Contextualized Emotion Recognition in Conversation as Sequence Tagging

Yan Wang Jiayu Zhang Jun Ma Shaojun Wang Jing Xiao
Ping An Technology
{wangyanj61,zhangjiayu470}@pingan.com.cn
{majun,wangshaojun851,xiaojing661}@pingan.com.cn

Abstract

Emotion recognition in conversation (ERC) is an important topic for developing empathetic machines in a variety of areas including social opinion mining, health-care and so on. In this paper, we propose a method to model ERC task as sequence tagging where a Conditional Random Field (CRF) layer is leveraged to learn the emotional consistency in the conversation. We employ LSTM-based encoders that capture self and inter-speaker dependency of interlocutors to generate contextualized utterance representations which are fed into the CRF layer. For capturing long-range global context, we use a multi-layer Transformer encoder to enhance the LSTM-based encoder. Experiments show that our method benefits from modeling the emotional consistency and outperforms the current state-of-the-art methods on multiple emotion classification datasets.

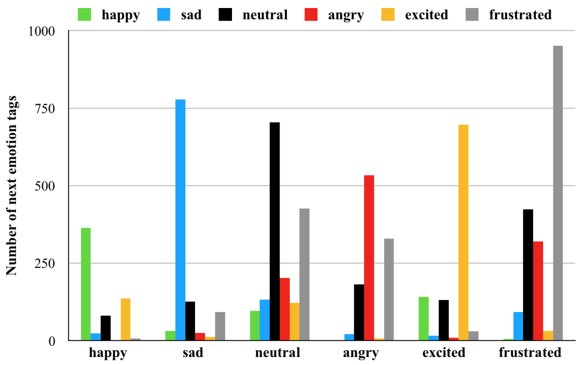

Figure 1: Emotional consistency on IEMOCAP (Busso et al., 2008). In a conversation, similar emotions tend to appear adjacently while dissimilar emotions seldom appear in the neighborhood. We call this phenomenon emotional consistency. For example, if the emotion of current utterance is happy, the tag of next utterance tends to be happy, excited or neutral rather than sad, angry or frustrated. This pattern also applies to other emotions.

1 Introduction

With the prevalence of conversation-based service, emotion recognition in conversation (ERC) has been attracting attention recently (Majumder et al., 2019; Zhong et al., 2019; Ghosal et al., 2019). Due to great potential in many scenarios such as recommendation system, customer service feedback and health-care, researchers keep focusing on empowering machine to understand emotions in conversation with emotional dynamics, which is a work with challenges lying in several aspects such as modeling the emotion inertia for each speaker and the influence of the interaction between speakers on emotional dynamics (Poria et al., 2019).

Recent works on ERC rely on recurrent neural networks (RNNs) to compute context-dependent representations of utterances (Poria et al., 2017; Majumder et al., 2019). Due to a carefully designed cell, RNNs like long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent unit (GRU) (Chung et al., 2014) memorize the sequential context to model the dependency between utterances. Such scheme of contextualized emotion recognition has shown its superiority in tracking emotional dynamics by modeling self and inter-speaker dependency in conversations.

Nevertheless, including LSTM and GRU, RNNs are limited in their capability to process tasks involving very long sequences in practice (Bradbury et al., 2016; Khandelwal et al., 2018). For mitigating this issue, the Transformer architecture (Vaswani et al., 2017) and graph convolution networks (GCNs) (Defferrard et al., 2016) have been introduced to ERC for propagating contextual information among distant utterances and yielded state-of-the-art performance (Zhong et al., 2019; Ghosal et al., 2019).

These approaches leverage contextualized utter-

, pages 332 - 341
July 1-3, 2020 ©2020 Association for Computational Linguistics

ance features to predict emotion tags, but they ignore the inherent relation between emotion tags. We observe, that the phenomenon of emotional consistency exists widely in conversations, that is, similar emotions are much more likely to appear adjacently than dissimilar emotions, as shown in Figure 1. We surmise modeling the emotional consistency is helpful to find a more reasonable distribution of emotion tags and thus further improves the performance of emotion classification.

In this work, we propose a method to address emotion classification as sequence tagging. For a given conversation, instead of predicting the distribution of emotion tags independently, we consider relations between nearby emotion tags and choose the globally best tag sequence for the entire conversation at once. Hence, we employ a CRF (Lafferty et al., 2001) to take into account the dependency between emotion tags in neighborhoods. Contextualized utterance representations fed into the CRF layer are computed by LSTM-based context encoders. By the aid of individual context encoder, our model tracks the self dependency which depicts emotional inertia of individual speakers. The inter-speaker dependency reflecting the influence of other speakers on a certain speaker is understood by the global context encoder. We use a multi-layer Transformer encoder to enhance the global context encoder so that our model can take advantage of long-range contextual information when computing contextualized utterance representations.

