Hybrid Curriculum Learning for Emotion Recognition in Conversation

Lin Yang*, **Yi Shen***, **Yue Mao, Longjun Cai**

Alibaba Group, Beijing, China
{yl176562, sy133447, maoyue.my, longjun.clj}@alibaba-inc.com

Abstract

Emotion recognition in conversation (ERC) aims to detect the emotion label for each utterance. Motivated by recent studies which have proven that feeding training examples in a meaningful order rather than considering them randomly can boost the performance of models, we propose an ERC-oriented hybrid curriculum learning framework. Our framework consists of two curricula: (1) conversation-level curriculum (CC); and (2) utterance-level curriculum (UC). In CC, we construct a difficulty measurer based on "emotion shift" frequency within a conversation, then the conversations are scheduled in an "easy to hard" schema according to the difficulty score returned by the difficulty measurer. For UC, it is implemented from an emotion-similarity perspective, which progressively strengthens the model's ability in identifying the confusing emotions. With the proposed model-agnostic hybrid curriculum learning strategy, we observe significant performance boosts over a wide range of existing ERC models and we are able to achieve new state-of-the-art results on four public ERC datasets.

Introduction

Emotion recognition in conversation (ERC) has attracted numerous interests from the NLP community in recent years due to its potential applications in many areas, such as opinion mining in social media (Chatterjee et al. 2019), dialogue generation (Huang et al. 2018) and fake news detection (Guo et al. 2019). The objective of ERC is to detect emotions expressed by the speakers in each utterance of the conversation. Previous works on ERC usually solve this problem with two steps. At the first step, each utterance is encoded separately into an utterance-level representation, which will be used as the input for sequence-based models (Majumder et al. 2019; Hazarika et al. 2018a; Jiao et al. 2019) or graph-based models (Ghosal et al. 2019; Ishiwatari et al. 2020) during the second step. Despite their success, previous works still have a lot of room for improvement (Poria et al. 2019b).

Curriculum learning (CL) (Bengio et al. 2009) is a training strategy which imitates the meaningful learning order in human curricula. The core idea of CL is to train the machine learning model with easier data subsets at first, and then gradually increase the difficulty level of data until the whole training dataset. As an easy-to-use plug-in, the CL strategy has demonstrated its power in improving the overall performance of various models in a wide range of scenarios (Wang, Chen, and Zhu 2020). Inspired by the success of CL in other NLP tasks (Zhou et al. 2020; Liu et al. 2018; Su et al. 2021), in this paper, we make effort to leverage the spirit of CL to improve the traditional ERC methods. Due to the hierarchical structure of the ERC datasets, we need to construct the curricula from two granularities: one curriculum sorts the conversations in the dataset from easy to hard, and the other sorts the utterances in each conversation from easy to hard.

The question arises how to measure the difficulty of conversations and utterances. Previous studies (Majumder et al. 2019; Shen et al. 2021a) have reported that most ERC methods mainly suffer from two issues: 1) *"emotion shift" problem*. It refers to that these methods cannot efficiently handle scenarios in which emotions of two consecutive utterances are different (Ghosal et al. 2021). 2) *"confusing label" problem*. Previous methods (Ghosal et al. 2019; Shen et al. 2021b) usually fail to distinguish between similar emotions very well. This is due to the subtle semantic difference between certain emotion labels such as *happy* and *exciting*. These two phenomena provide us the key to quantify the difficulty of conversations and utterances in ERC.

In this paper, we tailor-design a hybrid curriculum learning (HCL) framework for the ERC task. HCL framework consists of two complementary curriculum strategies, conversation-level curriculum (CC) and utterance-level curriculum (UC). In CC, we construct a difficulty measurer based on "emotion shift" frequency within a conversation, then the conversations with lower difficulty are presented to the model before harder ones. This way, the model gradually increases its ability to tackle the "emotion shift" problem.

While in UC, since ERC requires reasoning over multiple utterances in the conversation, we cannot directly schedule the utterances asynchronously in the "easy to hard" scheme. As a result, we design an emotion-similarity based curriculum (ESC) to implement utterance-level curriculum learning. Specifically, inspired by the *"confusing label" problem* mentioned above, we believe that in a conversation, those utterances with confusing labels are more difficult than others. Therefore, we make the model focus on the utterances with

*These authors contributed equally.

easily recognizable emotion labels in the early stage, then progressively strengthened the model's capability of identifying the confusing emotions.

More specifically, based on previous studies (Plutchik 1982; Mikels et al. 2005) on psychology, we employ the degree of intersection angle between different emotion labels in Valence-Arousal 2D emotion space (Guo et al. 2019; Yang et al. 2021) to measure the similarity between emotion labels. During ESC, instead of one-hot encoding, the target represents a probability distribution over all possible emotion labels. The probability of each label is determined by the similarity between current label and the gold label. In other words, instead of solely belonging to its true emotion label, each utterance can also belong to similar emotions to a lesser extent. In the beginning of the training process, the targets of utterances with emotions *happy* and *excited* should almost be the same, but always be very different from *sad*. During the training process, the label representation gradually shifted to the one-hot encoding. This way, small mistakes are corrected less than big mistakes in the beginning, which resembling a curriculum in which broad concepts are explained before subtle differences are emphasized.

