BERT-ERC: Fine-Tuning BERT Is Enough for Emotion Recognition in Conversation

Xiangyu Qin[1,2]*, Zhiyu Wu[1]*, Tingting Zhang[1], Yanran Li[2], Jian Luan[2]†, Bin Wang[2], Li Wang[3], Jinshi Cui[1]‡,

[1]School of Intelligence Science and Technology, Peking University
[2]Xiaomi AI Lab
[3]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University
2001213087@stu.pku.edu.cn, wuzhiyu@pku.edu.cn, zhangtingting3412@gmail.com, yanranli.summer@gmail.com, luanjian78@hotmail.com, wangbin11@xiaomi.com, liwang@pku.edu.cn, cjs@cis.pku.edu.cn,

Abstract

Previous works on emotion recognition in conversation (ERC) follow a two-step paradigm, which can be summarized as first producing context-independent features via fine-tuning pretrained language models (PLMs) and then analyzing contextual information and dialogue structure information among the extracted features. However, we discover that this paradigm has several limitations. Accordingly, we propose a novel paradigm, i.e., exploring contextual information and dialogue structure information in the fine-tuning step, and adapting the PLM to the ERC task in terms of input text, classification structure, and training strategy. Furthermore, we develop our model BERT-ERC according to the proposed paradigm, which improves ERC performance in three aspects, namely suggestive text, fine-grained classification module, and two-stage training. Compared to existing methods, BERT-ERC achieves substantial improvement on four datasets, indicating its effectiveness and generalization capability. Besides, we also set up the limited resources scenario and the online prediction scenario to approximate real-world scenarios. Extensive experiments demonstrate that the proposed paradigm significantly outperforms the previous one and can be adapted to various scenes.

Introduction

Emotion Recognition in Conversation (ERC) aims to identify the emotion of each utterance in the dialogue (Poria et al. 2019). This task has been popularly explored in the NLP research community (Ghosal et al. 2019a; Li et al. 2021; Gao et al. 2021), which has wide applications in building automatic conversational agents and mining user opinions.

Existing ERC algorithms reveal multiple influencing factors for understanding conversation emotion. As shown in Figure 1, we divided these factors into three groups: (1) **query utterance information** including the text of the query utterance; (2) **contextual information** including the text of the surrounding utterances (contexts); (3) **dialogue structure information** consisting of non-textual information of the conversation, such as the speaker information, the emotion states, and the relative position of the utterances. To exploit these three kinds of information, previous works (Ghosal et al. 2019b; Majumder et al. 2019; Ishiwatari et al. 2020; Shen et al. 2021) commonly follow a two-step paradigm of first extracting context-independent features via fine-tuning pretrained language models (PLMs) and then characterizing contextual information and dialogue structure information among the obtained features by their classifiers (models). For example, DialogRNN (Majumder et al. 2019) first extracts utterance features with RoBERTa-large (Liu et al. 2019) and then uses three GRUs (Chung et al. 2014) to encode contextual information, speaker state, and emotion state, respectively. To verify the contribution of the contextual information and dialogue structure information analyzed in step two, we design a pilot experiment. Specifically, baseline in the experiment uses RoBERTa-large and MLP as the PLM and classifier respectively, suggesting that these two kinds of information remain unexplored. Compared to the methods following the previous paradigm in Table 1, the baseline achieves comparable performance on MELD (Poria et al. 2018) dataset, indicating that the two kinds of information encoded in the second step only yields trivial improvement (e.g. DialogRNN only outperforms the baseline by 0.22%). Through analysis, the previous paradigm has two flaws. Firstly, the context-independent features obtained by the PLM are fairly abstract, and thus pose obstacles to analyzing contextual information and di-

Method		MELD
PLM	classifier	
RoBERTa-large	RGAT	62.80
	DialogGCN	63.02
	DAGNN	63.12
	DialogRNN	63.61
	DAG-ERC	63.65
RoBERTa-large	MLP	63.39

Table 1: Pilot experiment on MELD (%).

[*]These authors contributed equally.

[†]Corresponding author: luanjian78@hotmail.com

[‡]Corresponding author: cjs@cis.pku.edu.cn

alogue structure information. Secondly, the separation of fine-tuning step and training step leads to extra difficulty in modelling these two kinds of information. Considering these issues, EmoBERTa (Kim and Vossen 2021) discards the second step and uses entire contexts of the query utterance when fine-tuning. However, it lacks further reflections on the way to adapt the fine-tuning process to the ERC task. Thus, we raise several questions: How to use these three kinds of information when fine-tuning? How to optimize the fine-tuning process according to the characteristics of ERC?