In summary, our contributions are as follows:

- For the first time we model ERC task as sequence tagging and use CRF to model the emotional consistency in conversation. The CRF layer exploits past and future emotion tags to jointly decode the best tag sequence for the entire conversation.

- We apply a multi-layer Transformer encoder to enhancing the LSTM-based global context encoder. The enhanced encoder is able to capture long-range sequential context which is essential for computing contextualized utterance representations.

- Extensive experiments demonstrate that modeling the emotional consistency and long-range contextual dependency promotes the performance of emotion classification. Our method advances the state of the art for ERC on three conversation datasets.

The remainder of this paper is organized as follows. Section 2 discusses related works. Section 3 describes our sequence labeling architecture. Section 4 presents the experimental setting. Section 5 reports extensive experimental results and makes a detailed analysis. We conclude this paper in Section 6.

2 Related Work

Emotion Recognition in Conversation: Early researches on emotion recognition in conversation mainly use lexicon-based methods and audio features (Lee et al., 2005; Devillers and Vidrascu, 2006). Some open-source conversation datasets with visual, acoustic and textual features have been available in the past few years (Busso et al., 2008; Poria et al., 2018). Along with these datasets, a number of deep learning methods are applied to emotion recognition. Poria et al. (2017) proposes context LSTM to capture contextual information for sentiment classification. DialogueRNN (Majumder et al., 2019) models the emotional dynamics by its party GRU and global GRU. It employs attention mechanisms to pool information from global context for each target utterance. Zhong et al. (2019) proposes Knowledge-Enriched Transformer(KET), which learns structured conversation representation by hierarchical self-attention and external commonsense knowledge. DialogueGCN (Ghosal et al., 2019) applies the graph neural network to context propagation issues present in the current RNN-based methods for ERC and achieves the state-of-the-art performance on multiple conversation datasets.

Transformer: Transformer has achieved great success in various NLP tasks due to its rich representation and high computation efficiency. Self-attention mechanisms endow Transformer with the capability of capturing longer-range dependencies than RNNs. Recent works such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018) use Transformer encoder and decoder respectively to learn representations on large-scale datasets. These representations are transferred to down-stream tasks such as named entity recognition (NER) and question answering and achieves state-of-the-art results. Dai et al. (2019) introduces the notion of recurrence to address context fragmentation limitations of Transformer. Wang et al. (2019) explores Transformer with additional LSTM layers to better capture the sequential context while retaining the high computation efficiency.

Sequence Tagging: Sequence tagging has drawn research attention for a few decades. It includes a bunch of NLP tasks such as part of speech tagging (POS), chunking and NER. The most common statistical models for sequence tagging includes hidden Markov model (HMM), maximum entropy Markov model (MEMM) and CRF (Ratinov and Roth, 2009; Passos et al., 2014). These traditional sequence tagging methods rely heavily on hand-crafted features. In the past few years, convolutional neural networks (CNNs) and RNNs are introduced to tackle sequence tagging problems and achieves competitive performance against traditional methods (Graves et al., 2013; Chiu and Nichols, 2016). Huang et al. (2015) has pointed out that the combination of bidirectional LSTM and CRF can efficiently use both past and future input features as well as past and future tags information. Hence, BiLSTM-CRF model produces state-of-the-art results on many sequence tagging tasks.

3 CESTa: Contextualized Emotion Sequence Tagging

Existing works (Majumder et al., 2019; Zhong et al., 2019; Ghosal et al., 2019) define the ERC task as the prediction of emotion tags of constituent utterances. However, emotional consistency which is an important characteristic of the conversation is not taken into consideration. CESTa differs from those methods in that it treats ERC as a task of sequence tagging of which performance is generally improved by choosing the globally best set of tags for the entire sequence at once. To this end, CESTa employs a CRF to take advantage of past and future tags to predict the current tag. For the tth utterance in a conversation, the textual feature $\mathbf{u}_t$ is extracted by a single-layer CNN and fed into the global and individual context encoders which learn inter-speaker and self dependency respectively. Moreover, the global context encoder is enhanced by a number of Transformer blocks to propagate long-range contextual information effectively. The concatenation of the global context encoding $\mathbf{g}_t$ and individual context encoding $\mathbf{s}_t$ is considered as a matrix of scores and fed into the final CRF layer. The overall architecture is shown in Figure 2.

