Towards Contrastive Context-Aware Conversational Emotion Recognition

Hanqing Zhang and Dawei Song

Abstract—Conversational Emotion Recognition (CER) aims at classifying the emotion of each utterance in a conversation. For a target utterance, its emotion is jointly determined by multiple factors, such as conversation topics, emotion labels and intra/inter-speaker influences, in the conversational context of it. Then an important research question arises: can the effects of these contextual factors be sufficiently captured by the current CER models? To answer this question, we carry out an empirical study on four representative CER models by a context-replacement methodology. The results suggest that these models either exhibit a label-copying effect, or rely heavily on the intra/inter-speaker dependency structure within the conversation, but do not make a good use of the semantics carried by the conversational context. Thus, there is a high risk that they overfit certain single factors, yet lacking a holistic understanding of the semantic context. To tackle the problem, we propose a semantic-guided contrastive context-aware CER method, namely C3ER, to augment/regularize a backbone CER model, which can be any neural CER framework. Specifically, C3ER takes the hidden states of utterances from the CER model as input, extracts the contrast pairs consisting of relevant and irrelevant utterances to the conversational context of a target utterance, and uses contrastive learning to establish a soft semantic constraint between the target utterance and its context. It is then jointly trained with the main CER model, forcing the model to gain a semantic understanding of the context. Extensive experimental results show that C3ER can significantly boost the accuracy and improve the robustness of the representative CER models.

Index Terms—Conversational emotion recognition, conversational context, semantic constraint, contrastive learning

1 Introduction

Conversational emotion recognition (CER) is aimed at utterance-level emotion classification for a conversation. It has attracted an increasing attention from both academia and industry in recent years. Effective CER is crucial for building advanced dialogue systems that would become more empathetic and engaging by taking into account user's emotional state [1], [2]. In addition, the practical demand of CER is growing in many application domains, e.g., online health care, education and legal trails [3].

Compared with traditional sentence/document-level emotion recognition, the main challenge of CER lies in the fact that CER is governed by different factors of context, such as topic, interlocutors' personality, intra/inter-personal dependencies, argumentation logic, viewpoint, and intent, etc [4]. Generally speaking, for a target utterance to be classified, the utterances before and after it in the conversation can be regarded as its "conversational context". The conversational context can contain different aspects that influence CER from different perspectives. In this paper, we divide them into three types: (1) the emotion labels of context utterances, (2) the semantics carried by the actual content of utterances (e.g., topic or dialogue intent), and (3) the relationship between speakers, i.e., intra/inter-speaker influence. For the convenience of presentation, we refer the first two collectively to as semantic context, and the third as structure context.

With the advance of deep learning techniques, neural CER models have achieved certain performance breakthroughs. Poria et al. [5] proposed various early-stage neural CER models based on the long short-term memory (LSTM) structures, to capture the conversational contextual information and get utterance-level representation for emotion classification. After then, numerous neural CER methods have emerged. Most of them are dedicated to building a more solid utterance representation to better model the impact of conversational context. More concretely, they treat the utterances in a conversation as a sequence, and utilize the sequence models commonly used in Natural Language Processing (NLP), such as recurrent neural networks (RNN) [6], [7], [8], [9], Transformer [10], [11], [12], and GCN [13], [14], [15], to aggregate the conversational context of each target utterance to get its final vector representation.

Some recent studies [17], [18] pointed out that the neural networks such as the RNN and Transformer, are hard to fully capture conversational dynamics. As for the field of CER, Ghosal et al. [19] also found that certain CER models exhibit a "label copying" effect, i.e., an effect of mimicking the affective states of the context utterances when classifying the target utterance, rather than understanding the actual semantics of the context. For example, the models in effect tend to "copy" the major emotion label of the conversational context as the predicted label of the target utterance, or directly "copy" the

• *The authors are with the School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100811, China. E-mail: {zhanghanqing, dwsong}@bit.edu.cn.*

Manuscript received 24 November 2021; revised 10 September 2022; accepted 3 October 2022. Date of publication 10 October 2022; date of current version 15 November 2022.
This work was supported in part by Natural Science Foundation of Beijing under Grant 4222036 and in part by Huawei Technologies under Grant TC20201228005.
(Corresponding author: Dawei Song.)
Recommended for acceptance by C.-C. Lee.
Digital Object Identifier no. 10.1109/TAFFC.2022.3212994

emotion transfer patterns (e.g., a negative emotion in a conversation tends to appear after the emotion "anger") in training dataset. However, the conclusion in [19] was limited to the a simple LSTM-based model and lacks a systematic analysis of more representative and state-of-the-art CER models.

In order to fill this gap, we conduct an empirical study on four representative CER methods to further explore what the models actually learn from the conversational context, which are reported in more detail in Sections 4 and 6.6. We find that in general the current CER frameworks fall short in understanding of the semantics of a conversation. On the one hand, LSTM- and RNN-based CER frameworks trend to overfit the label patterns of conversational context, and show the "label copying" effect. Specifically, for each target utterance, we replace its context utterances with different content that yet carry the same emotions as the original ones. We found that such replacement can hardly affect the classification accuracy of the model. However, the performance degraded dramatically, when the replacement content is carried different emotion labels from the original context utterances so that the label patterns of the conversational context are destroyed. This suggests that the models are overly sensitive to the emotion labels instead of the actual semantics of the conversational context. On the other hand, the GCN-based CER models are not label-copying, and instead they rely more on the intra/inter-speaker dependency structure within a conversation. As a result, the performance decreased sharply when the structural context of a conversation is missing. In summary, the representative CER models studies tend to fit certain single aspects of context, i.e., the label-copying or structure-dependency effects, yet lacking a holistic understanding of the semantic context. This limits the accuracy and robustness of the CER models to some extend.

To alleviate the problems mentioned above, we further propose a semantic-guided context-enhanced mechanism to regularize a CER model and facilitate a more effective understanding of conversational context. The intuition is that, in most cases, the semantic context in a conversation tends to be consistent, i.e., an utterance can be predicted by its preceding utterances to a certain extent [18]. As illustrated in Fig. 1, the utterances u_1 to u_3 are about the topic of "name", which is semantically consistent with u_4. In this sense, we call u_4 as "context-relevant". From a deep learning model's perspective, given the context utterances $u_{1:3}$ and with the utterance-level representations generated by a CER framework, if a model can correctly predict u_4 by distinguishing the context-relevant utterance u_4 from the randomly sampled "context-irrelevant" candidates, then the CER framework as more context-aware. Heuristically, a model that fully perceive conversational context would be more helpful for utterance-level emotion analysis. Let us take semantic context in term of topic as an example. Under the topic of "funerals", the utterance is more inclined to a negative emotion; yet the emotion of the same utterance content tend to be positive when the topic is "weddings".

In this paper, we incorporate the above ideas into a contrastive learning scheme, and propose a contrastive context-aware CER method, namely C3ER, to augment a CER framework. Specifically, the representation for a sequence

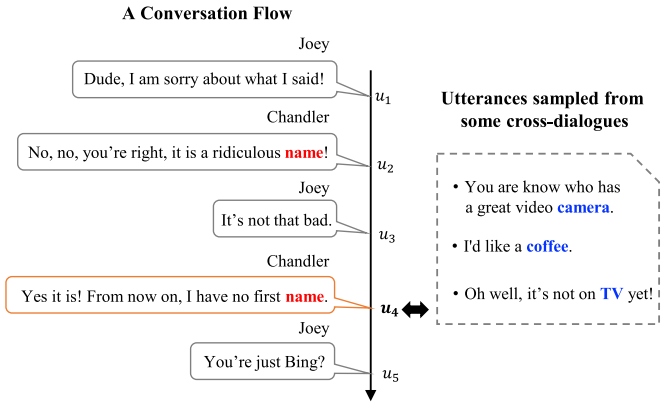


Fig. 1. A toy example of conversational context. The actual content of a conversation (*left-hand side*) is a snippet from the MELD dataset [16]. Assume u_4 is the target utterance that is to be classified. According to its historical utterances u_1 to u_3, we can infer that the conversation is talking about "name", instead of "camera", "coffee" or "TV". Therefore, u_4 is context-relevant. The *right-hand side* of the figure shows some context-irrelevant utterances which are randomly sampled from cross-dialogues and irrelevant to the historical context of u_4.

of utterances can be extracted from any existing CER model. We then map the historical information for a target utterance onto a compact latent semantic space, which can be regarded as a summary of its historical context. Then we take the target utterance itself and its subsequent ones within a certain proximity as the context-relevant (positive) samples, and randomly sample context-irrelevant utterances from other conversations in the dataset (termed as "cross-dialogues") as the negative samples, to construct contrast pairs. Meanwhile, the above contrast pair construction process can also be performed in the opposite direction of the utterance sequence. Finally, contrastive learning is employed, with the objective to make each target utterance's representation semantically closer to positive samples but away from the negative samples in the semantic space. This implicitly achieves the goal that each utterance in a conversation can be predicted by its conversational context. The contrast learning objective could be regarded as a soft semantic constraint for CER and is added to emotion classification task for jointly training, allowing a standard CER framework to better capture the useful semantic information in the conversational context for more effective utterance-level emotion recognition.