Motivated by these questions, we propose a new paradigm for ERC: integrating query utterance information, contextual information, and dialogue structure information when fine-tuning, and adapting the PLM to the ERC task in terms of input text, classification structure and training strategy. The comparison between the proposed paradigm and the previous one is shown in Figure 1. Furthermore, we develop our model BERT-ERC according to the proposed paradigm, which promotes performance in three aspects, namely suggestive text, fine-grained classification module, and two-stage training. (1) Regarding suggestive text, we use the utterances within a certain distance from the query utterance and several indicative tokens, such as speaker name and <*mask*>, to form the input text, thereby indicating speaker information and highlighting the query utterance emotion. (2) The fine-grained classification module considers the temporal structure (past-query-future) of the conversation and generates position-aware features. (3) Concerning the two-stage training, we first train a coarse teacher model via fine-tuning the PLM with the above strategies. Then, we explicitly interpolate the predictions into the input text of the fine student model, allowing it to obtain contextual emotion state. Compared to existing algorithms, both teacher model and student model achieve substantial improvement on four datasets, indicating the effectiveness of these strategies.

In addition to achieving higher accuracy, we note the constraints of ERC in applications, which are ignored by previous works. Thus, we conduct extensive applicability experiments and adapt our paradigm to different scenes. Specifically, we set up the limited resource scenario and the online prediction scenario to approximate real-world scenes. For the former, we design a concise input text structure based on the speaker information to promote the performance of limited-scale PLMs. For the latter, we choose the large-scale PLM and tiny-scale PLM as the coarse teacher and fine student respectively to meet the real-time requirement.

Overall, our contributions can be summarized as follows: (1) We reveal the limitations of the previous ERC paradigm with a pilot experiment. (2) We propose a new paradigm for ERC: integrating three influencing factors when fine-tuning, and adapting the PLM to the ERC task in terms of input text, classification structure, and training strategy. (3) We develop a new model in three aspects, namely suggestive text, fine-grained classification module, and two-stage training. Moreover, it outperforms existing methods and achieves the accuracy of 71.70% on IEMOCAP, 67.11% on MELD, 61.42% on DailyDialog, 39.84% on EmoryNLP. (4) We conduct numerous applicability experiments and adapt the proposed paradigm to different scenarios.

Related Work

Emotion Recognition in Conversation

ERC has received extensive attention in the past decades given its wide applications. Most existing algorithms follow a fixed paradigm that can be generalized as first producing context-independent features and then analyzing contextual information and dialogue structure information. Basically, these methods can be divided into two groups: recurrent-based methods and graph-based methods.

Regarding the recurrent-based methods, HiGRU (Jiao et al. 2019a) uses two GRUs to explore utterance emotion and conversation emotion, respectively. Moreover, DialogRNN (Majumder et al. 2019) employs three GRUs to encode context state, speaker state, and emotion state, respectively. COSMIC (Ghosal et al. 2020) is the latest recurrent-based algorithm, which introduces external knowledge into DialogRNN to achieve better performance.

For the graph-based methods, DialogGCN (Ghosal et al. 2019b) treats the conversation as a directed graph, where each utterance is connected with the contexts. Differently, DAG-ERC (Shen et al. 2021) uses a directed acyclic graph to model the dialogue, where each utterance only receives information from the past utterances. Besides, some methods apply Transformer (Vaswani et al. 2017), in which self-attention can be viewed as a graph. Specifically, KET (Zhong, Wang, and Miao 2019) combines extra knowledge and transformer encoder to boost performance. DialogXL (Shen et al. 2020) adapts the transformer to the ERC task via dialog-aware self-attention.

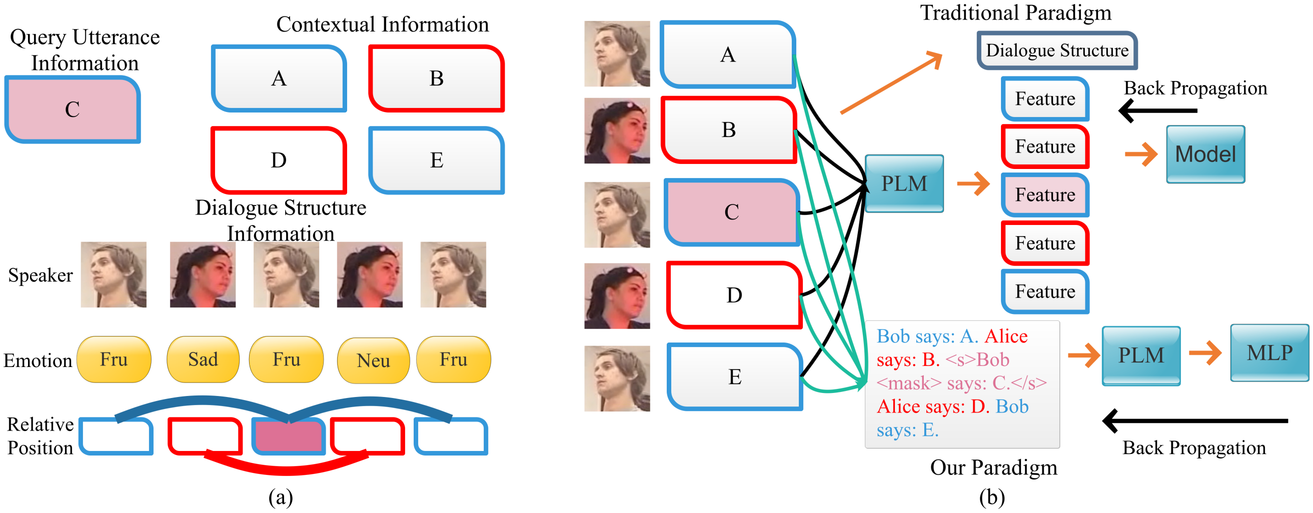
Unlike the above methods, EmoBERTa (Kim and Vossen 2021) feeds the contexts of the query utterance into the PLM and explores contextual information when fine-tuning.