Our hybrid curriculum learning framework is model-agnostic. We evaluate our approach on five representative ERC models. Results on four benchmark datasets demonstrate that the proposed hybrid curriculum learning framework leads to significant performance improvements.

In summary, our main contributions are as follows:

- We propose a hybrid curriculum learning framework to tackle the task of ERC. At conversation-level curriculum, we utilize an emotion-shift frequency to measure the difficulty of each conversation.
- We propose emotion-similarity based curriculum learning to achieve utterance-level curriculum learning. It implements the basic idea that at early stage of training it is less important to distinguish between similar emotions compared to separating very different emotions.
- We conduct experiments on four ERC benchmark datasets. Empirical results show that our proposed hybrid curriculum learning framework can effectively improve the overall performance of various ERC models, including the state-of-the-art.

Related Work

Emotion Recognition in Conversation

Emotion recognition in conversations (ERC) has been widely studied due to its potential application prospect. The key point of ERC is how to effectively model the context of each utterance and corresponding speaker. Existing works generally resort to deep learning methods to capture contextual characteristics, which can be divided into sequence-based and graph-based methods. Another direction is to improve the performance of existing models by incorporating various external knowledge, which we classified as knowledge-based methods.

Sequence-based Methods Many previous works consider contextual information as utterance sequences. ICON (Hazarika et al. 2018a) and CMN (Hazarika et al. 2018b) both utilize gated recurrent unit (GRU) to model the utterance sequences. DialogueRNN (Majumder et al. 2019) employs a GRU to capture the global context which is updated by the speaker state GRUs. Jiao et al. (2019) propose a hierarchical neural network model that comprises two GRUs for the modelling of tokens and utterances respectively. Hu, Wei, and Huai (2021) introduce multi-turn reasoning modules on Bi-directional LSTM to model the ERC task from a cognitive perspective.

Graph-based Methods In this category, some existing works (Ghosal et al. 2019; Ishiwatari et al. 2020; Zhang et al. 2019) utilize various graph neural networks to capture multiple dependencies in the conversation. DialogXL (Shen et al. 2021a) modifies the memory block in XLNet (Yang et al. 2019) to store historical context and leverages the self-attention mechanism in XLNet to deal with the multi-turn multi-party structure in conversation. Shen et al. (2021b) design a directed acyclic graph (DAG) to model the intrinsic structure within a conversation, which achieves the state-of-the-art performance without considering the introduction of external knowledge.

Knowledge-based Methods KET (Zhong, Wang, and Miao 2019) employs hierarchical transformers with concept representations extracted from the ConceptNet (Speer and Lowry-Duda 2017) for emotion detection, which is the first ERC model integrates common-sense knowledge. COSMIC (Ghosal et al. 2020) adopts a network structure very close to DialogRNN and adds external commonsense knowledge from ATOMIC (Sap et al. 2019) to improve its performance. TODKAT (Zhu et al. 2021) leverages an encoder-decoder architecture which incorporates topic representation with commonsense knowledge from ATOMIC for ERC.

Curriculum Learning

Starting from the work by Bengio et al. (2009), a variety of curriculum learning approaches (Wang, Chen, and Zhu 2020; Soviany et al. 2021) has been studied. In the field of NLP, curriculum learning has been used for various tasks such as neural machine translation (Zhou et al. 2020; Liu et al. 2020), relation extraction (Huang and Du 2019) and natural answer generation (Liu et al. 2018). To the best of our knowledge, we leverage curriculum learning for the first time in the ERC task.

Proposed Framework

Task Definition

In ERC, a conversation *C* contains a sequence of textual utterances $\{u_1, u_2..., u_n\}$, where n denotes the number of utterances. Each utterance $u_i = \{w_{i,1}, w_{i,2}..., w_{i,t(u_i)}\}$ consists of $t(u_i)$ tokens, where $t(u_i)$ is the length of u_i. There are m participants $P = \{p_1, p_2..., p_m\}(m \geq 2)$ in *C*. Each utterance u_i is uttered by one participant in P . Given a predefined emotion label set $E = \{y_1, y_2, ..., y_r\}$, the objective of the ERC task is to predict the emotion label of each utterance in *C* with the information provided above.