We conduct experiments on two datasets (i.e., IEMOCAP [20] and MELD [16]), with four representative CER frameworks in two scenarios including real-time and non-real-time. Experimental results show that C3ER can significantly improve the accuracy of the CER frameworks. Perturbation tests based on replacement of context for each target utterance further reveal that C3ER can help the label copy CER framework flexibly deal with perturbations, and avoid overfitting label patterns. As for the non-label copy CER method, it also can help them capture more useful semantic information and reduce dependence on the structural context of a conversation, thus improving the robustness of the frameworks.

In a nutshell, our main contributions are summarized as follows:

(1) We motivate and explore the problem of semantic understanding the conversational context in CER. To the best of our knowledge, this is the first work to systematically

investigate the influence of semantic context on different CER frameworks, which stimulate the reconsideration of whether the existing CER models make a good use of conversational context and really understand the its semantics, so as to develop more robust and efficient CER methods in the future.

(2) We propose a pluggable approach, namely C3ER, to enhance the contextuality in the utterance representation, using contrastive learning. C3ER can be flexibly used to regularize any existing CER framework during the model training phrase without participating in the inference stage, allowing the CER model to have a holistic understanding of the semantic context.

(3) We conduct a series of experiments on two datasets, with four representative CER frameworks. The results show that the proposed C3ER improves the effectiveness and robustness of the CER frameworks, by alleviating the problem of over-fitting the label patterns for the "label-copying" CER models and strengthening the understanding of semantic context for the "non-label-copying" CER models.

The remainder of this paper is organized as follows: Section 2 gives a brief literature review on conversational emotion recognition and contrastive learning. Section 3 recaps the preliminaries of CER. And then, we choose four representative CER methods as baselines to conduct a empirical study in Section 4, and describe the details of the proposed method in Section 5. We further presents an extensive empirical evaluation of our method and a series of in-depth analysis and empirical studies in Section 6. Finally, we conclude the paper in Section 7 and discuss the promising directions for future research in Section 8.

2 Related Work

Our work is closely related to two areas: Conversational Emotion Recognition and Contrastive Learning. We will discuss the related work in these two fields separately below.

Conversational Emotion Recognition. In recent years, most state-of-the-art (SOTA) CER models are based on deep learning. Early work uses LSTM to capture contextual information of conversations to improve the performance [5]. In order to further model the context, memory networks that were previously used in question answering [21], have also been adapted for emotional reasoning. The core idea is to treat an utterance as a query and its context as a document, and then perform multi-hop inferences to recognize the corresponding utterance's emotion [6], [9], [22]. Later, various context factors that are important emotion recognition have been explored, such as factors about the speakers (e.g., inter-speaker influence, intra-speaker influence, personality, etc.) [10], [11], [13], [14], [15], external knowledge [23], [24], and conversational topic [25], [26]. Inspired by the cognitive theory of emotion, Hu et al. [27] design multi-turn reasoning modules to extract and integrate emotional clues. These factors can be extracted through various neural networks such as RNN, GCN, Transformer, etc., and have continuously improved the CER performance. Ma et al. [28] propose multi-view network (MVN), a real-time CER method, which explores the emotion representation from word- and utterance-level views. A parameter-efficient method, Bidirectional emotional recurrent unit(BiERU) is proposed to model the conversational context Li et al. [29]. Compared with sentence-level or document-level emotion recognition, conversational emotion recognition are more sensitive to conversational context modeling. However, existing CER methods pay more attention to heuristically incorporating emotional factors into the network model to improve the performance, rather than exploring what the models actually learn from the conversational context and developing a solution to capture that. This paper is the first attempt to fill this gap.

Contrastive Learning. The concept of learning useful patterns from contrast pairs has been extensively studied in the literature of contrastive learning [30], [31], [32] and more recently has been applied to a range of applications. Contrastive visual representation learning has been shown to achieve the equivalent effectiveness of supervised learning methods by constructing contrast samples and learning visual representation in a self-supervised fashion [33], [34]. Oord et al. [35] uses a probabilistic contrastive loss to learn sequence representations in a latent space, and the experimental results demonstrate that the approach is able to learn useful representations and achieve a strong performance on several domains. Logeswaran and Lee [36] use the contrast pairs to learn sentence representation, for which two contiguous sentences are considered as positive pairs, and the sentences from other documents are regarded as negative pairs. Cheng et al. [37] use contrastive learning to eliminate bias (such as gender prejudice and racial prejudice) in text representations generated by pre-trained language models. Furthermore, an adversarial perturbation method [38] is proposed for contrastive learning to solve the problem of exposure bias in text generation, with aim to make positive samples have a higher likelihood and negative samples have a lower likelihood by adding perturbations. Xiong et al. [39] propose an approximate nearest neighbor negative contrastive learning (ANCE) approach for dense retrieval, and its main contribution is to select hard training negatives globally from the entire corpus. As contrastive learning does not rely on labeled data in most case, it has been successfully applied in an increasing number of fields.

In the field of conversation modeling, there are also some related works using contrastive learning. Cai et al. [40] propose a group-wise contrastive dialogue learning approach by maximizing positive responses and minimizing negative responses, to tackle the low-diversity problem of dialogue generation. Wu et al. [41] propose a contrastive objective function to simulate the response selection task, and incorporate it into the BERT pre-trained language model for task-oriented dialogue systems (TOD-BERT). Liu et al. [42] proposes two topic-aware contrastive learning objectives, namely coherence detection and sub-summary generation objectives to implicitly model the topic change and handle the information scattering problem for the dialogue summarization task. How to construct the contrast sample pairs is the key for contrastive learning. Compared with the application in images, the construction of positive samples for text is more challenging. In addition, simple contrast pairs that are easy to distinguish have brought about limited gains. Thus applying contrastive learning to natural language processing tasks is still an open problem to be further explored.

To the best of our knowledge, we are the first to use contrastive learning in CER. Different from the existing CER models which mainly focus on developing new neural network structures for performance improvement, we create a semantic context modelling method using contrastive learning to enhance a CER model, which can be incorporated into and jointly learnt with any existing CER framework to improve the classification accuracy and robustness.

3 Preliminaries of CER

3.1 Task Definition

The task of conversational emotion recognition is established in a dialogue scenario and aims at an utterance-level emotion classification. Assume that a conversation l is composed of a sequence of utterances $\{u_j\}_{j=1}^{T_l}$, where T_l is the number of utterances in the conversation. The goal is to design and train a model to predict the emotion e_j of each utterance u_j, where e_j belongs to a finite set of emotion labels $\mathcal{E}$. In general, the whole conversation can be used as context when classifying e_j, while for a real-time scenario, only the historical utterances $u_{1:j-1}$ can be used as context for emotion inference.

3.2 Typical Neural CER Frameworks

The C3ER could be regarded as a soft semantic constraint for an existing CER framework. In order to better illustrate our method, we first introduce the general structure of typical Neural CER Frameworks. We generally divide a typical neural CER framework into two parts: context modeling network and classifier. The former uses various sequence modeling networks such as RNN, memory network and GCN, to aggregate conversational context for each utterance u_t. Formally, it takes a target utterance u_t as input, and outputs its emotional hidden state $\mathbf{h}_t$

$$\mathbf{h}_t = CER(u_t), \tag{1}$$

where $\mathbf{h}_t \in \mathbb{R}^d$ with dimensionality d. CER could be the part in any CER framework before the classification head layer. A typical classifier is composed of a fully connected layer and softmax, which are then used to predict the target utterance's emotion label. The formulation is as follows:

$$\hat{\mathbf{y}}_t = \text{softmax}(W\mathbf{h}_t + \mathbf{b}), \tag{2}$$

where $W \in \mathbb{R}^{|\mathcal{E}|\times d}$, $\mathbf{b} \in \mathbb{R}^{|\mathcal{E}|}$, and $|\mathcal{E}|$ is the number of emotion classes. Typically, cross-entropy is used as the loss function, and the loss for the target utterance u_t can be computed as

$$L_e(u_t) = -\sum_{e=1}^{|\mathcal{E}|} \mathbf{y}_t^e \log\left(\hat{\mathbf{y}}_t^e\right), \tag{3}$$

where $\hat{\mathbf{y}}_t^e$ is the logit output of the CER model, and $\mathbf{y}_t^e$ is the one-hot vector of the target label for the emotion class e.