Fine-Tuning Methods

Given the effectiveness of PLMs, researchers typically adapt them to downstream tasks via fine-tuning for better performance. Through our research, existing fine-tuning methods can be divided into three groups: text-based methods, structure-based methods and distillation-based methods.

For text-based methods, Prompt (Kumar et al. 2016; McCann et al. 2018; Radford et al. 2019; Schick and Schütze 2020) allows the similar structure between the input text of the downstream task and pretraining task. For example, when analyzing the emotion of "Alice: I did well in the exam", we may attach the prompt "Alice felt <*mask*>". The <*mask*> token enables the PLM to work in a familiar setting and thus improves its performance in the downstream task. Inspired by Prompt, we design the suggestive text, which indicates the dialogue structure via special tokens.

Structure-based methods facilitate fine-tuning mainly in two ways: introducing external knowledge and enabling parameter-efficient transfer learning. For the former, K-BERT (Liu et al. 2020) injects expertise into the PLM by constraining the self-attention module with a knowledge graph. Besides, prefix tuning (Li and Liang 2021) utilizes a domain word initialized module to emphasize the key content of the downstream task. For the latter, Adapter tuning (Houlsby et al. 2019) attaches small neural modules to



(a)

(b)

Figure 1: (a) Influencing factors in ERC. (b) Different paradigms for ERC.

each layer of the PLM. Moreover, LoRA (Hu et al. 2021) proposes trainable rank decomposition matrices to reduce trainable parameters. In our work, we design a classification structure based on the characteristics of the ERC task.

Concerning distillation-based methods, they aim to maximally compress PLM size at the cost of limited performance loss. Specifically, TINYBERT (Jiao et al. 2019b) proposes a two-stage distillation framework for transformer-based models. Besides, DistilBERT (Sanh et al. 2019) puts forward a lighter BERT (Devlin et al. 2018) via the knowledge distillation strategy. However, they regard distillation loss as the only knowledge transfer pathway. Unlike the above methods, we shift our focus on promoting the performance of the student model and transfer knowledge via the input text of the student model for the first time.

Methodology

Task Definition

Given a dialogue script along with the speaker information about each constituent utterance, ERC aims to analyze the sentiment of each utterance from a predefined set of emotions. Let $[(u_1, s_1), ..., (u_N, s_N)]$ denote a conversation containing N utterances, where s_i represents the speaker of u_i. As illustrated in Figure 2, given an utterance u_i, the object of ERC is to predict its emotion label $y_i \in Y$ according to the contexts $[u_1, ..., u_N]$ and the corresponding speaker information, where Y denotes the emotion set. In addition to the offline prediction, we also investigate online ERC (OERC) for practical needs. As shown in Figure 3, given an utterance u_i, OERC predicts the emotion label $y_i \in Y$ based on the preceding utterances $[u_1, ..., u_i]$ and speaker information.

In this paper, we define the ERC task as $P(Y|X, M, S)$, where Y, X, M, S denote predictions, input text generation approach, PLM, and training strategy respectively. Furthermore, we design a new paradigm for ERC, which can be summarized as integrating three influencing factors (query utterance information, contextual information, and dialogue structure information) during fine-tuning, and adapting the PLM to the ERC task in terms of input text, classification structure and training strategy. In other words, we select the most appropriate (X, M, S) in different scenarios. According to the proposed paradigm, we develop our model BERT-ERC in three aspects: suggestive text, fine-grained classification module, and two-stage training. Details of these strategies will be presented as follows.

Suggestive Text

Let $[(u_1, s_1), ..., (u_N, s_N)]$ denote the conversation, and x_i represents the input text of the query utterance u_i. Traditional algorithms only feed the query utterance into the PLM, i.e., $x_i = u_i$, which proved to be a suboptimal strategy. Thus, we use utterances within a certain distance from the query utterance to form the input text. Nonetheless, directly splicing different utterances probably yields negligible improvement, as the PLM comprehends limited knowledge of dialogue structure in pretraining. Accordingly, we explicitly introduce dialogue structure information and contextual information into the input text:

$$\begin{aligned} x_i =&[X_p(u_a, s_a), ..., X_p(u_{i-1}, s_{i-1}), \\ &X_q(u_i, s_i), \\ &X_f(u_{i+1}, s_{i+1}), ..., X_f(u_b, s_b)] \end{aligned} \quad (1)$$

where a, b denote the range of the contexts, X_p, X_q, and X_f denote the corresponding strategy for past utterances, query utterance, and future utterances. We provide an example in Figure 2, and details will be presented as follows.