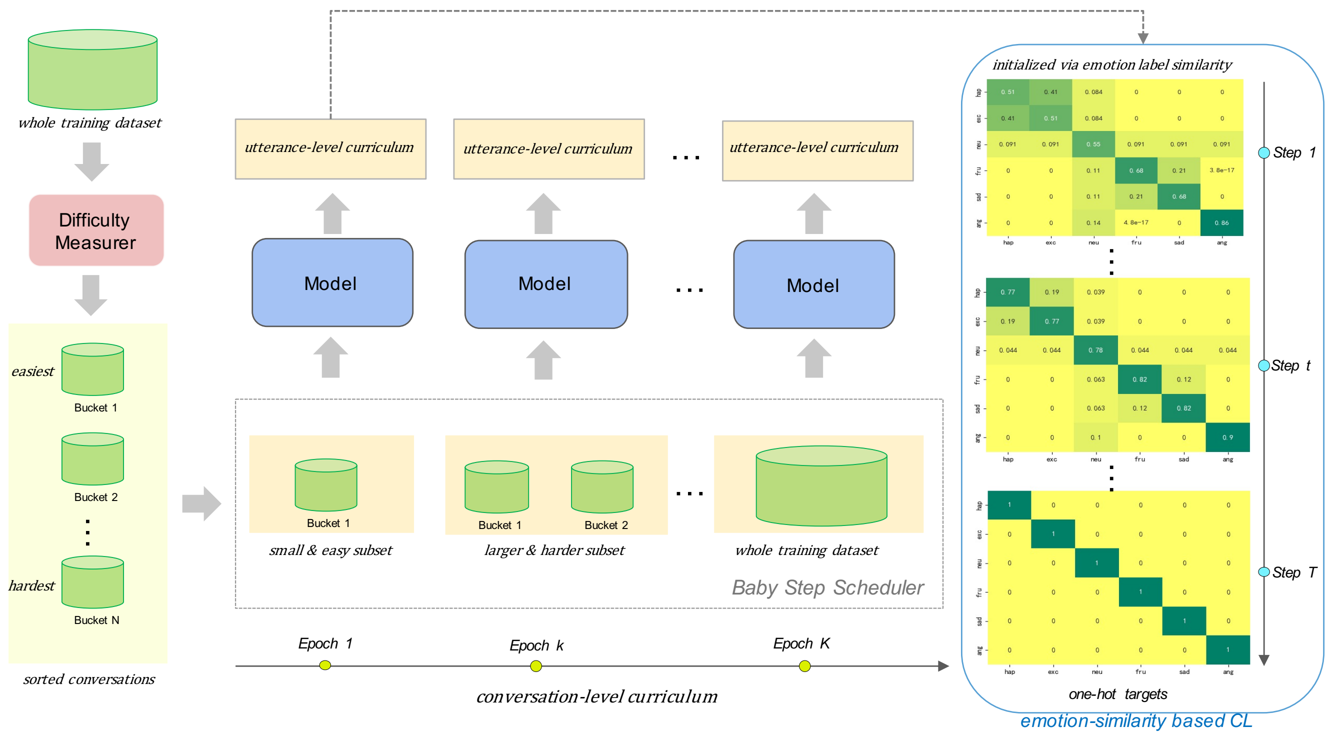

Figure 1: The proposed hybrid curriculum learning (HCL) framework for ERC.

Overview

In curriculum learning, a typical curriculum design consists of two core components: difficulty measurer and training scheduler (Bengio et al. 2009). Difficulty Measurer is used to quantify the relative “easiness” of each data example. The training scheduler arranges the sequence of data subsets throughout the training process based on the judgment from the difficulty measurer. For ERC oriented curriculum learning, the challenge is how to design suitable difficulty measurer and training scheduler for emotion recognition.

A conversation consists of a sequence of utterances. This hierarchical structure inspired us to construct two curricula for scheduling conversations and utterances respectively. Therefore, our framework consists of two nested curricula, conversation-level curriculum (CC) on the outside and utterance-level curriculum (UC) on the inside.

For CC, we design an emotion-shift based difficulty measurer. A widely used CL strategy called *baby step* (Spitkovsky, Alshawi, and Jurafsky 2010) is leveraged as training scheduler. For UC, due to the characteristics of the ERC task, the utterances in the same conversation must be input into a batch simultaneously during the training process. As a result, it is infeasible to employ traditional training scheduler such as *baby step* to arrange the training order of the utterances. We proposed emotion-similarity based curriculum learning to address this issue.

The proposed HCL framework is illustrated in Figure 1 and the details of CC and UC are elaborated in following two subsections, respectively.

Conversation-level Curriculum

To design conversation-level curriculum for ERC, we need to answer a question: what kind of conversation is supposed to be easier than other conversations? Since we have mentioned that previous ERC models (Majumder et al. 2019; Shen et al. 2021a) tend to suffer from emotion-shift issue, we adopt emotion-shift frequency to measure the difficulty of each conversation. The main idea is that, the more frequent emotion-shift in conversation c_i occurs, the more difficult it is. Therefore, the conversation-level difficulty score of c_i is defined as

$$d_{cc}(c_i) = \frac{N_{es}(c_i) + N_{sp}(c_i)}{N_u(c_i) + N_{sp}(c_i)}, \tag{1}$$

where $N_{es}(c_i)$ and $N_u(c_i)$ denote the number of emotion-shift occurrences in c_i and the total number of utterances in c_i, respectively. $N_{sp}(c_i)$ is the number of speakers take part in c_i and it acts as a smoothing factor.

We leverage baby step training scheduler (Spitkovsky, Alshawi, and Jurafsky 2010) to arrange conversations and organize the training process. Specifically, the whole training set D is divided into different buckets, i.e. $\{D_1, \cdots, D_T\}$, in which those conversations with similar difficulty scores are categorized into the same bucket. The training starts from the easiest bucket. After a fixed number of training epochs or convergence, the next bucket is merged into the current training subset. Finally, after all the buckets are merged and used, the whole training process further continues several extra epochs. Our HCL framework is described in Algorithm 1 and the process of CC is illustrated as *Line 1-Line 5*.