4 An Empirical Study of Representative CER Models

4.1 Context Replacement Methods

Because of the black-box characteristics of neural networks, it is difficult to directly figure out what existing neural CER models have learned from conversational context information. Instead, we propose a method based on context-replacement to indirectly infer what a model has learned from the conversational context by changing the input of the model. Concretely, we aim to observe the classification performance of the trained CER models under the setting of deliberately modified conversational context of each target utterance. To do so, we have designed three different context replacement methods. The first is to replace each target utterance's conversational context with different content yet keeping their original emotional labels (**EM**), in order to test the model's sensitivity to the actual content of the conversational context. The second is to replace the context with utterances randomly sampled from other conversations (**RM**), and another is the extreme case of RM, which replaces conversational context for each target utterance with the ones which have exact different emotion labels and conversational context(**AM**). As such both content and emotion labels of the context are modified, in order to test the model's sensitivity to the emotion labels associated to the conversational context.

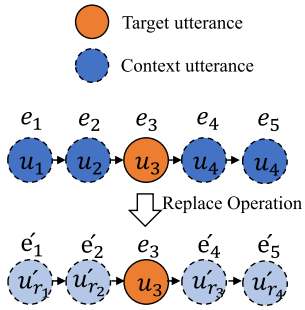


Fig. 2. The diagram of conversational context modification operation. Assume u_3 is regarded as target utterance in a conversation, its history and future utterances are replace with certain constraints. The EM mode represents u_t is replaced with u'_t that has the same emotional label but different content. The RM mode represents u_t is replaced with u'_t that only has different content regardless of its emotional labels. The AM mode represents u_t is replaced with u'_t that has both different content and emotional labels.

Formally, for each target utterance u_t in a conversation, its bi-directional context utterances in the past and future are modified with certain constraints while u_t is kept unchanged. As the illustrated in Fig. 2 , We design three context replacement modes

(1) *Emotion-Relevant Modification (EM)*: For a target utterance u_t, the content for each of its context utterances $c(u_1, \ldots, u_{t-1}, u_{t+1}, \ldots, u_n)$ will be replaced with new utterances drawn from cross-dialogues in the test data, while the emotion labels are kept the same as those associated to the original utterances, to maintain the emotional relevance of the replacement content. (2) *Random Modification (RM)*: Similarly, we need to replace the conversational context for each target utterance. However, the replacement utterances are randomly sampled from the test set, without considering whether or not their emotion labels are related to the original contextual utterances'. (3) *Altering Modification (AM)*: As a more extreme case of RM, we replace conversational context for each target utterance with the ones which have exact different emotion labels and conversational context.

TABLE 1
Conversational Context Replacement Experiment

Methods\Operation	No operation	EM / decline	RM / decline	AM / decline
BC-LSTM[5]	57.3	54.6 / 2.7↓	40.5 / 16.8↓	34.3 / 23.0↓
DialogueRnn[8]	62.2	60.0 / 2.2↓	35.1 / 27.1↓	30.5 / 31.7↓
AGHMN[9]	59.1	54.0 / 5.1↓	31.1 / 28.0↓	25.7 / 33.4↓
DAG[15]	67.3	66.0 / 1.3↓	63.2 / 4.1↓	62.3 / 5.0↓

4.2 Experimental Setup

We implement four representative CER models including a real-time method, i.e., AGHMN [9] and three non-real-time methods, i.e., LSTM [5], DialogueRnn [1], and DAG [15].

BC-LSTM [5]. This is one of the most classic early CER models, which uses LSTM to capture the conversational contextual information and get utterance-level representation for emotion classification.

DialogueRnn [8]. It is one of the most typical CER approaches, in which the attention mechanism and RNNs are used to track the conversational global state and speaker-dependent state. We adopt a specific variant of DialogueRnn, namely Bi-DialogueRnn (publicly available at github[1]), which was reported to have the best performance in previous studies.

AGMHN [9]. It is a SOTA framework using Memory Network for the real-time setting of CER. In this model, an attention gated hierarchical memory network is used to better extract the utterance features under the condition without the future context. We used the variants with the best performance in the actual test, according to the original open-source code (publicly available at github[2]), namely UniF-BiAGRU for MELD and AGRU-BiCNN for IEMOCAP.

DAG [15]. It is the SOTA framework with the best reported performance so far. A directed acyclic neural network is used to model the structure information within a conversation, that is, the relationship between speakers. It achieves the SOTA performance in CER up to now. We follow DAG to use the RoBERTa[43] to extract utterance feature, and choose the variants reported with best performance, namely DAG-EEC provide by the original open-source code (publicly available at github[3]) during our experiments.

Then, we test the performance of these models, under the different context replacement modes, on the widely used IEMOCAP dataset [20] which contains dyadic conversations among 10 speakers and is proved to sensitive to conversational context [23].

4.3 Results

The experimental results, in term of F1 scores, are shown in Table 1. We can find that most of the CER models are predominately sensitive to the emotion labels of conversational context, rather than the utterance's content itself. More specifically, the performance of BC-LSTM, DialogueRnn and AGHMN drops a lot when label patterns of conversational context are destroyed (under the **RM** and **AM** operation). However, when these label patterns are maintained (under the **EM** operation), the performance of those models decreases very little. Hence, they tend to learn the "label copying" patterns.

The only exception is DAG. Its performance does not plummet when the label patterns are destroyed, showing that DAG has almost no "label copy" effect. DAG is a graph-based model to capture conversational structure, it naturally has a structure-dependence. The particularity of this model will be discussed in more detail in Section 6.6.

This experiment reveals quantitatively that the representative CER models tend to focus on the single aspect of emotion labels or dialogue structure, but lack a holistic understanding of the semantics of conversational context. While these individual aspects are important for CER, there is a risk of overfitting. Then a research problem arises: can these neural CER models be semantic-regularized for improved effectiveness and robustness?

In the next sections, we will address this problem by proposing a semantic-enhanced regularization method to augment the CER models via contrastive learning and joint training.

5 Contrastive Context-Aware CER (C3ER)

5.1 Overall Framework

The overall framework of C3ER is shown in Fig. 3. It is pluggable and in principle could be injected to any existing neural CER model. Specifically, it consists of four major components. First, C3ER extracts a sequence of hidden states generated by an existing CER framework for emotion classification. Then a set of contrast pairs that capture the context-aware semantic patterns in a conversation are constructed, based on which a contrastive loss is developed to serve as a semantic-enhanced regularization. Finally, the

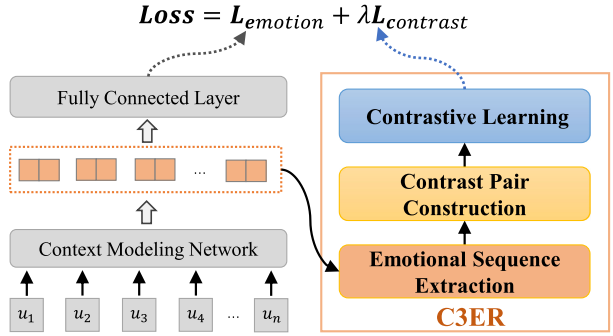


Fig. 3. The overall framework of C3ER. It can be injected into any CER framework (left-hand), and contains 4 components: Emotional sequence extraction, Contrast pair construction, Contrastive learning, and Joint training.

1. https://github.com/SenticNet/conv-emotion
2. https://github.com/wxjiao/AGHMN
3. https://github.com/shenwzh3/DAG-ERC

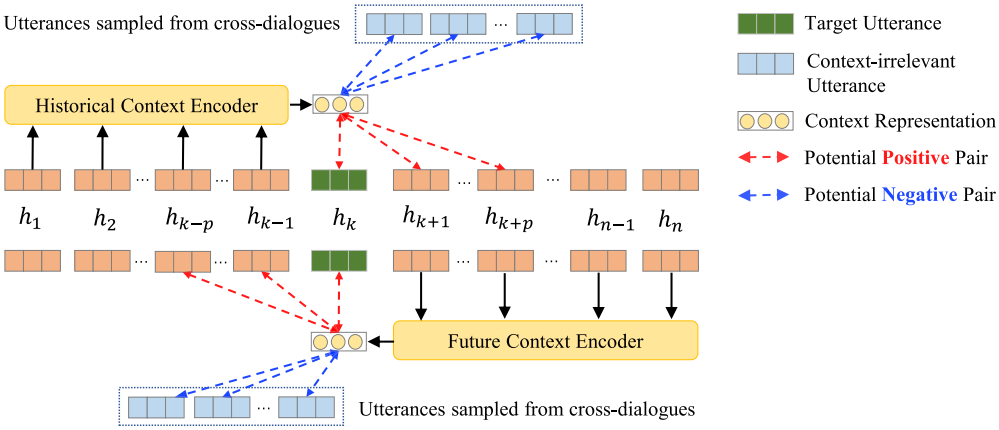


Fig. 4. Contrast pair construction process in C3ER.

contrast loss is added to the existing CER's emotion classification loss for joint training.