Regarding X_q, we exploit three kinds of special tokens. Firstly, we place *speaker says:* ahead of the query utterance to provide speaker information. Secondly, *<s>* and *</s>* are employed to enclose the query utterance for emphasis. Thirdly, we apply the *<mask>* token to focus the model on the emotion state of the query utterance. In other words, the suggestive query utterance can be expressed by:

$$X_q(u_i, s_i) = \textit{<s>} s_i \ \textit{<mask> says: } u_i \textit{</s>} \quad (2)$$

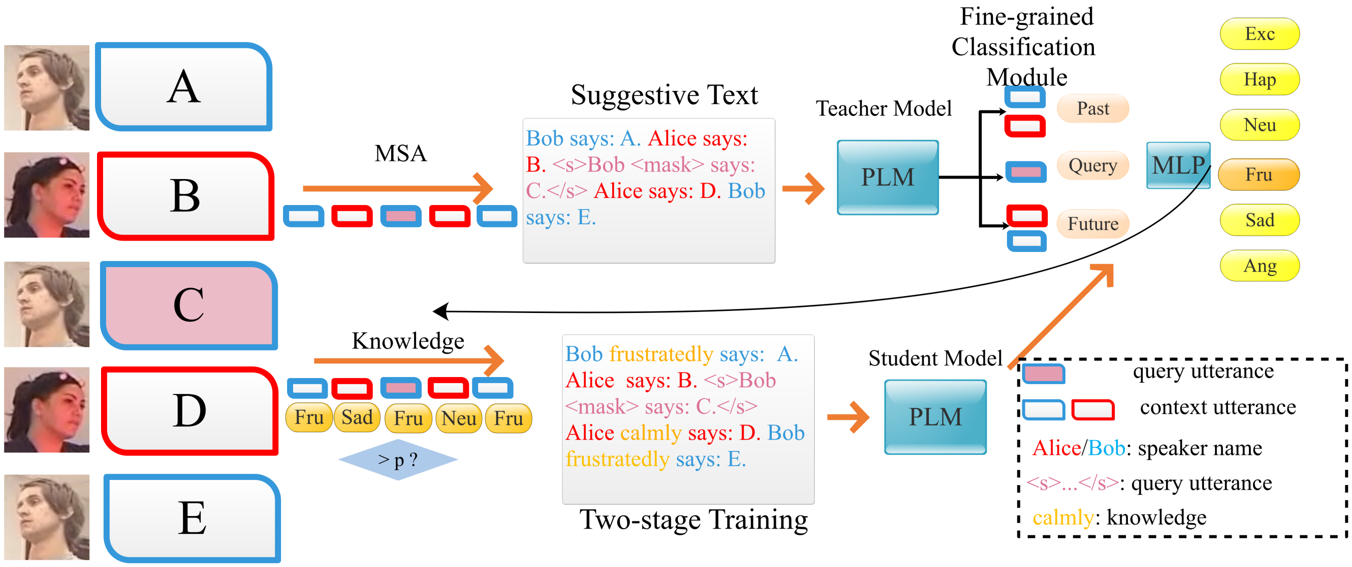

A
B
C
D
E
MSA
Suggestive Text
Bob says: A. Alice says: B. <s>Bob <mask> says: C.</s> Alice says: D. Bob says: E.
Teacher Model
PLM
Fine-grained Classification Module
Past
Query
Future
MLP
Exc
Hap
Neu
Fru
Sad
Ang
Knowledge
Fru
Sad
Fru
Neu
Fru
> p ?
Bob frustratedly says: A. Alice says: B. <s>Bob <mask> says: C.</s> Alice calmly says: D. Bob frustratedly says: E.
Two-stage Training
Student Model
PLM
query utterance
context utterance
Alice/Bob: speaker name
<s>...</s>: query utterance
calmly: knowledge

Figure 2: The pipeline of BERT-ERC.

For X_p and X_f, *speaker says:* serves as the only indication given the supporting role of the contexts:

$$X_p(u_j, s_j) = s_j \textit{ says: } u_j \tag{3}$$

$$X_f(u_j, s_j) = s_j \textit{ says: } u_j \tag{4}$$

We denote the above method as multi-speaker aggregation, as it involves all contexts within a certain distance from the query utterance. However, limited-scale PLMs still perform poorly in modelling contextual information and dialogue structure information even with these special tokens. Considering that utterances of the same speaker as the query one can better reflect the emotion state of the speaker, we propose single-speaker aggregation, thereby reducing the task to exploring the mood swings of a specific speaker. Formally, let X_p^s, X_q^s, and X_f^s denote the corresponding strategy for the three kinds of utterances in single-speaker aggregation. Thus, the aggregated input text can be expressed by:

$$X_q^s(u_i, s_i) = \textit{<s>}s_i \textit{ <mask> says: } u_i\textit{</s>} \tag{5}$$

$$X_p^s(u_j, s_j) = s_j \textit{ says: } u_j \quad \textit{if } s_j == s_i \textit{ else None} \tag{6}$$

$$X_f^s(u_j, s_j) = s_j \textit{ says: } u_j \quad \textit{if } s_j == s_i \textit{ else None} \tag{7}$$

Fine-Grained Classification Module

Given the input text x_i of the query utterance u_i, PLM generates the features of each constituent token. Traditional fine-tuning methods generally utilize the *class* token for classification, as the input of most tasks is a piece of text without any special structures. However, x_i has principal-subordinate structure (query utterance in leading position, contexts in supporting status) and temporal structure (past-query-future), indicating that the previous classifier is sub-optimal for our model. Accordingly, we propose a fine-grained classification module according to the traits of ERC, whose details will be presented as follows.