Utterance-level Curriculum

As it is infeasible to employ a traditional CL training scheduler to asynchronously arrange the order of the utterances, the question arises how to measure the difficulty of the utterances and establish a feasible curriculum at utterance-level.

We address this problem by assuming that the utterances with confusing emotion labels are more difficult for prediction and our utterance-level curriculum is based on the pairwise similarities between the emotion labels.

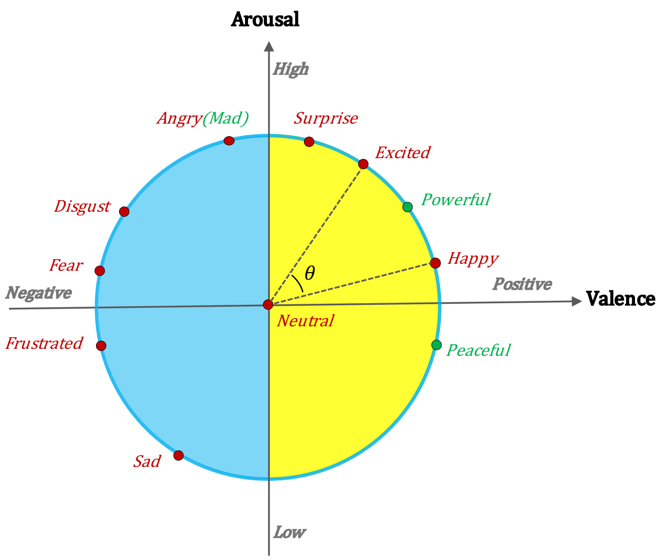

Figure 2: The 2D arousal-valence emotion wheel proposed by us. Each emotion label is corresponding to a point on the unit circle. This wheel has integrated the versions from (Jing, Mao, and Chen 2019; Yang et al. 2021; Toisoul et al. 2021). The emotions in red color have appeared in previous versions. The emotions in green color is what we added (these emotions only appear in EmoryNLP dataset). θ denotes the intersection angle between *happy* and *excited*. The angle between similar emotions will be relatively small.

Previous studies (Plutchik 1982; Mikels et al. 2005; Russell 1980) on psychology believe that emotion contains two dimensions: arousal and valence, and they are used to leverage a wheel-like 2D coordinate system to describe emotions. Inspired by these works, we propose a new emotion wheel as Figure 2, which contains all emotions in the standard ERC datasets. As depicted in Figure 2, each emotion label can be mapped to a point on the unit circle. Then we calculate the similarity between emotion labels as in Equation 2.

$$s_{ij} = \begin{cases} \max(\cos(\theta_{ij}), 0) & v_i \cdot v_j > 0 \\ 0 & v_i \cdot v_j < 0 \\ 1/N & v_i \cdot v_j = 0 \end{cases} \tag{2}$$

Here, s_{ij} stands for the similarity of label i and label j. v_i denotes the valence value of i. We take the cosine of the included angle θ_{ij} between i and j as their similarity. If $\theta_{ij} > 90°$ (i.e., $cos\theta_{ij} < 0$) the similarity is set to 0. If the valence polarities of i and j are opposite, then the similarity is also set to 0. The similarity between label *neutral* and other labels is defined as $1/N$, where N is the total number of emotions in corresponding datasets.

The process of emotion-similarity based curruclum learning (ESC) is described as *Line 6 - Line 13* in Algorithm 1. We first calculate the similarity between each emotion label pair as Equation 2 and generate the emotion similarity matrix $\mathcal{M}_{sim}$, then $\mathcal{M}_{sim}$ is normalized as $\mathcal{M}_{target}$. At the beginning of ESC training, we take the rows of $\mathcal{M}_{target}$ as the initial target probability distributions over all possible classes for training, and each row corresponds to an emotion label. That is, instead of solely belonging to its ground-truth label, each input utterance can also belong to similar labels to a lesser extent. During the training process, this label representation is gradually shifted towards the standard one-hot-encoding. We define the update strategy as in *Line 9 - Line 11*, where $m_{i,j}$ denotes the probability of j-th element of i-th row in $\mathcal{M}_{target}$ at training step t. The constant parameter $\epsilon \in (0, 1)$ controls how quickly the label vectors converge to the one-hot-encoded labels. Row-wise normalization is performed after each update. This update strategy leads to a proper label-weighting curriculum.

$$\mathcal{L}(\theta) = -\sum_{c=1}^{z}\sum_{i=1}^{n}\sum_{k=0}^{m} \mathcal{M}_{target}[y_{u_i}^c]_k \log \mathcal{P}_{u_i}^c[k] \tag{3}$$

For each training step, the predicted probability distribution of utterance u_i defined as $\mathcal{P}_{u_i}$. Finally, the model is trained with the standard cross-entropy loss function as Equation 3, where $\mathcal{P}_{u_i}^c[k]$ denotes the predicted probability that the label of u_i in conversation c is k. $\mathcal{M}_{target}[y_{u_i}^c]_k$ denotes the target probability of label k in current label-similarity matrix at training step t. z is total number of conversations in training set, n is the utterance number of conversation c. In this way, we implement UC through ESC.