Emotional Sequence Extraction. For better flexibility, C3ER directly takes a sequence of hidden states from a backbone CER model, which is described as Equation 1 for every conversation, as its inputs. This allows C3ER to share the neural network structure of the CER model before the classification layer, forcing the network model to learn a more contextual emotional representation.

Contrast Pair Construction. This part is the core process of C3ER. Based on the assumption that consistency of semantic context tends to be maintained in the conversational flow, we construct positive (relevant) and negative (irrelevant) contrast samples from both forward and reverse directions. Those contrast pairs aim to capture semantic patterns from the conversational context, in order for the C3ER model to better understand and comply with the semantic contexts of a conversation.

Contrastive Learning. In this part, we design a contrastive loss to push a target utterance's context semantically closer to its relevant utterances but away from irrelevant ones. This mechanism explicitly establishes a soft semantic constraint between the utterance and its context, pushing a CER model have a better understanding of the conversational context.

Joint Training. We use multi-task learning to achieve the ultimate goal of semantic semantic context based augmentation of the CER model. The contrastive loss is combined with the loss of emotion classification for joint training to improve the performance of CER.

More technical details of these key components are presented in the following subsections.

5.2 Emotional Sequence Extraction

The proposed C3ER is a semantic context based enhancement for a backbone CER model, which in principle can be any neural CER model, e.g., bc-LSTM [5] and DialogueRnn [8]. For a full utilization of the backbone CER model's ability and ensuring the adaptability of the C3ER method, we directly take the utterance-level hidden sequence from the CER model, as input to C3ER. In other words, we injecting the C3ER module onto the classification head of the CER model. This could avoid the adjustment of network structure to a great extent. In addition, our method shares parameters with the CER's context modeling module, so that the two tasks can be integrated and interact with each other to achieve the purpose of mutual promotion.

Specifically, given a conversation composed of a sequence of utterances $l = [u_1, u_2 \ldots .u_n]$, where n is the number of turns in a conversation, it is fed into the CER model to get the emotional hidden state $\mathcal{H} = [\mathbf{h}_1, \mathbf{h}_2 \ldots .\mathbf{h}_n]$ according to the Equation (1). Next, contrast pairs will be constructed based on the obtained utterance-level representation.

5.3 Contrast Pair Construction

Constructing the contrast pairs is at the core of our method. The basic idea is that, for the conversational context of a target utterance, we can identify its relevant utterances as positive examples and its irrelevant utterances as negative example. The contrast pairs thus contain contextual semantic patterns that can enhance the semantic contextuality of the utterance's emotional representation.

As illustrated in Fig. 4, given a sequence of utterance representations $\mathbf{h}_{1:n}$ extracted from an existing CER framework, the historical conversational context for $\mathbf{h}_k$ refers to its historical utterances $\mathbf{h}_{1:k-1}$ in the same dialogue, encoded by a historical context encoder. The relevant utterances are the intra-dialogue utterances adjacent to $\mathbf{h}_k$ within certain distance at the same direction, and irrelevant ones refer to utterances which are sampled from cross-dialogues in the dataset. Heuristically, the same goes for constructing from the opposite direction of a conversation, and the counterpart of the historical context encoder is a future context encoder.

Formally, given the sequence of emotional representation $\mathcal{H} = [\mathbf{h}_1, \mathbf{h}_2 \ldots .\mathbf{h}_{T_l}]$, contrast pairs are generated for every target utterance's emotional representation $\mathbf{h}_k$, where $k \in \{1, 2 \ldots T_l\}$. For a target utterance's hidden state $\mathbf{h}_k$, we construct contrast pairs from two directions. We call the direction in which the dialogue proceeds in chronological order as forward direction, and the reverse as backward direction.

Positive Pairs. For the forward direction, we use a forward $\overrightarrow{GRU}$ model to encode the historical utterances of $\mathbf{h}_k$ as the context representation

$$\overrightarrow{\mathbf{c}}_k = \overrightarrow{GRU}([\mathbf{h}_1, \mathbf{h}_2, \ldots \mathbf{h}_{k-1}]), \tag{4}$$

It is natural to assume that the most relevant to the context $\overrightarrow{\mathbf{c}}_k$ are the w-nearest neighboring utterances in the forward direction. Therefore we construct the context and its relevant utterances as positive pairs, formalized as

$$\overrightarrow{P}_k = \{p_i, p_i \in (\overrightarrow{\mathbf{c}}_k, \mathbf{h}_{k+w})\}, \tag{5}$$

where $\overrightarrow{P}_k$ denotes the set of positive pairs in the forward direction, and $k + w \leq T_l$.

Similarly to the above procedure, we use a backward $\overleftarrow{GRU}$ to encode the future context of the target utterance h_k, and obtain a backward direction context representation $\overleftarrow{\mathbf{c}}_k$. Then, the corresponding backward positive pairs set $\overleftarrow{P}_k$ is constructed. Now, we get all positive pairs for the target utterance

$$P_k = \{\overrightarrow{P}_k, \overleftarrow{P}_k\}. \tag{6}$$

Negative Pairs. Considering that different dialogues have different backgrounds, topics and dialogue intents, a natural idea is to directly use the utterances from different dialogues (i.e., cross-dialogues) as reference negative samples. Therefore, based on the contextual representation $\overrightarrow{\mathbf{c}}_k$ and $\overleftarrow{\mathbf{c}}_k$ described earlier, we can obtain irrelevant utterances randomly sampled from the cross-dialogues to construct negative pairs, formalized as

$$\overrightarrow{N}_k = \{n_i, n_i \in (\overrightarrow{\mathbf{c}}_k, \hat{\mathbf{h}}')\}, \tag{7}$$

$$\overleftarrow{N}_k = \{n_j, n_j \in (\overleftarrow{\mathbf{c}}_k, \hat{\mathbf{h}})\}, \tag{8}$$

where $\hat{\mathbf{h}}'$ and $\hat{\mathbf{h}}$ are the utterances randomly sampled from some other cross-dialogues. The negative pairs of the target utterance can be represented as

$$N_k = \{\overrightarrow{N}_k, \overleftarrow{N}_k\}. \tag{9}$$

The positive pairs P_k and negative pairs N_k contain semantically relevance and irrelevance information about the conversational context of a target utterance. They can be put together to form contrast pairs, which allow the model to perceive up and down at a conversation. For each utterance's hidden emotional representation h_k, its contrast pairs can be represented as

$$D_k = \{N_k, P_k\}. \tag{10}$$

In this section, given the hidden emotional sequences extracted from the backbone CER model, we construct a series of context-irrelevant/relevant contrast pairs for each utterance. In the next section, we will use them within the contrastive learning setting to regularize the CER model and force the model to learn the characteristics of these semantic patterns.

5.4 Contrastive Learning

As illustrated in Fig. 5, given a contrast pair sample, we first design a matching network to calculate two semantic matching scores, one for the positive pair (i.e., between the context representation and the relevant utterance) and the other for the negative pair (i.e., between the context representation and the irrelevant utterance), and then build a contrastive loss based on the scores. Concretely, for every target utterance's hidden emotional representation $\mathbf{h}_k$, it is

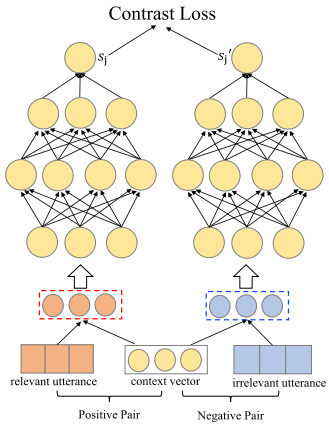


Fig. 5. The schematic diagram of the contrastive learning in C3ER. The contrast samples are feed into a MLP network to calculate the matching scores. And the contrast loss is applied to give higher matching scores for the positive contrast sample, and lower matching scores for the contrast negative sample. In practice, the MLP for calculating matching scores for positive and negative samples is parameter-sharing.

accompanied by a set of contrast pairs $D_k = \{P_k, N_k\}$. Given a pair $d_j = (\tilde{\mathbf{c}}_j, \tilde{\mathbf{u}}_j)$ selected from D_k, we first concatenate the vector representations of a contrast sample, and then a MLP (Multi-layered Perceptron) network is used to compute the matching scores for the pair

$$o_j = \text{mlp}([\tilde{\mathbf{c}}_j \circ \tilde{\mathbf{u}}_j]). \tag{11}$$

In our method, we use a 2-layer MLP, which inputs the concatenated vector and output a real value as the matching score. The sigmoid function is used to normalize the score to convert the matching scores into the interval [-1,1]

$$s_j = \text{sigmoid}(o_j). \tag{12}$$

Intuitively, the goal of our proposed approach is to give higher matching scores for the positive samples, and lower matching scores for those negative samples. Accordingly, for the target utterance $\mathbf{h}_k$, its contrastive loss is given by

$$L_c(D_k) = -\frac{1}{|P_k|}\sum_{i=1}^{|P_k|} s_i^+ - \frac{1}{|N_k|}\sum_{j=1}^{|N_k|}(1 - s_j^-), \tag{13}$$

where s_i^+ and s_j^- represent the matching scores for a positive pair from P_k and a negative pair from N_k respectively. $|P_k|$ denotes the number of positive sample pairs in D_k, and $|N_k|$ is that of the negative pairs.