Let $[f_1, ..., f_l]$ denote the features of x_i, where $[f_a, ..., f_b]$ correspond to the query tokens. We first divide the tokens into past tokens, query tokens, and future tokens based on the position. Then, past features F_p, query features F_q, and future features F_f are generated via mean operation:

$$F_p = mean([f_1, ..., f_{a-1}]) \tag{8}$$

$$F_q = mean([f_a, ..., f_b]) \tag{9}$$

$$F_f = mean([f_{b+1}, ..., f_l]) \tag{10}$$

Afterwards, we get the concatenated feature $F_{cls} = [F_p, F_q, F_f]$, which contains both principal-subordinate structure information and temporal structure information.

Similar to the processing of the *class* token, a fully connected layer followed by the *Tanh* activation function is utilized for projection. Finally, we use the Dropout (Srivastava et al. 2014) layer to prevent overfitting and the MLP for classification. In other words, prediction $\hat{y}_i$ can be computed by:

$$\hat{y}_i = MLP(Dropout(Tanh(FC(F_{cls})))) \tag{11}$$

Two-Stage Training

Given the significance of contextual emotion state, existing algorithms implicitly exploit it via modelling dialogue structure information. Differently, the proposed paradigm makes it possible to explicitly introduce contextual emotion state into the input text. Inspired by knowledge distillation (Hinton et al. 2015), we utilize the coarse teacher - fine student framework. Specifically, we first train a teacher model via fine-tuning the PLM with aforementioned strategies and then interpolate the predictions with high confidence into the input text of the student model. It is worth noting that the two-stage training strategy imposes no constraints on the two models. Accordingly, we test various combinations of PLMs to meet the requirements of different scenarios. The pipeline of the two-stage training is shown in Figure 2, and

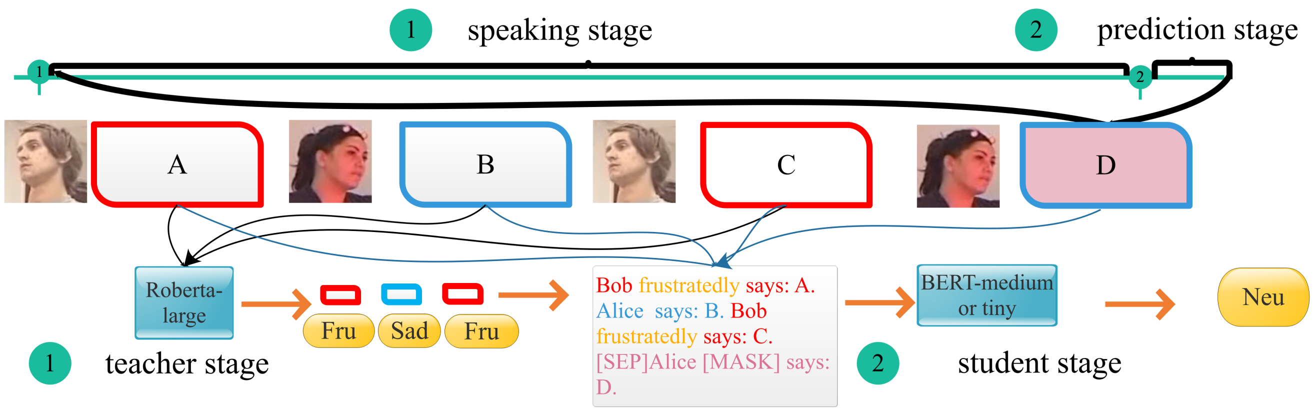


Figure 3: Two-stage training in OERC scenario.

we will illustrate it in terms of knowledge, framework, and combination of different PLMs.

Knowledge According to the emotion set Y of the teacher model, we divide the knowledge into task driven knowledge and common-sense-based knowledge. For the former, Y includes all classes involved in the task. In other words, the teacher model first completes the task and then transfers the predictions with high confidence as the knowledge to the student model. For the latter, we draw inspiration from the way humans perceive emotion state. Specifically, given the predominance of *neutral* emotion in daily conversations, people first simplify ERC as a binary (neutral, emotional) or ternary (positive, neutral, negative) problem, and then conduct further classification. Similarly, the teacher model completes a simplified ERC task and then imparts the predictions to the student model to conduct fine-grained classification.