Experimental Settings

Datasets

We evaluate our method on the following four published ERC datasets [1]: **IEMOCAP** (Busso et al. 2008), **MELD** (Poria et al. 2019a), **DailyDialog** (Li et al. 2017), **EmoryNLP** (Zahiri and Choi 2018). The detailed statistics of the datasets are reported in Table 1 [2].

Following previous works (Ghosal et al. 2019; Zhong, Wang, and Miao 2019; Ishiwatari et al. 2020), the evaluation metrics are chosen as micro-F1 excluding the extremely high majority class (neutral) for DailyDialog and weighted-F1 for other three datasets.

Baselines

Since HCL is a model-agnostic framework, we choose the following five ERC models to verify whether HCL is able to further improve the performance of these models.

[1] These datasets are multi-modal datasets, we only focus on the textual information so as to be consistent with previous works.

[2] Some baseline methods made slight adjustments in data splits, we keep exactly the same settings as corresponding methods respectively for fair comparison.

Algorithm 1: Training Process with HCL

Input:
$\mathcal{D}$: whole training dataset;
$\mathcal{F}$: the difficulty measurer in CC;
k : the number of buckets in baby step scheduler;
$\mathcal{M}_{sim}$: the emotion similarity matrix in ESC
T : the max training step for ESC;
Δt: interval step for updating the targets in ERC;
ϵ: decay factor in ESC;
Output: M^* : the optimal model.

$\mathcal{D}^{'} = \text{sort}(\mathcal{D},\ \mathcal{F})$
$\{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^k\} = \mathcal{D}^{'}$ where $\mathcal{F}(d_a) < \mathcal{F}(d_b),\ d_a \in \mathcal{D}^i,\ d_b \in \mathcal{D}^j, \forall i < j$
$\mathcal{D}^{train} = \emptyset$
for $s = 1 \ldots k$ **do**
    $\mathcal{D}^{train} = \mathcal{D}^{train} \cup \mathcal{D}^s$
    $\mathcal{M}_{sim} = \{s_{ij}\},\ i,\ j = 1, \cdots, m$
    $\mathcal{M}_{target} = \frac{m_{ij}^{sim}}{\sum_{j=1}^{m} m_{ij}^{sim}},\ i, j = 1, \cdots, m$
    for $t = 1 \cdots T$ **do**
        if $t \,\%\, \Delta t = 0$ **then**
            $$m_{ij}^{'} = \begin{cases} \frac{1}{1+\epsilon \sum_{j \neq j} m_{i,j}},\ if\ ,\ i = j \\ \frac{\epsilon m_{ij}^{t}}{1+\epsilon \sum_{j \neq j} m_{i,j}},\ if\ ,\ i \neq j \end{cases}$$
            $\mathcal{M}_{target} = \{\frac{m_{ij}^{'}}{\sum_{j=1}^{m} m_{ij}^{'}}\},\ i, j = 1, \cdots, m$
        end
        train(M, $\mathcal{D}^{train}$, $\mathcal{M}_{target}$)
    end
end
return M

DialogueRNN (Majumder et al. 2019) This is a famous sequence-based ERC model, which uses three GRUs to model the speaker, the context given by the preceding utterances, and the emotion behind the preceding utterances.

DialogueGCN (Ghosal et al. 2019) This is a representative graph-based ERC model. It captures self-dependency and inter-speaker dependency by using two-layer graph neural networks.

DAG-ERC (Shen et al. 2021b) It is the state-of-the-art of all the ERC models that do not employ external knowledge. DAG-ERC utilizes directed acyclic graph to model the structure of a conversation.

COSMIC (Ghosal et al. 2020) It is a representative knowledge-based ERC model. It leverages external commonsense knowledge to improve the performance.

TODKAT (Zhu et al. 2021) This is the state-of-the-art knowledge-based ERC model. Besides commonsense knowledge, it also incorporates topic information.

Implementation Details

All of the baseline models mentioned above have released their source codes. We keep exactly the same settings as reported in the original papers during our experiments. For HCL, the tunable hyperparameters include number of buckets in CC, max training epochs during each baby step, interval steps for training target updating in UC, decay factor in UC. These hyperparameters are manually tuned on each dataset with hold-out validation. The results reported in our experiments are all based on the average score of 5 random runs on the test set. Our experiments are conducted on a single Tesla V100M32 GPU.

Results and Analysis

Overall Results

The overall experimental results are reported in Table 2, where "X+HCL" means training the model X with the proposed HCL framework. We can see that HCL has improved the performance of all baseline models, showing the robustness and universality of our approach.

In general, the performance boosts achieved by HCL on models with simpler feature extractor (i.e., DialogueRNN and DialogueGCN) are more remarkable. An exception is that TODKAT+HCL achieves significant improvements on three datasets. The reason may be that the original TODKAT model does not take account of the speaker information, while our CC has introduced the inter-speaker emotion-shift in the difficulty measurer, which is equivalent to considering speaker information to a certain extent and is beneficial for TODKAT.

Ablation Study

To reveal the individual effects of CC and UC, we try different variants of HCL on TODKAT by removing either CC or UC. The experimental results on IEMOCAP and EmoryNLP are shown in Table 3, from which we see that both CC and UC make positive contributions to the overall performance when used alone. Although only utilizing UC leads to larger improvements than only using CC, the optimal performance is achieved when CC and UC are combined, indicating that CC and UC are complementary to each other.