5.5 Joint Training

Compared with the original CER framework, the C3ER enhanced CER has an additional context encoding module composed of a GRU and a matching network. The parts of context modeling and emotion classifier are shared with the original CER framework. The entire network's loss function integrates the emotion classification loss and the contrast loss. The former is defined as in Equation (3), and the contrastive loss is added to the original CER model for joint training in a form that is similar to a regularization

TABLE 2
The Statistics of the Used Datasets

Dataset		#dialogues	#utterances	Avg. #turns
IEMOCAP	train/val	120	5,810	48.4
	test	31	1,623	52.4
MELD	train	1,039	9,989	9.6
	dev	114	1,109	9.7
	test	280	2,610	9.3

"#" Represents the number of corresponding items.

$$L(\theta) = \frac{1}{\sum_{l=1}^{L} T_l} \sum_{t=1}^{T_l} (L_e(u_t) + \lambda L_c(D_t)), \tag{14}$$

where θ denotes the set of parameters of C3ER, T_l is the number of utterances in the lth conversation, and L is the total number of conversations in the training set. $L_e(u_i)$ is the loss function that is described in Equation (3). $L_c(D_t)$ is the contrasitve loss for C3ER module, and λ is a coefficient which determines the intensity of semantic context-enhancement.

6 Experiments

6.1 Datasets

As shown in Table 2, two commonly used benchmarking datasets: IEMOCAP [20] and MELD [16], are chosen to evaluate our proposed method. IEMOCAP contains videos of dyadic conversations among 10 speakers under diverse scenarios. MELD is a multi-party conversation dataset crawled from the Friends TV series. Both datasets are originally for multi-modal dialogue emotion classification. Since this work focuses on the modeling of the conversational context, only textual modality in the original data is used.

6.2 Baselines

Consistent with Section 4, four typical CER models are chosen as baselines, including two classic early methods (namely BC-LSTM [5] and DialogueRnn [8]) and two SOTA methods with leading performance so far, namely AGMHN [9] and DAG [15].[4] They are described in Section 4.

We further incorporate the proposed contrastive learning module (i.e., C3ER) into each baseline model and explore if it can lead to a performance improvement.

For a fair comparison, the experiments with DialogueRnn and bc-LSTM use the pre-trained utterance features provided by the authors of DialogueRNN [8] which is available at github[1]. We just use the textual feature in both of IEMOCAP and MELD. As for AGMHN [9] and its variants, we also follow the data pre-processing method provided in the original source code of the model[2]. In the same way, we use the pre-trained features provided by the open source[3] for DAG [15].

6.3 Experimental Settings

Precision (P), Recall (R) and weighted F-score (F1) are used as performance measures to evaluate the single-category emotion classification. Then the overall performance of each comparative model is evaluated by Accuracy (Acc) and weighted F-score.

With respect to the modeling direction (i.e., forward or backward) of the contrast pair construction process, we consider two variations for C3ER: namely Uni-C3ER and Bi-C3ER. The former means to construct contrast pairs only from the forward direction, while the latter is from both directions simultaneously. In addition, we conduct the significant test for each CER model with/without the Bi-C3ER module on IEMOCAP and MELD, the statistical significance result is reported by permutation test with $p < 0.1$.

The hyper-parameters of the baseline CER models are mainly obtained by referring to the source codes and academic papers. The specific parameters are determined by taking the best results actually measured under different seeds. On this basis, our proposed contrastive module (C3ER) is added, and its external hyper-parameters are obtained through a grid search. We use full conversations as the counting unit of batches. The range of parameters λ is [0.001,0.01,0.1,1.0,10], and the size of negative samples is varied in the range between 1 and batch size minus 1. During our experiments, we only choose 8 and 16 as batch size. We record the hyper-parameters of the Bi-C3ER branch for reproducing the experiment result in Tables 3 and 4. All the parameters are reported in order bs-LSTM, DialogueRnn, AGMHN, and DAG. For IEMOCAP dataset, $\lambda = [10.0, 1.0, 0.01, 0.1]$, batch size is [8,16,16,8] and size of negative samples is [5,9,8,1]; For the MELD dataset, $\lambda = [1.0, 1.0, 1.0, 10]$, the batch size of all baselines is set to 8, and size of negative samples is [5,4,3,5].

6.4 Experimental Results

Tables 3 and 4 show the experimental results on IEMOCAP and MELD dataset respectively. We can draw the following observations:

(1) C3ER can improve the classification accuracy over all the baseline methods. On IEMOCAP, it improves the performance of the original AGMHN and DialogueRnn by 3.4%-4.8%. This result proves that the semantic context of an utterance in the conversation is important for CER. The improvement on the IEMOCAP dataset is more significant than MELD, which is consistent with the conclusion from previous works [23] due to the nature of data. Compared with IEMOCAP, the MELD dataset has a shorter dialogue length, and the context has less influence on emotion classification.

(2) Given the improvement over DAG, we have achieved a new SOTA performance on both IEMOCAP and MELD. DAG was the SOTA model according to the reported results so far. After injecting C3ER into DAG, the classification accuracy performance is further improved by about 1.6% on IEMOCAP and 1.3% on MELD, achieving a new SOTA performance. It proves the superior effectiveness of C3ER.

(3) Among the variants of C3ER, the effectiveness of Bi-C3ER is better than Uni-C3ER. According to the experimental results, the performance of using Uni-C3ER is between the use of Bi-C3ER and the original methods in the most cases, which indicates that the bidirectional contrast pairs can provide more useful semantic information for model training and thus lead to more gain in performance. In particular, this phenomenon would be beneficial for the real-time CER scenario, because our method only operates in the

4. We also implemented the DialogueGCN [13], but we find it easy to collapse during the model training phrase, so we do not include it as our baselines.

TABLE 3
Performances on IEMOCAP Dataset

Model	Happy			Sad			Neural			Angry			Excitd			Frustrated			Avg	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc	F1
bc-lstm	**41.0**	34.7	37.6	**69.4**	66.5	67.9	51.6	**53.9**	52.7	60.0	**67.1**	**63.3**	**66.4**	56.2	60.9	56.2	62.2	59.0	57.8	57.9
+Uni-C3ER	37.3	26.4	30.9	67.3	69.8	68.5	**53.8**	53.1	53.5	60.1	64.7	62.3	63.4	**67.2**	**65.3**	57.2	58.3	57.7	58.2	57.8
+Bi-C3ER	37.7	**41.7**	**39.6**	68.8	**71.0**	**69.9**	53.5	53.6	**53.6**	**66.9**	59.4	62.9	66.2	52.5	58.6	**58.2**	**67.0**	**62.3**	**58.7**†	**58.7**†
																			1.6%↑	**1.4**%↑
DialogueRnn	52.4	**30.6**	**38.6**	86.9	64.9	74.3	54.9	57.8	56.3	**69.0**	64.1	**66.5**	**69.6**	74.2	71.8	56.4	**70.3**	**62.6**	63.1	62.9
+Uni-C3ER	53.6	25.7	34.7	53.6	25.7	83.7	59.1	**60.9**	60.0	68.2	60.6	64.2	64.9	89.6	64.9	**61.8**	53.5	57.4	65.2	64.0
+Bi-C3ER	**55.9**	26.4	35.8	**89.7**	**78.4**	**83.7**	**63.7**	58.1	**60.8**	64.1	**68.2**	66.1	65.1	**90.3**	**75.6**	59.2	61.4	60.3	**66.1**†	**65.2**†
																			4.8%↑	**3.7**%↑
AGMHN	50.3	**53.9**	**52.0**	**81.6**	**66.9**	**73.5**	49.6	58.6	53.7	69.5	**61.8**	65.4	75.6	55.9	64.2	56.1	65.1	60.1	60.8	61.3
+Uni-C3ER	50.9	38.46	43.8	81.5	64.9	72.3	50.1	58.1	53.8	**82.1**	56.5	**66.9**	**76.2**	62.5	68.6	54.8	**73.5**	**62.8**	61.9	61.7
+Bi-C3ER	**55.0**	38.5	45.3	80	63.7	70.9	**54.0**	**64.1**	**58.6**	74.6	60.6	**66.9**	72.8	**69.9**	**71.3**	**58.1**	68.0	62.6	**63.4**†	**63.4**†
																			4.3%↑	**3.4**%↑
DAG	43.0	36.4	39.4	**83.7**	77.6	50.5	63.6	**76.6**	69.5	**71.7**	64.1	67.7	70.8	64.9	67.7	68.9	**69.8**	69.4	68.1	67.9
+Uni-C3ER	48.8	**42.0**	45.1	79.1	80.4	**79.8**	65.7	74.2	69.7	69.6	64.7	67.1	**73.1**	68.2	**70.6**	**69.1**	68.8	68.9	68.9	68.8
+Bi-C3ER	**50.0**	41.3	**45.2**	78.0	**81.2**	79.6	**66.4**	74.5	**70.2**	68.2	**68.2**	**68.2**	70.9	**69.2**	70.1	71.6	67.0	**69.2**	**69.2**†	**68.9**†
																			1.6%↑	**1.5**%↑