Framework In two-stage training, we follow the coarse teacher - fine student framework. Firstly, the teacher model generates a pseudo label for each utterance and filters out the low confidence predictions. For a given dialogue $[(u_1, s_1), ..., (u_N, s_N)]$, let $\hat{y}_i$ denotes the prediction of u_i and p_i represents the prediction confidence of $\hat{y}_i$. Thus, the utterance i embedded with knowledge can be expressed by $f(u_i, s_i, \hat{y}_i, p_i)$, and f denotes the screening strategy:

$$f(u_i, s_i, \hat{y}_i, p_i) = \begin{cases} (u_i, s_i, \hat{y}_i) & \textit{if } p_i \geq p \\ (u_i, s_i) & \textit{otherwise} \end{cases} \tag{12}$$

where p is a hyperparameter in our model. Secondly, we introduce the knowledge into the input text of the student model and change the suggestive text strategy of the student:

$$X_q(u_i, s_i, \hat{y}_i) = \textit{<s>}s_i \textit{ <mask> says: } u_i\textit{</s>} \tag{13}$$

$$X_p(u_i, s_i, \hat{y}_i) = s_i \textit{ <emo>says: } u_i \tag{14}$$

$$X_f(u_i, s_i, \hat{y}_i) = s_i \textit{ <emo>says: } u_i \tag{15}$$

where *<emo>* corresponds to the emotion label of $\hat{y}_i$. For example, we set *<emo>* to *angrily* if and only if $\hat{y}_i$ denotes *anger*. In such a manner, the explicitly indicated contextual emotion states improve the performance of the student model. In test phase, we first make predictions with the teacher model and then generate the knowledge, which will be imparted to the student for refined predictions.

Combination of Different PLMs As mentioned above, we choose the combination of PLMs according to the experimental scenario. To achieve optimal accuracy, we use RoBERTa-large as the PLM of both teacher and student. However, large-scale PLMs cannot meet the real-time requirement of OERC, which is also a common issue in prevailing algorithms. To solve this problem, we divide OERC into the speaking stage and the prediction stage. The model is unoccupied in the former stage (3-5 seconds), as it needs the query utterance for emotion recognition. For the latter, the model is required to assess the emotion in 50-100 milliseconds. Considering the long duration of the first stage, the teacher-student framework perfectly fits OERC scenario. Specifically, in the speaking stage, a large-scale PLM is used to generate past emotion states. Then, in the prediction stage, we use a tiny-scale PLM as the student to conduct online prediction based on the text embedded with knowledge.

Experiments

Datasets

Our experiments involve four datasets, whose information is as follows. **IEMOCAP** (Busso et al. 2008) is a multi-modal dataset, where each utterance is labelled with one of six emotions, namely *neutral, happiness, sadness, anger, frustrated, and excited*. Following previous works, dialogues of the first four sessions are used as the training set and the rest are used as the test set. **MELD** (Poria et al. 2018) is a multi-modal dataset extracted from the TV show *Friends*. It contains seven emotion labels: *anger, disgust, fear, happiness, sadness, surprise, and neutral*. **DailyDialog** (Li et al. 2017) collects conversations of English learners. It includes the same seven emotion labels as MELD. **EmoryNLP** (Zahiri and Choi 2018) is also built on the TV show *Friends*, but differs from MELD in scenes and labels. It contains seven types of labels: *neutral, sad, mad, scared, powerful, peaceful, and joyful*. In our experiments, we only utilize textual modality. Regarding evaluation metrics, we follow previous works and choose micro-averaged F1 excluding *neutral* for DailyDialog and weighted-average F1 for the rest datasets.

Method	IEMOCAP	MELD	DailyDialog	EmoryNLP
DialogRNN + RoBERTa (Majumder et al. 2019)	64.76	63.61	57.32	37.44
DialogGCN + RoBERTa (Ghosal et al. 2019b)	64.91	63.02	57.52	38.10
RGAT + RoBERTa (Ishiwatari et al. 2020)	66.36	62.80	58.08	37.78
KET (Zhong, Wang, and Miao 2019)	59.56	58.18	53.37	33.95
DialogXL (Shen et al. 2020)	65.94	62.41	54.93	34.73
DAGNN (Shen et al. 2021)	64.61	63.12	58.36	37.98
COSMIC (Ghosal et al. 2020)	65.28	65.21	58.48	38.11
DAG-ERC (Shen et al. 2021)	68.03	63.65	59.33	39.02
EmoBERTa (Kim and Vossen 2021)	68.57	65.61	-	-
CoMPM (Lee and Lee 2021)	69.46	66.52	60.34	38.93
T-GCN (Lee and Choi 2021)	-	65.36	**61.91**	39.24
BERT-ERC (teacher)	69.43	66.15	60.71	39.73
BERT-ERC (student)	70.84	66.65	61.42	39.84
BERT-ERC (best)	**71.70**	**67.11**	61.42	**39.84**

Table 2: Comparison with the state-of-the-art methods on four datasets (%).

