the final hidden state through a feed-forward neural network to get the predicted emotion:

$$P_i = softmax(W_o o_i(task) + b_o), \quad (15)$$

$$\hat{y}_i = \underset{c \in C_{task}}{argmax}(P_i[c]), \quad (16)$$

where $W_o \in \mathbb{R}^{2d \times |C_{task}|}$ and $b_o \in \mathbb{R}^{|C_{task}|}$ denote learnable parameters, $P_i \in \mathbb{R}^{|C_{task}|}$, $|C_{task}|$ denotes the number of labels for each task, c is the class with the highest probability for the task, and $\hat{y}_i$ is the predicted label per task.

Training

We use the cross-entropy loss to train each task:

$$L_{task} = -\frac{1}{\sum_{k=1}^{D} N(k)} \sum_{i=1}^{D} \sum_{t=1}^{N(i)} \log(P_{i,t})[y_{i,t}], \qquad (17)$$

where D is the number of dialogue samples, $N(i)$ is the number of utterances in dialogue i, $P_{i,t}$ is the probability distribution of the label of utterance t in dialogue i, $y_{i,t}$ is the true label of utterance t of dialogue i.

We train our Multi-Task Learning (MTL) model by combining the loss functions of individual tasks into a single global loss function. The combined objective function of the MTL model is then optimized using stochastic gradient descent with backpropagation. Following this strategy, we define the total loss L_{total} as follows:

$$L_{total} = L_{ERC} + L_{SC} + L_{ES}^{g} + L_{ES}^{s}, \qquad (18)$$

where g and s represent ES tasks at the global level and speaker level, respectively. In order to confirm the role of ES and SC tasks on ERC tasks, we posit that each task in our multi-task learning framework holds equal importance. Consequently, we assign a weight of 1 to each task. This uniform weight assignment ensures balanced attention across all tasks during training, thereby preventing any individual task from dominating the learning process at the expense of others.

4 Experiment

4.1 Dataset

We conduct experiments on four standard ERC datasets. Table 3 shows the statistics of the datasets.
IEMOCAP (Busso et al., 2008): This dataset includes 151 binary dialogues, with backgrounds consisting of specific scripts or improvisational performances. The dataset contains a total of 10 speakers. Each utterance in the dialogues is labeled with one of the following six emotions: *happy, sad, neutral, angry, excited,* and *frustrated*.
MELD (Poria et al., 2019a): This dataset contains 1433 dialogues that typically involve more than two speakers. The data is taken from the TV show *Friends*. Each utterance in the dialogues is tagged with one of the following seven labels: *neutral, happiness, surprise, sadness, anger, disgust,* and *fear*.
EmoryNLP (Zahiri and Choi, 2018): This dataset also comes from the TV show *Friends*, the dialogues typically involve more than two speakers. Each utterance is labeled with one of the following seven categories: *angry, disgust, sad, joy, neutral, surprise,* and *fear*.
DailyDialog (Li et al., 2017): This dataset is composed of human-written communications, covering various topics of daily life. Each utterance is labeled with one of the following seven emotion labels: *anger, disgust, fear, joy, neutral, sadness,* and *surprise*.

Table 3: Data distribution of the datasets.

dataset	**dialogues**			**utterances**		
	train	val	test	train	val	test
IEMOCAP	120	12	31	5810		1623
MELD	1039	114	280	9989	1109	2610
EmoryNLP	659	89	79	7551	954	984
DailyDialog	11118	1000	1000	87832	7912	7863

4.2 Baseline Models

To validate the efficacy of our model, we compared it to several previous studies in the field of Emotion Recognition in Conversation (ERC). **DialogueRNN RoBERTa** (Majumder et al., 2019) employs several different GRUs to model the speaker and the global context, and another GRU to model the global emotional interaction state. **DialogueGCN RoBERTa** (Ghosal et al., 2019) uses the graph-based network to model the utterances in a conversation. **COSMIC** (Ghosal et al., 2020) leverages COMET to extract commonsense knowledge, and uses GRU to fuse this external knowledge with contextual information. **DAG-ERC** (Shen et al., 2021) constructs a directed acyclic graph (DAG) based on the dialogue structure, and uses DAGNN to aggregate information between distant and nearby contexts. **DialogueCRN** (Hu et al., 2021) designs a multi-round reasoning module that includes the reasoning process simulated by the LSTM network and the retrieving process using the attention mechanism to extract contextual clues. **DialogueEIN** (Liu et al., 2022) employs a Transformer encoder to simulate semantic interaction within dialogues, and designs an Emotional Interaction Network to capture dependencies and interactions in the conversation. All baseline models utilize the RoBERTa model (Liu et al., 2019) to extract utterance-level features for this task.