It shows the overall performance of the four baselines and two variants Uni-C3ER and Bi-C3ER, which we added on the basis. † represents the Bi-C3ER achieves significant improvements over the baselines among the average Acc and F1 metrics.

training phase and it can effectively utilize bidirectional dialogue information during the model training, but still carries on the single direction in the reasoning stage.

(4) The performance of C3ER on bc-LSTM is weaker than that of the other three more complex CER frameworks. Specifically, the improvements over bc-LSTM on both datasets are less than 1.6%, which is less than the best improvement performance (close to 5%) on DialogueRnn and AGMHN. This phenomenon shows that the proposed C3ER module also relies on the complexity of the backbone model to a certain extent, because the conversational context modeling itself is a important and challenging task.

6.5 Model Analysis

Coefficient λ for Contrast Loss. we explore how the regularization coefficient λ affects the emotion classification performance. The range of λ is $[1e^{-2}, 1e^{-1}, 0, 1e^{1}, 1e^{2}]$. As shown in Fig. 6, the performance trends of DialogueRnn and AGHMN on both datasets are consistent. With the increase of λ, F1-

TABLE 4
Performances on MELD Dataset

Model	Sadness			Neural			Angry			Surprise			Fear			Disgust			Joy			Avg	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc	F1
bc-lstm	30.2	7.7	12.3	**72.0**	81.6	76.5	37.5	**42.3**	**39.8**	**46.2**	51.9	**48.9**	-	-	-	-	-	-	**50.7**	54.0	52.3	59.3	56.3
+Uni-C3ER	**42.3**	5.3	9.4	70.8	**83.6**	76.7	39.2	39.1	39.1	44.2	49.8	46.8	-	-	-	-	-	-	50.6	55.0	52.7	59.6	56.0
+Bi-C3ER	36.5	**11.1**	**17.0**	71.9	82.2	**76.7**	**40.0**	37.1	38.5	44.6	**53.3**	48.6	-	-	-	-	-	-	50.1	**56.7**	**53.2**	**59.8**	**56.8**
																						0.8%↑	**0.9**%↑
DialogueRnn	25.9	13.5	17.7	71.2	81.4	76.0	**46.1**	35.7	40.2	44.4	49.1	46.6	-	-	-	-	-	-	48.1	**58.7**	**52.9**	59.3	56.4
+Uni-C3ER	**26.5**	13.9	18.2	70.9	81.8	76.0	45.4	**44.1**	**44.7**	**45.9**	**50.2**	**48.0**	-	-	-	-	-	-	51.7	52.5	52.1	59.8	57.1
+Bi-C3ER	26.4	**15.4**	**19.5**	**71.3**	**83.3**	**76.7**	43.1	41.0	41.8	45.0	50.9	47.7	-	-	-	-	-	-	**54.0**	50.5	52.2	**60.0**†	**57.2**†
																						1.2%↑	**1.4**%↑
AGMHN	**36.5**	18.8	24.8	**73.7**	77.6	75.6	42.6	40.6	41.5	53.5	49.5	51.4	13.6	12	12.8	12.9	**11.8**	**12.3**	50.3	**60.7**	55.0	59.4	58.4
+Uni-C3ER	32.2	22.1	26.2	71.5	**81.8**	**76.3**	**50.3**	28.1	36.1	46.0	**61.6**	**52.7**	16.0	8.0	10.7	7.1	1.5	2.5	52.6	55.2	53.9	60.2	57.8
+Bi-C3ER	30.5	**28.9**	**29.6**	72.6	79.9	76.1	42.9	**43.8**	**43.3**	**55.7**	45.6	50.1	**19.5**	**16.0**	**17.6**	**29.4**	7.4	11.8	**55.9**	54.5	**55.2**	**60.3**†	**59.2**†
																						1.5%↑	**1.4**%↑
DAG	43.4	30.3	35.7	**77.0**	77.4	77.2	57.1	42.0	48.4	49.7	67.3	57.1	21.2	**28.0**	**24.1**	43.5	**25.0**	**31.8**	**57.5**	**66.4**	61.7	63.9	63.3
+Uni-C3ER	**49.3**	32.2	**39.0**	76.9	**79.2**	**78.0**	56.7	**44.1**	49.6	49.9	68.0	57.5	20.0	12.0	15.0	52.4	16.2	24.7	55.4	65.9	60.2	64.6	63.6
+Bi-C3ER	46.9	**32.7**	38.5	76.9	78.8	77.8	**57.4**	43.8	**49.7**	**50.3**	**68.1**	**57.8**	**25.0**	16.0	19.5	**61.9**	19.1	29.2	55.5	**66.4**	60.5	**64.7**†	**63.8**†
																						1.3%↑	**0.8**%↑

The content is the same as Table 3. "-" indicates that the classification performance of the model is close to 0 in corresponding items, since MELD dataset have only a very small samples of "fea" and "disgust."

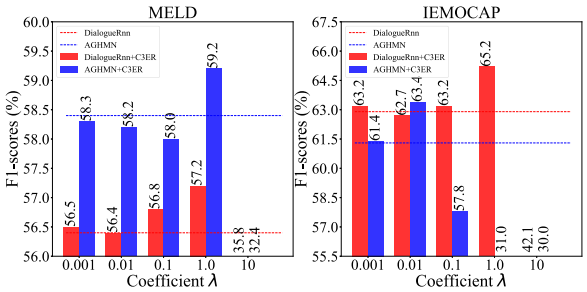


Fig. 6. Influence of contrast loss coefficient λ on performance of different models (DialogueRnn & AGHMN) and datasets (MELD & IEMOCAP).

score shows a trend of increasing first and then decreasing. This is in line with our expectation: when λ is too small, the model only focuses on the emotion classification task, so that our C3ER module has little impact on the emotion classification task; when λ is too large, the main task of the model will be ignored; therefore a moderate value tends to achieve a better effectiveness of emotion classification. When the value of λ is 1.0, DialogueRnn achieves the best enhancement effect on IEMOCAP and MELD. As for AGHMN, the best values of λ is 1.0 and 0.01, respectively.

However, as shown in Table 5, it is interesting to see that the performance of DAG is quite different from DialogueRnn and AGHMN. For DAG, the contrast loss coefficient λ has relatively less effect on performance of emotion recognition than DialogueRnn and AGHMN. Even when the $\lambda = 100$, the decrease of F-scores is less than 4.0% for MELD and less than 7.0% for IEMOCAP. This may indicate that DAG itself have modeled well the context information which is relevant to emotion classification, and then context-enhanced module have a strong consistency with the optimization objective of DAG.

Size of Negative Samples. We also investigate the influence of the number of negative samples on the emotion classification results for the IEMOCAP dataset. In order to better control the variables, we select the nearest 3 utterances adjacent to the target as the positive samples, and set the batch size values for all CER methods to 8. Negative samples are obtained from other cross-dialogues in the same batch that are not in the same conversation as the target utterance. The number of conversations used to construct negative samples ranges from 1 to 7. The test results are shown in Table 6. Both methods achieve the best performance when the size of negative sample is 4. This indicates that when the number of positive samples is fixed, the selection of matching negative cases is conducive to better results, and it is not appropriate to have too many or too few negative samples. A similar trend also holds for the MELD dataset.