In addition, we also tried another two strategies to combine CC and UC: CC-First (CCF) and UC-First (UCF). CCF performs CC and UC in a pipeline manner. In UCF, the execute order of CC and UC is reversed. The results of CCF and UCF are also outlined in Table 3. It shows that UCF is better than CCF and HCL outperforms both CCF and UCF. This is intuitive, because UCF follows the order from fine-grained to coarse-grained, which is more in line with the "easy to hard" scheme in CL. Compared with UCF, HCL makes UC and CC interact with each other during the training process, which is more consist with the hierarchical structure of conversation, so the performance is even better than UCF.

Performance for Emotion-shift

To verify the effect of HCL in the emotion-shift scenario, we summarize the results of TODKAT+HCL on different types of utterances. The results are presented in Table 4, where

Datasets	Conversations			Utterances			classes	avg_utt	Evaluation
	Train	Val	Test	Train	Val	Test			
IEMOCAP	120		31	5810		1623	6	66.8	Weighted-F1
MELD	1038	114	280	9989	1109	2610	7	9.57	Weighted-F1
EmoryNLP	713	99	85	9934	1344	1328	7	14.05	Weighted-F1
DailyDialog	11118	1000	1000	87170	8069	7740	7	7.85	Micro-F1

Table 1: The statistics of datasets. avg_utt denotes the average number of utterances.

METHOD	IEMOCAP	MELD	DailyDialog	EmoryNLP
DialogueRNN	62.75	57.03	-	-
DialogueGCN	64.18	58.10	-	-
COSMIC	65.28	65.21	58.48	38.11
DAG-ERC	68.03	63.65	59.33	39.02
TODKAT	61.33	68.23	58.47	43.12
DialogueRNN+HCL	64.62 (↑ 1.87)	58.30 (↑ 1.27)	-	-
DialogueGCN+HCL	65.41 (↑ 1.23)	59.31 (↑ 1.21)	-	-
COSMIC+HCL	66.23(↑ 0.95)	65.85(↑ 0.64)	59.54(↑ 1.06)	38.96(↑ 0.85)
DAG-ERC+HCL	**68.73**(↑ 0.70)	63.89(↑ 0.24)	59.64(↑ 0.31)	39.82(↑ 0.80)
TODKAT+HCL	63.03(↑ 1.70)	**68.96**(↑ 0.73)	**59.76**(↑ 1.29)	**46.11**(↑ 2.99)

Table 2: The overall results on different methods on four datasets. The results of baseline methods are from the original papers.

METHOD	IEMOCAP	EmoryNLP
TODKAT	61.33	43.12
TODKAT+CC	61.83(↑ 0.50)	44.20(↑ 1.08)
TODKAT+UC	62.01(↑ 0.68)	45.19(↑ 2.07)
TODKAT+CCF	62.07(↑ 0.74)	45.06(↑ 1.94)
TODKAT+UCF	62.76(↑ 1.43)	45.47(↑ 2.35)
TODKAT+HCL	**63.03**(↑ 1.70)	**46.11**(↑ 2.99)

Table 3: Ablation study on TODKAT

METHOD	IEMOCAP		EmoryNLP	
	ES (41.2%)	N-ES (58.8%)	ES (69.2%)	N-ES (30.8%)
TODKAT	56.24	64.62	39.36	51.51
TODKAT+HCL	**56.91**	**67.01**	**42.40**	**54.02**

Table 4: The performance of TODKAT+HCL on utterances which exhibits emotion-shift. Numbers in parenthesis indicate the percentage in the test dataset.

ES and *N-ES* denote utterances with emotion-shift and utterances without emotion-shift, respectively. HCL improves the performance of TODKAT on both *ES* and *N-ES* of the two datasets. The improvement on ES in EmoryNLP is more significant than on ES in IEMOCAP.

A plausible explanation is that the training set of IEMOCAP contains much less conversations and the average length of conversations is much longer, so the difficulty scores of conversations in IEMOCAP are usually lower. Therefore, for IEMOCAP, the difficulty discrimination between different buckets in the training scheduler is not as obvious as EmoryNLP.

Performance on Different Emotions

In this subsection, we aim to verify whether HCL can improve the performance of baseline model on "confusing labels". For each pair of emotion labels in ERC dataset, if their similarity (defined in Equation 2) is larger than 0, then both of them are regarded as confusing labels in our setting. [3]. We report the results of DAG-ERC and DAG-ERC+HCL on every emotion label in IEMOCAP. There are a total of four confusing labels in this dataset: *happy*(H), *excited*(E), *sad*(S) and *frustrated*(F). As presented in Table 5, DAG-ERC+HCL outperforms DAG-ERC on all emotion labels other than *neutral* and the overall performance on the confusing labels is better (*69.37* vs *67.88* on weighted-F1). This shows that HCL does strengthen the ability on distinguishing the confusing emotion labels of DAG-ERC. However, the performace is limited by *neutral*, the reason is that *neutral* is similar to every other label to some extent as in Equation 2, which increases the difficulty for recognition.