4.3 Implementation

The RoBERTa-large model fine-tuned by the emotion classification is used as the feature extractor. For each utterance $u_i \in U$, the last layer of $[CLS]$ embedding is used as the underlying feature repre-

Table 4: ERC performance of different models on four datasets. Micro average F1-score is used on DailyDialog, with the neutral labels excluded. The maximum value is in bold and the second best value is underlined.

ERC Models	IEMOCAP	MELD	EmoryNLP	DailyDialog	
	Weighted-F1	Weighted-F1	Weighted-F1	Macro-F1	Micro-F1
DialogueRNN RoBERTa	64.76	63.61	37.44	-	57.32
DialogueGCN RoBERTa	64.91	63.02	38.10	-	57.52
COSMIC	65.28	65.21	38.11	51.05	58.48
DAG-ERC	68.03	63.65	<u>39.02</u>	-	59.33
DialogueCRN	67.53	<u>65.77</u>	-	-	-
DialogueEIN	**68.93**	65.37	38.92	-	**62.58**
Mtl-ERC-ES	<u>68.68</u>	**66.50**	**39.46**	53.06	<u>60.10</u>

sentation of u_i. u_i's dimension is 1024. We trained the model for 100 epochs, utilizing Adam as the optimizer. The learning rates employed for IEMOCAP, MELD, EmoryNLP, and DailyDialog were 0.0001, 0.00004, 0.0003, and 0.00005 respectively. The batch size is 32, and the dropout rate is 0.2. All experiments were carried out on a 3080 GPU.

5 Result and Analysis

5.1 Overall performance

Following previous works, for IEMOCAP, MELD, and EmoryNLP datasets, we choose the Weighted-average F1 score (Weighted-F1) to measure the overall performance. For the DailyDialog dataset, we employed both the Micro-average F1 score (Micro-F1) and the Macro-averaged F1 score (Macro-F1) to assess the model's performance. Consistent with prior studies, neutral emotions were excluded when evaluating the DailyDialog dataset using Micro-F1.

The results are shown in Table 4. Our model Mtl-ERC-ES has achieved the best performance on both the MELD and EmoryNLP datasets and in particular, makes significant progress on the MELD dataset. Both MELD and EmoryNLP datasets are annotated from the TV show "Friends", where utterances are typically short and casual. This may indicate that our model, Mtl-ERC-ES, is capable of handling daily, shorter dialogues, but whether this can be generalized to all similar conversation environments still requires further testing and validation. In the IEMOCAP dataset, our model is slightly inferior to the advanced model DialogueEIN and similarly falls behind DialogueEIN in the DailyDialog dataset. However, our model still outperforms other baseline models. One possible explanation for this discrepancy lies in the specific strengths of our model, which is excellent at capturing dynamic emotional transitions in dialogues. However, the DailyDialog and IEMOCAP datasets exhibit an extreme imbalance in the ES labels, with the "No Shift" label predominating. This is illustrated by the distribution of the ES labels in Table 2. In the DailyDialog test set, the quantities of the "No Shift" label at the global and speaker levels are [5574, 4849], representing [72%, 62%] of the total samples respectively. Similarly, in the IEMOCAP test set, the "No Shift" label represents [69%, 82%] of the total samples. In contrast, in the test set of the MELD dataset, where our model showed superior performance, the "No Shift" label represents only [49%, 41%] of the cases, and in the EmoryNLP dataset, it represents [41%, 35%]. Thus, our model shows significant advantages in everyday dialogues where emotion shifts occur frequently. This result is consistent with our original intention in designing Mtl-ERC-ES. Experiments on four ERC benchmark datasets - IEMOCAP, MELD, EmoryNLP, and DailyDialog - demonstrate that our multi-task model, which takes into account elements of emotion shift, can achieve outstanding performance.

5.2 Ablation Study

We conducted ablation experiments to verify the effect of multi-task learning on the model. We tested on the IEMOCAP and EmoryNLP datasets, examining the performance with the ES task, the SC task, and both tasks removed. The evaluation metric is weighted F1 scores. The experimental results are shown in Table 5. As can be seen from the table, the model performs less well when trained with only a

Table 5: Results of ablation study on the IEMOCAP dataset, with or without SC (Sentiment Classification) and ES (Emotion Shift) tasks.

Model	IEMOCAP	EmoryNLP
Mtl-ERC-ES	68.68	39.46
w/o ES	68.05	38.91
w/o SC	68.01	38.74
w/o ES&SC	67.75	38.88

Table 6: F1-weighted scores for the task of Emotion Shift on the IEMOCAP dataset.