TABLE 5
Influence of Contrast Loss Coefficient λ on Performance of DAG [15] Between Different Datasets (MELD & IEMOCAP)

Dataset \ λ	$1e^{-3}$	$1e^{-2}$	$1e^{-1}$	$1e^{0}$	$1e^{1}$	$1e^{2}$
IEMOCAP	68.1	68.2	**68.9**	67.2	65.3	61.9
MELD	63.7	63.6	63.7	63.5	**63.8**	60.0

The F-scores of the methods are reported.

TABLE 6
The Influence of the Number of Negative Samples on the Emotion Classification Performance

Method \ #dialogue	1	2	3	4	5	6	7
DialogueRnn+C3ER	63.5	63.7	64.0	**65.0**	64.0	64.9	64.5
AGHMN+C3ER	60.6	60.8	60.5	**62.4**	60.9	61.1	60.8
DAG+C3ER	68.5	68.2	68.2	**68.9**	68.6	68.3	68.5

F1-scores is used as evaluation index. "#dialogue" represents the number of conversation selected as negative samples.

6.6 Perturbation Test

In order to figure out where the performance gain of the C3ER comes from and quantify the influence of C3ER on the backbone CER framework's performance, we carry out a series of perturbation tests on IEMOCAP dataset, which is known more sensitive to contextual information.

6.6.1 Test Setting

Since it is hard to quantify the context semantic consistency between target utterance and its conversational context, we follow the context-replacement approach used in our empirical study Section 4, to indirectly validate the performance of C3ER through perturbation tests by replacing the conversational context of the target utterance under different context-replacement modes.

More concretely, the main idea of the perturbation test is to modify the context of the target utterance in the test data, and then find out the performance change of the trained model. When the trained CER models classify the target utterance in a conversation, its conversational context is replaced with cross-dialogue content, which can indirectly reflect the how the CER model deal with its context. We also designed three different replacement modes, which are detailed in Section 4. The EM Mode maintains the original label patterns, and the RM and AM is designed to destroy the patterns inversely, AM is a more extreme case of RM. The performance degradation of different models under those settings can indirectly indicate how they react to the influence of "label copy" patterns. Whether the C3ER is able to significantly improve the performance of CER models under the different perturbation settings, will serve as a strong evidence to prove the function of C3ER.

During the experiments, the checkpoints for each model are chosen by training with same hyper-parameters, instead of the grid search. Furthermore, we chose 5 different random seeds to test results and report the average F1-score. We also conduct the significant test under the hypothesis that the CER models without C3ER have a better performance than the models with C3ER, and the statistical significance result is reported by permutation test with $p < 0.1$.

6.6.2 Results and Analysis

The perturbation test results are shown in Table 7. According to the observations we have made in the empirical analysis of the baseline models (see Section 4), we divide them into two types, called *Label Copying Approaches* (i.e., BC-LSTM, DialogueRnn and AGHMN) and *Non-label Copying Approaches* (i.e., DAG) respectively. We will discuss them separately below.

TABLE 7
Comparison of Different Models' F1-Scores Performance and Corresponding Significance Test Under the Perturbation Test

Methods \ Operation	No operation	EM	RM	AM
BC-LSTM	57.3	54.6	40.5	34.3
BC-LSTM+C3ER	57.6†	56.8†	47.2†	40.7†
DialogueRnn	62.2	60.0	35.1	30.5
DialogueRnn+C3ER	63.9†	62.2†	52.2†	41.1†
AGHMN	59.1	54.0	31.1	25.7
AGHMN+C3ER	60.3†	56.2†	36.5†	31.3†
DAG	67.3	66.0,	63.2	62.3
DAG+C3ER	68.5†	67.4	63.0	62.4

† *represents the C3ER achieves significant improvements over the baselines.*

TABLE 8
Comparison of F1-Scores Performance for DAG [15] With/Without C3ER Module Under Different Modes of Conversational Context Modification

Operation \ Model	DAG	DAG+C3ER
No Operation	67.3	68.5†
Structure	63.1	65.5†
Content	63.2	63.0
Structure+Content	62.9	63.5

"Structure" represents the structural information of a conversation is broken; and "Content" represents the actual content of a target utterance is replaced; "Structure+Content" means two operations are performed simultaneously. † *represents the C3ER achieves significant improvements over the baselines.*

Label Copying Approaches. BC-LSTM, DialogueRnn and AGHMN show similar trends under the different settings regardless of whether C3ER is injected. When the EM perturbation is applied, the performance of the baseline methods with the C3ER module added almost do not decrease, and the decrease is slight when C3ER is not added. In case of the RM and AM perturbation, the F1-scores of BC-LSTM, DialogueRnn and AGHMN are reduced sharply, and the performance deteriorates more in AM than RM mode. After adding the C3ER module, the performance degradation is eased significantly. And significant test further confirms the above trend.

Therefore, C3ER is like a regularization module, helping label-copying CER frameworks handle conversational context information more flexibly and at the same time enhance their robustness. For the baseline models that tend to learn the "label copying" patterns, the performance after the emotion-relevant perturbation will be less degraded, while the random perturbation will destroy the label pattern, resulting in a sharp performance decline. On the contrary, for a context-aware model where C3ER is injected, in the case of emotion-related context perturbation, i.e., EM mode, the model can make reasonable use of label patterns to maintain a high classification accuracy. When the utterance's context is randomly or entirely replaced with both content- and emotionally irrelevant information (i.e., RM and AM mode), the C3ER method could help CER model reduce dependency on the label pattern, so as to maintain classification performance to some extent.

Non-Label Copying Approaches. For the baseline DAG [15], we observe an interesting phenomenon that the perturbation no matter under EM, RM, AM operation, has less significant influence on its emotion classification performance compared to other three baselines. This indicates that DAG has little "label copy" effect compared to the other three models.

More experimental results of DAG with respect to extra modes of perturbation on structure and content are reported in the Table 8. According to the principles and framework structure of DAG, we know that DAG mainly focuses on the construction of structural information of a conversation, that is, the relationship between speakers established by a directed acyclic graph. Therefore, in order to figure out where the gains come from, we do a further ablation experiment and design 3 different modes to break the conversational context for DAG. The "content" mode is regarded as the destruction of actual content of target utterance's conversational context, which is same with "RM" mode described in Section 4.1. For the "structure" mode, we take randomly replace the edges constructed by the DAG and speaker information as the destruction of the structural context of the dialogue. The third mode (i.e., structure+content) is a mixture of the above two.

The experimental results are shown in Table 8. In case where both of the structure and actual content information of conversational context are broken, the F-scores of the model with or without C3ER is injected shows no significant difference. When only the structural information of a dialogue is destroyed, the model with C3ER injected have a significant improvement. We suspect that C3ER can help the CER model capture the structure of a conversation inferred from the semantic context, so as to reduce dependence on the conversational structure. Hence, DAG with C3ER can maintain high classification accuracy even when the structural context of a dialogue is destroyed. This phenomenon also supports the conjecture in Section 6.5.

In a summary, we have conducted a quantitative analysis of how different CER models use the conversational contextual information. The experimental results prove that our C3ER method has a certain degree of versatility, which can help understand various types of conversational context and improve the backbone CER model's performance. For the label-copying CER methods, C3ER mainly aims to avoid over-fitting the label patterns. For non-label copying methods(i.e., DAG), it trends to act as a semantic enhancement module in helping the CER model to reduce dependence on the conversational structure.

6.7 Case Study

In order for a more intuitive understanding of the effect of our method, we select an example conversation from IEMOCAP as a case, to show how C3ER influences its attention in comparison with DialogueRnn, which correspond to a large performance improvement. Fig. 7 shows the emotion classification results and attention visualization of the selected case. In this conversation, 13 out of 23 utterances are labeled with emotion "frustrated". DialogueRnn exhibits a label-copy effect, which directly classifies almost all utterances into "frustrated" to obtain the highest classification accuracy. When C3ER is added, the performance of the model has been improved. Not only the prediction is correct for most of the

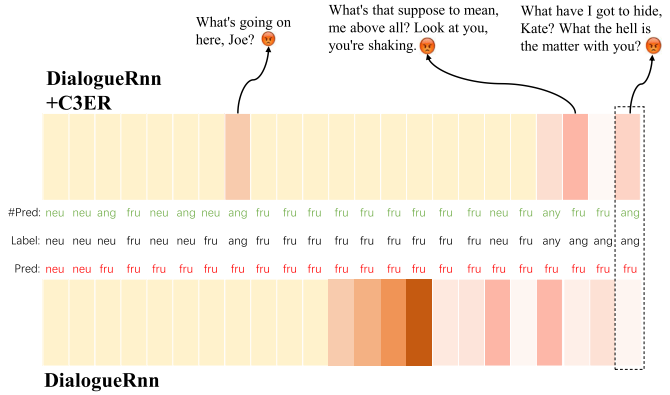


Fig. 7. Visualization of the DialogueRnn's attentions under the different settings whether C3ER is injected. The lines "#Pred" represents the emotion classification results of DialogueRnn with C3ER, "Pred" is that of without C3ER, and between them is the ground-truth labels. The attention values are from the models when predicting the emotion of 23th utterance marked with a dashed box in the dialogue.