Case Study

Figure 3(a) shows a conversation passage sampled from the IEMOCAP dataset. The goal is to predict the emotion label of the last utterance in the blue box. Due to emotion-shift occurs, all the baseline methods in our experiment are easy to mistakenly identify the emotion as *frustrated*. Most of our "X+HCL" methods are able to recognize the emotion of this utterance correctly, which indicates that HCL alleviates this problem to some extent. Figure 3(b) depicts a case with confusing labels. The gold emotion label of the last utterance in the red box is *excited*. Some of the baseline models such as DialogueGCN and DAG-ERC mistook the emotion as *happy*. After following HCL framework, Di-

[3] *Neutral* is not included

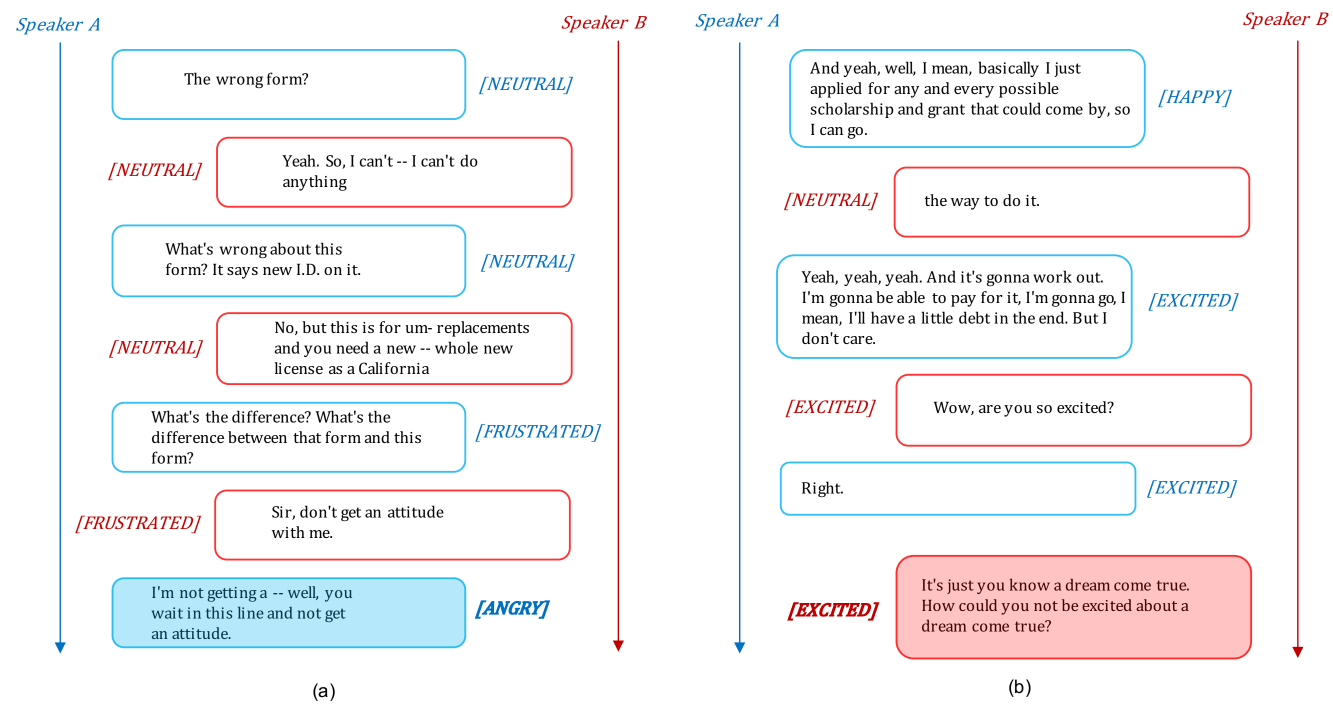

Figure 3: Two conversation passages from IEMOCAP for case study. The ground-truth emotion label of each utterance is given in the corresponding position. (a) An emotion-shift case. (b) A confusing label case.

Method	Happy (8.8%)	Sad (15.1%)	Neutral (23.7%)	Angry (10.5%)	Excited (18.4%)	Frustrated (23.5%)	HESF (65.8%)	NA (34.2%)
DAG-ERC	47.59	79.83	**69.36**	66.67	66.79	68.66	67.88	**68.53**
DAG-ERC+HCL	**48.97**	**82.21**	68.08	**66.72**	**69.43**	**68.73**	**69.37**	67.66

Table 5: Comparison of DAG-ERC and DAG-ERC+HCL on different emotions. Here *HESF* and *NA* denote "*Happy* + *Excited* + *Sad* + *Frustrated*" and "*Neutral* + *Angry*", respectively. Numbers in parenthesis indicate the percentage of each emotion label in the test dataset.

alogueGCN+HCL and DAG-ERC+HCL successfully identified the emotion as the correct label *excited*.

Why Curriculum Learning Works?