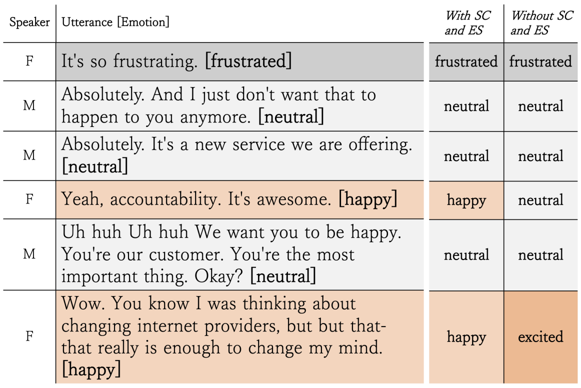
Label	Global	Speaker
Strongly negative shift	76.47	79.04
Negative shift	32.17	21.19
No shift	80.87	86.12
Positive shift	34.57	19.82
Strongly positive shift	47.37	58.82

single task than trained with multiple tasks. This proves that multi-task learning for these three tasks can effectively improve the model's performance.

We also validated the performance of the ES task. The performance on the IEMOCAP dataset is shown in Table 6. From the results, we observe that the model is better at recognizing significant negative and positive ES than regular negative and positive ES, Which means that our model excels at detecting significant changes in emotional fluctuations. However, when identifying subtle changes, such as the shift from neutral to positive, the model is insufficient. On the one hand, this is because the emotion recognition of subtle changes is more challenging than significant changes. On the other hand, our definition of strong ES may benefit the model more, which can be explored in future research.

5.3 Case Study

Figure 3 illustrates a snippet of dialogue taken from a test sample in the IEMOCAP dataset. This case is selected to demonstrate how capturing information about the ES can help the model to accurately identify emotional changes in a dialogue. The background of this dialogue is that service personnel M is trying to solve the problem of speaker F. In this conversation, F's emotions gradually change from initial frustration and despondency to relief and satisfaction. At the beginning of the conversation, F expresses her feelings of frustration. M responds that he fully understands her problem and mentions a new type of service. F is satisfied with this response because she feels understood and respected, so her emotions change due to external influences. When M reaffirms their service philosophy, that they want customer F to be happy, F's emotions remain in a positive state. In this dialogue, if our model relies solely on the single ERC task, it fails to recognize the transition to happiness in the utterance, 'Yeah, accountability. It's awesome.' The model carries forward the emotional judgment from the prior context, even when the emotion has shifted. However, under multi-task learning, our model successfully identifies this emotion shift and accurately detects that F maintains this emotion in the subsequent conversation. This case shows the model's successful identification of F's emotional changes, indicating its ability to accurately capture the causes of ES, which is crucial for predicting dynamic emotions in dialogue.



Speaker	Utterance [Emotion]	*With SC and ES*	*Without SC and ES*
F	It's so frustrating. [frustrated]	frustrated	frustrated
M	Absolutely. And I just don't want that to happen to you anymore. [neutral]	neutral	neutral
M	Absolutely. It's a new service we are offering. [neutral]	neutral	neutral
F	Yeah, accountability. It's awesome. [happy]	happy	neutral
M	Uh huh Uh huh We want you to be happy. You're our customer. You're the most important thing. Okay? [neutral]	neutral	neutral
F	Wow. You know I was thinking about changing internet providers, but but that-that really is enough to change my mind. [happy]	happy	excited

Figure 3: A dialogue snippet of speakers M and F from the IEMOCAP dataset.

6 Conclusion

We discussed a new way to define emotion shifts (ES) based on the ERC datasets, where the defined ES can represent the tendency and magnitude of emotional changes. Moreover, we modeled the tasks of emotion recognition and ES separately at the global dialogue level and the speaker level. Through multi-task learning, the three tasks of Emotion Recognition in Conversation (ERC), Sentiment Classification (SC), and ES are trained simultaneously, effectively improving the performance of ERC. Experimental results show that ES, as defined based on the polarity of sentiment, is effective in improving ERC. Mtl-ERC-ES outperforms the baseline on multiple datasets. We also verified the performance of the ES task, and the results show that under this definition, strong ES can be effectively recognized by the model, which will be helpful for further research on the role of ES in ERC.

Acknowledgement

This work was supported in part by JSPS KAKENHI No. JP21H00907. The authors would like to thank Seita Shimada and Yoshiaki Matsukawa, Rakuten Card Co. Ltd., for their helpful comments on this study.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-task learning in natural language processing: An overview. *arXiv preprint arXiv:2109.09138*.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Joosung Lee and Wooin Lee. 2022. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679, Seattle, United States. Association for Computational Linguistics.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv preprint arXiv:2003.01478*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. 2022. DialogueEIN: Emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11595–11603.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.