Implementation Details

We conduct experiments in three application scenarios. For offline prediction (Section 4.3), we fix all parameter settings and assess our model on four datasets. Moreover, in Section 4.4 and 4.5, we conduct experiments on IEMOCAP in limited resources scenario and OERC scenario to approximate real-world scenes respectively. Details are as follows.

Regarding **offline prediction**, the proposed model predicts the emotion of each utterance with sufficient time, space and computational resources. To achieve the optimal performance, we use RoBERTa-large (Liu et al. 2019) with the first 8 encoder layers frozen as the PLM. Moreover, we utilize the multi-speaker aggregation and set the knowledge confidence p to 0.7. For the **limited resources scenario**, such as mobile devices, due to the limitations of space and computational resources, we have to use limited-scale models. Besides, we expect more frozen parameters when fine-tuning for parameter reuse. Thus, we choose RoBERTa-base (Liu et al. 2019) with the first 6 or 10 encoder layers frozen as the PLM. Concerning **OERC**, to exploit the time in the speaking stage, we use RoBERTa-large with the first 8 layers frozen as the PLM of the teacher. Besides, we employ BERT-tiny (Turc et al. 2019) and BERT-medium (Turc et al. 2019) as the PLM of the student. Moreover, we set p to 0.5.

Scenario-independent settings are listed as follows. We truncate the input text to meet the requirement of PLMs. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of 9e-6 in experiments. Besides, we utilize a 1-layer MLP as the classifier unless otherwise specified. For all datasets, we train 10 epochs with the batch size of 8. Focal Loss (Lin et al. 2017) is applied to alleviate the class imbalance problem. We implement all experiments on 4 NVIDIA Tesla V100 GPUs with the Pytorch framework.

Offline Prediction Scenario

Comparison with the State-of-the-Art Methods We compare BERT-ERC with several state-of-the-art methods on four datasets in Table 2. The first 8 lines present the performance of several traditional algorithms, which model contextual information and dialogue structure information based on the context-independent features. Through comparison, algorithms using contexts when fine-tuning (lines 9-14) outperform traditional methods, suggesting the significance of exploring contextual information and dialogue structure information in the extraction stage. Moreover, compared to EmoBERTa (Kim and Vossen 2021), our model achieves substantial improvement as we adapt the fine-tuning process to the ERC task. Concurrently, CoMPM (Lee and Lee 2021) and T-GCN (Lee and Choi 2021) insert contextual information into PLMs by extra models. Nonetheless, BERT-ERC still leads in most cases, which mainly benefits from the suggestive text and the fine classifier. Furthermore, we use the two-stage training strategy to generate the fine student model in line 13, which outperforms the coarse teacher in all datasets, indicating that proper knowledge can further boost the performance of our paradigm. Besides, we use a 2-layer MLP as the classifier on IEMOCAP and freeze more PLM layers on MELD to achieve the optimal performance on these two datasets (line 14). Overall, our model achieves leading performance on four datasets, demonstrating the advantages of the proposed paradigm.

Ablation Study We design an ablation study on IEMOCAP to diagnose the proposed modules, whose results are shown in Table 3. (1) The baseline model uses the query utterance as the input and employs the *class* token for classification. (2) The <*mask*> token emphasizes the emotion state of the query utterance and improves the performance by 0.76%. (3) We get the improvement of 10.83% via the contexts of the query utterance, demonstrating that exploring contextual information and dialogue structure information in the fine-tuning step is the most critical strategy in ERC. (4) FCM (line 4) further gains the progress of 0.88%, which can be credited to the introduced dialogue structure. (5) Inspired by ViT (Dosovitskiy et al. 2020), we believe that 2-layer classification MLP outperforms 1-layer MLP in high quality datasets. Compared to the latter, the former obtains the improvement of 1.72%. (6) Compared to the teacher, the fine student model achieves the improvement of 0.55%, suggesting the effectiveness of the two-stage training.

<*mask*>	contexts	FCM	2-layer MLP	two-stage training	IEMOCAP
					56.96
✓					57.72
✓	✓				68.55
✓	✓	✓			69.43
✓	✓	✓	✓		71.15
✓	✓	✓	✓	✓	71.70

Table 3: Ablation study on IEMOCAP (%). (FCM: Fine-grained classification module)

Method	IEMOCAP
DialogRNN + RoBERTa-large	64.76
DialogRNN + RoBERTa-base	62.75
DialogGCN + RoBERTa-large	64.91
DialogGCN + RoBERTa-base	64.18
RGAT + RoBERTa-large	66.36
RGAT + RoBERTa-base	65.22
BERT-ERC + RoBERTa-large + MSA	69.43
BERT-ERC + RoBERTa-base (fr6) + MSA	66.87
BERT-ERC + RoBERTa-base (fr6) + SSA	68.98
BERT-ERC + RoBERTa-base (fr10) + MSA	63.22
BERT-ERC + RoBERTa-base (fr10) + SSA	66.16

Table 4: Performance comparison on IEMOCAP in the limited resources scenario (%). (MSA: multi-speaker aggregation; SSA: single-speaker aggregation; RoBERTa-base (fr$n$): RoBERTa-base with the first n layers frozen;)