According to the theory of curriculum learning (Bengio et al. 2009), the curriculum will work only if the entropy of data distributions increases during the training process. In HCL, conversation-level curriculum leverages the emotion-shift frequency to measure the difficulty. The more frequent the emotion-shift occurs in a conversation, the greater the diversity of the emotion labels, in other words, the higher the entropy. For utterance-level curriculum, since emotion-similarity based CL does not distinguish similar emotion in the early stage, it is equivalent to merging some emotion labels and could be considered as reducing the diversity of emotions. As a result, it also meets the condition which the entropy should be increased gradually.

Conclusions

In this paper, we propose simple but effective hybrid curriculum learning (HCL) for emotion recognition in conversations. HCL is a flexible framework independent of the original training models. During training, HCL simultaneously employs conversation-level and utterance-level curricula to execute the training process as an easy to hard schema. Conversation-level curriculum consists of an emotion-shfit based difficulty measurer and a baby step scheduler. Utterance-level curriculum is implemented as emotion-similarity based CL. Experiments on four benchmark datasets have proved the generality and effectiveness of HCL.

In the future, we plan to improve our method in three directions. First, we will attempt to seek other suitable features to construct difficulty measurer for ERC. Second, we aim to introduce other training schedulers for CL to further improve the performance. Finally, we aim to apply a learning-based approach to model the similarity between emotion labels.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *ICML*, 41–48.

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.

Chatterjee, A.; Narahari, K. N.; Joshi, M.; and Agrawal, P. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *SemEval2019*, 39–48.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *EMNLP Findings*, 2470–2481.

Ghosal, D.; Majumder, N.; Mihalcea, R.; and Poria, S. 2021. Exploring the Role of Context in Utterance-level Emotion, Act and Intent Classification in Conversations: An Empirical Study. In *ACL Findings*, 1435–1449.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *EMNLP*, 154–164.

Guo, C.; Cao, J.; Zhang, X.; Shu, K.; and Liu, H. 2019. Dean: Learning dual emotion for fake news detection on social media. *arXiv preprint arXiv:1903.01728*.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018a. ICON: interactive conversational memory network for multimodal emotion detection. In *EMNLP*, 2594–2604.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *NAACL*, 2122–2132.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *ACL*, 7042–7052.

Huang, C.; Zaiane, O. R.; Trabelsi, A.; and Dziri, N. 2018. Automatic dialogue generation with expressed emotions. In *NAACL*, 49–54.

Huang, Y.; and Du, J. 2019. Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. In *EMNLP*, 389–398.

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 *EMNLP*, 7360–7370.

Jiao, W.; Yang, H.; King, I.; and Lyu, M. R. 2019. HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition. In *NAACL*, 397–406.

Jing, S.; Mao, X.; and Chen, L. 2019. Automatic speech discrete labels to dimensional emotional values conversion method. *IET Biometrics*, 8(2): 168–176.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP*, 986–995.

Liu, C.; He, S.; Liu, K.; Zhao, J.; et al. 2018. Curriculum Learning for Natural Answer Generation. In *IJCAI*, 4223–4229.

Liu, X.; Lai, H.; Wong, D. F.; and Chao, L. S. 2020. Norm-Based Curriculum Learning for Neural Machine Translation. In *ACL*, 427–436.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI*, 6818–6825.

Mikels, J. A.; Fredrickson, B. L.; Larkin, G. R.; Lindberg, C. M.; Maglio, S. J.; and Reuter-Lorenz, P. A. 2005. Emotional category data on images from the International Affective Picture System. *Behavior research methods*, 37(4): 626–630.

Plutchik, R. 1982. A psychoevolutionary theory of emotions. *Social sciences information*, 21(4-5): 529–553.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019a. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*, 527–536.

Poria, S.; Majumder, N.; Mihalcea, R.; and Hovy, E. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7: 100943–100953.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, 3027–3035.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021a. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *AAAI*, 13789–13797.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021b. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *ACL*, 1551–1561.

Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2021. Curriculum learning: A survey. *arXiv preprint arXiv:2101.10382*.

Speer, R.; and Lowry-Duda, J. 2017. ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *SemEval2017*, 85–89.

Spitkovsky, V. I.; Alshawi, H.; and Jurafsky, D. 2010. From baby steps to leapfrog: How “less is more” in unsupervised dependency parsing. In *NAACL*, 751–759.

Su, Y.; Cai, D.; Zhou, Q.; Lin, Z.; Baker, S.; Cao, Y.; Shi, S.; Collier, N.; and Wang, Y. 2021. Dialogue response selection with hierarchical curriculum learning. In *ACL*, 1740–1751.

Toisoul, A.; Kossaifi, J.; Bulat, A.; Tzimiropoulos, G.; and Pantic, M. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1): 42–50.

Wang, X.; Chen, Y.; and Zhu, W. 2020. A Survey on Curriculum Learning. *arXiv preprint arXiv:2010.13166*.

Yang, J.; Li, J.; Li, L.; Wang, X.; and Gao, X. 2021. A Circular-Structured Representation for Visual Emotion Distribution Learning. In *CVPR*, 4237–4246.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zahiri, S. M.; and Choi, J. D. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In *AAAI Workshops*.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *IJCAI*, 5415–5421.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *EMNLP*, 165–176.

Zhou, Y.; Yang, B.; Wong, D. F.; Wan, Y.; and Chao, L. S. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *ACL*, 6934–6944.

Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; and He, Y. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *ACL*, 1571–1582.