"frustrated" utterances, but a considerable part of "angry" and "neutral" utterances are also classified correctly.

Specifically, we analyze the attention distribution of the two models (i.e., DialogueRnn with and without C3ER) in recognizing the last utterance of the conversation. DialogueRnn without C3ER incorrectly classifies the emotion of "angry" as "frustrated". Its attention is more local, focusing more on the continuous utterance with main emotion label "frustrated". After adding the C3ER module, the model seems to be able to capture more longer-term dependencies, focusing most of the attention on the 8th and 21st utterances marked with the text as shown in Fig. 7, thereby achieving correct classification. It is interesting that the utterances which the C3ER-enhanced model mainly focuses on have a similar language expression in addition to the fact that their emotions are all "angry". This indicates that the C3ER-enhanced model conducts a context-aware emotional reasoning by synthesizing context and the utterances whose expressions are similar to its own.

7 Conclusion

In this paper, we fisrt conduct an extensive empirical investigation of the effect of conversational context on the performance of conversational emotion recognition models. The results reveal that the representative CER models we have studied tend to overfit certain individual aspects of context, e.g., the emotion labels and intra/inter-speaker structures, but lack a holistic understanding of the conversational context, specially semantic context.

To solve this problem, we proposed a semantic-guided regularization method, namely C3ER, which can be integrated into a baseline CER model via contrastive learning and joint training. C3ER is based on contrastive learning in a self-supervised manner. It extracts the context-relevant/irrelevant utterances from intra/cross-dialogues for each utterance context as contrast pairs. Then contrastive learning is employed to establish the explicit semantic connection between utterances and their conversational context. This mechanism can be injected into a CER framework and jointly trained with emotion classification through multi-task learning, forcing the CER model to perform an emotional reasoning from the perspective of context understanding. Extensive experiments demonstrate that the idea of explicitly adding context understanding constraints to CER models is helpful to improve the classification accuracy and robustness.

8 Future Work

(1) Building a comprehensive CER framework that can integrate multiple influence factors. Conversational emotion recognition is a task largely affected by multiple factors such as semantic context (e.g., content emotion, dialogue topic, intent, etc.) and conversational structure (e.g., inter-speaker influence, intra-speaker influence, etc.). In the future, we will establish more quantitative analysis mechanisms for each influencing factor, rather than just heuristic exploration. Based on the findings, we can construct an holistic emotion decision-making framework that considers a range of factors to improve the performance of CER.

(2) Constructing more hard contrast samples to promote contextual learning. How to construct a more effective set of contrast pairs is the key to contrastive learning. In the future, it is promising to construct more complex and indistinguishable contrast sample pairs based on the characteristics of conversational emotion recognition, to further improve the CER model's understanding of conversational context.

(3) Incorporating external commonsense knowledge to conversational emotion recognition. Commonsense knowledge is also an important factor affecting the emotion of the conversation beyond the conversational context explored in our paper. Studying the relationship and a seamless integration between exploiting the commonsense knowledge and understanding the conversational context is an interesting topic that deserves a more in-depth further investigation.

References

[1] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell., 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 730–739.

[2] J. Wang, X. Sun, M. Wang, and M. Wang, "Emotional conversation generation with bilingual interactive decoding," *IEEE Trans. Comput. Soc. Syst.*, vol. 9, no. 3, pp. 818–829, Jun. 2022.

[3] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1426–1439, 3rd Quart. 2022.

[4] S. Poria, N. Majumder, R. Mihalcea, and E. H. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[5] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

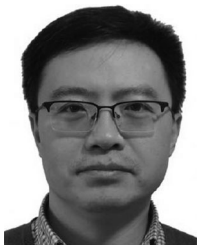
[6] ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection, Brussels, Belgium.

[7] HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition, Minneapolis, Minnesota.

[8] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. F. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell.*, 2019, pp. 6818–6825.

[9] W. Jiao, M. R. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," in *Proc. 34th AAAI Conf. Artif. Intell., 32nd Innov. Appl. Artif. Intell. Conf., 10th AAAI Symp. Educ. Adv. Artif. Intell.*, 2020, pp. 8002–8009.
[10] J. Li, Z. Lin, P. Fu, Q. Si, and W. Wang, "A hierarchical transformer with speaker modeling for emotion recognition in conversation," *CoRR*, 2020.
[11] DialogXL: All-in-one XLNet for multi-party conversation emotion recognition, vol. 35.
[12] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 985–1000, Jan. 2021.
[13] DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation," Hong Kong, China.
[14] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 5415–5421.
[15] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 1551–1560.
[16] MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations, Florence, Italy.
[17] C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio, "Do neural dialog systems use the conversation history effectively? An empirical study," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 32–37.
[18] C. Hao, L. Pang, Y. Lan, F. Sun, J. Guo, and X. Cheng, "Ranking enhanced dialogue generation," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 465–474.
[19] D. Ghosal, N. Majumder, R. Mihalcea, and S. Poria, "Utterance-level dialogue understanding: An empirical study," *CoRR*, 2020.
[20] IEMOCAP: Interactive emotional dyadic motion capture database, vol. 42.
[21] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," *CoRR*, 2015.
[22] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1426–1439, 3rd Quart. 2022.
[23] D. Ghosal, N. Majumder, A. F. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: Commonsense knowledge for emotion identification in conversations," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 2470–2481.
[24] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 165–176.
[25] J. Wang et al., "Sentiment classification in customer service dialogue with topic-aware multi-task learning," in *Proc. 34th AAAI Conf. Artif. Intell., 32nd Innov. Appl. Artif. Intell. Conf., 10th AAAI Symp. Educ. Adv. Artif. Intell.*, 2020, pp. 9177–9184.
[26] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 1571–1582.
[27] D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7042–7052.
[28] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang, "A multi-view network for real-time emotion recognition in conversations," *Knowl.-Based Syst.*, vol. 236, 2022, Art. no. 107751.
[29] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, 2022.
[30] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.
[31] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *J. Mach. Learn. Res.*, vol. 13, pp. 307–361, 2012.
[32] J. Y. Zou, D. J. Hsu, D. C. Parkes, and R. P. Adams, "Contrastive learning using spectral methods," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2238–2246.
[33] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
[34] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[35] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, 2018.
[36] L. Logeswaran and H. Lee, "An efficient framework for learning sentence representations," in *Proc. 6th Int. Conf. Learn. Representations*, 2018.
[37] P. Cheng, W. Hao, S. Yuan, S. Si, and L. Carin, "FairFil: Contrastive neural debiasing method for pretrained text encoders," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.
[38] S. Lee, D. B. Lee, and S. J. Hwang, "Contrastive learning with adversarial perturbations for conditional text generation," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.
[39] L. Xiong et al., "Approximate nearest neighbor negative contrastive learning for dense text retrieval," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.
[40] H. Cai et al., "Group-wise contrastive learning for neural dialogue generation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 793–802. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.70
[41] C.-S. Wu, S. C. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 917–929.
[42] J. Liu et al., "Topic-aware contrastive learning for abstractive dialogue summarization," *CoRR*, 2021.
[43] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, 2019.

Hanqing Zhang received the BE degree from Harbin Engineering University, Harbin, China, in 2017, and the ME degree from the Information System and Security and Countermeasures Experimental Center, Beijing Institute of Technology, in 2020. Now he is working toward the PhD degree with the School of Computer Science & Technology, Beijing Institute of Technology. His current research interests include affective computing, controlled text generation and emotional dialogue generation.



Dawei Song received the PhD degree from The Chinese University of Hong Kong, in 2000, and became a full professor with The Robert Gordon University, U.K., in 2008. He is currently a professor in computer science with the Beijing Institute of Technology, China. He has also worked as a professor with The Open University U.K., and Tianjin University, China. His research interest has been focused on information retrieval and natural language processing, especially novel models and computational methods to support intelligent access, retrieval and understanding of complex and multi-modal information in a way that is compatible with human cognitive information processing. He has published more than 200 research papers, including those on prestigious journals such as *ACM Transactions on Information Systems* and *IEEE Transactions on Knowledge and Data Engineering*, and top conferences in natural language processing and artificial intelligence such as SIGIR, WWW, ACL, AAAI, IJCAI and NeurIPS.