Acc (%) PLM Method	BERT-tiny	BERT-medium
MSA + C	42.58	61.64
MSA + K + C	48.22	**69.62**
SSA + K + C	56.27	69.02
SSA + K	**63.27**	68.37

Table 5: Performance comparison on IEMOCAP in OERC scenario (%). (MSA: multi-speaker aggregation; SSA: single-speaker aggregation; K: knowledge; C: contexts)

Limited Resources Scenario

To approximate the scene of conducting ERC on mobile devices, we set up the limited resources scenario and conduct experiments in Table 4. According to the first 6 lines, limited-scale PLMs hazard existing algorithms, possibly due to the reduced modelling capability. Besides, our model suffers from a more severe performance drop, as it is built entirely on the PLM. However, BERT-ERC with a small PLM still outperforms most previous works, suggesting that integrating query utterance information, contextual information and dialogue structure information when fine-tuning is the most critical strategy in ERC. Moreover, single-speaker aggregation (line 9) obtains the progress of 2.11% compared to multi-speaker aggregation (line 8), indicating that focusing only on the mood swings of one speaker improves ERC performance in the resource-limited condition. Besides, we freeze the first 10 layers of RoBERTa-base for more parameter reuse and achieve the accuracy of 66.16% with single-speaker aggregation, which is comparable to the performance of existing methods. Insofar as we know, we are the first to consider PLM size and parameter reuse in ERC, and results show that our paradigm can be adapted to various scenarios by altering the text generation strategy.

OERC

In OERC scenario, we choose RoBERTa-large as the coarse teacher and use BERT-tiny and BERT-medium as the fine student. Turc et al. (Turc et al. 2019) point out that these two PLMs are 3× and 65× faster than RoBERTa-large in inference respectively. The experimental results are shown in Table 5. Baseline model does not use the knowledge (line 1) and fails to meet practical needs. Compared to the baseline, two-stage training introduces contextual emotion state as the knowledge (line 2) and promotes the performance by 5.64% and 7.98%. It is worth noting that BERT-medium with teacher achieves 3× faster inference while still outperforming the state-of-the-art methods in offline prediction scenario. To further boost the performance of BERT-tiny, we utilize single-speaker aggregation (line 3) and achieve the improvement of 8.05%. Furthermore, we simplify the input text by replacing the preceding utterances with the revealed emotion states and achieves an accuracy of 63.27% with BERT-tiny. Reviewing the strategies in line 2 to 4, we discover that simpler input text advances limited-scale PLMs. However, these two strategies fail to benefit BERT-medium, as it has sufficient parameters to model dialogue structure and past utterances. Overall, the proposed two-stage training strategy yields great inference speed gains with limited accuracy loss and achieves great performance on different PLMs with the corresponding text generation approach.

Conclusion

Given the flaws of the previous ERC paradigm, we put forward a new one in this paper and further develop the BERT-ERC model according to the proposed paradigm. Our model utilizes three strategies (suggestive text, fine-grained classification module, and two-stage training) to introduce dialogue structure information and contextual information into the PLM. Through extensive experiments, BERT-ERC achieves state-of-the-art performance on four datasets. Besides, we set up the limited resources scenario and OERC scenario to approximate real-world scenes. Overall, comprehensive experiments demonstrate the generalization ability and effectiveness of the proposed paradigm.

Acknowledgments

We would like to thank the intelligent voice assistant *XiaoAi* and the *Xiaomi AI Lab* for their guidance and inspiration to our emotion conversation work.

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, J.; Liu, Y.; Deng, H.; Wang, W.; Cao, Y.; Du, J.; and Xu, R. 2021. Improving Empathetic Response Generation by Recognizing Emotion Cause in Conversations. In *EMNLP*.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019a. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *EMNLP*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019b. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; and Goto, J. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7360–7370.

Jiao, W.; Yang, H.; King, I.; and Lyu, M. R. 2019a. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. *arXiv preprint arXiv:1904.04446*.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2019b. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Kim, T.; and Vossen, P. 2021. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa. *arXiv preprint arXiv:2108.12009*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, 1378–1387. PMLR.

Lee, B.; and Choi, Y. S. 2021. Graph based network with contextualized representations of turns in dialogue. *arXiv preprint arXiv:2109.04008*.

Lee, J.; and Lee, W. 2021. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. *arXiv preprint arXiv:2108.11626*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y.; Li, K.; Ning, H.; Xia, X.; Guo, Y.; Wei, C.; Cui, J.; and Wang, B. 2021. Towards an Online Empathetic Chatbot with Emotion Causes. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; and Wang, P. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2901–2908.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6818–6825.

McCann, B.; Keskar, N. S.; Xiong, C.; and Socher, R. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

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

Poria, S.; Majumder, N.; Mihalcea, R.; and Hovy, E. H. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7: 100943–100953.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schick, T.; and Schütze, H. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *arXiv preprint arXiv:2012.08695*.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.