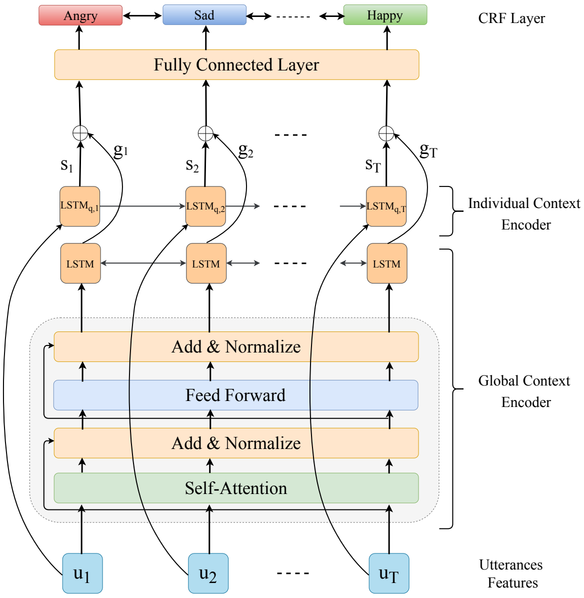

Figure 2: Overview of CESTa. The Transformer-enhanced global context encoder takes the textual feature $\mathbf{u}_t$ of the tth utterance in a conversation as input and produces encoding $\mathbf{g}_t$. Also, $\mathbf{u}_t$ is fed into the individual context encoder to update states for the corresponding speaker of which index is $q = q(\mathbf{u}_t)$ and outputs another encoding $\mathbf{s}_t$. A CRF layer is applied over the concatenation of each $\mathbf{g}_t$ and $\mathbf{s}_t$ to obtain the final prediction for each utterance in the conversation.

3.1 Utterance Feature Extraction

We employ convolutional neural networks (CNNs) to extract textual features for each utterance. Following Kim (2014), we use a simple architecture consisting of a single convolutional layer followed by max-pooling layer and a fully connected layer. Specifically, three distinct convolutional filter region sizes of 3, 4, 5 are used to obtain n-gram features. For each region size, we use 100 filters to learn complementary features. The max-pooling results of each feature map are activated by a rectified linear unit (RELU) and concatenated before fed into a fully connected layer consisting of 100 hidden units, of which the activation forms the utterance representation.

We explore two methods to train this network. It can be trained jointly with CESTa and thus its gradients will be updated during the training of the whole architecture. On the other hand, it also can be trained as an individual task of utterance classification with emotion tags. According to characteristics of different datasets, we choose pertinent strategies for the utterance feature extraction. The strategy choices are reported in Section 4.3.

3.2 Global Context Encoder

It is essential to take the contextual information into account when classifying an utterance in a sequence since other utterances in this sequence have a substantial effect on the emotion of current utterance. In other words, the emotion of current speaker can be forced to change by utterances of counterparts. This fact reflects the inter-speaker dependency which is closely related to the tendency for speakers to mirror their counterparts during the conversation (Navarretta, 2016) and is crucial to model emotional dynamics in a conversation.

Given the sequential nature of the conversation, we employ a bidirectional LSTM (BiLSTM) to capture the contextual information. However, modeling the long-range contextual information is a weakness of RNNs. Due to self-attention mechanisms, the Transformer is superior to RNN-based models in processing long-range context. Hence, we use a multi-layer Transformer encoder to enhance the context encoder. Specifically, the enhanced context encoder takes textual features of utterances as input, applies a multi-head self-attention operation (Vaswani et al., 2017) over them followed by point-wise fully connected feed-forward layers to produce contextualized vectors of utterances. Finally, contextualized utterance representations are fed into the BiLSTM layer which fuses long-range sequential contextual information to produce the context encoding:

$$\begin{aligned}
\mathbf{h}_0 &= (\mathbf{u}_1, \ldots, \mathbf{u}_T) \\
\mathbf{h}_l &= TransformerBlock(\mathbf{h}_{l-1}), l \in [1, N] \\
\mathbf{g}_t &= BiLSTM_t(\mathbf{h}_N^t), t \in [1, T]
\end{aligned} \tag{1}$$

where N is the number of Transformer layers, T is the length of conversation, $\mathbf{g}_t$ is the context encoding that is formed by the concatenation of left context vector $\overrightarrow{\mathbf{g}_t}$ and right context vector $\overleftarrow{\mathbf{g}_t}$, which is generated by a forward LSTM and a backward LSTM respectively.

3.3 Individual Context Encoder

Individual context encoder keeps track of the self dependency which reflects the emotional influence that speakers have on themselves during the conversation. Under the effect of emotional inertia, each individual speaker in a conversation tends to maintain a stable emotional state during the conversation until counterparts lead into changes (Poria et al., 2019). Since our model is only evaluated on textual modality, we hypothesize the self-dependency of each individual speaker could be deduced by its own textual utterances. This leads to an effective but simpler speaker-level context encoder than those used in other works (Majumder et al., 2019; Ghosal et al., 2019).

We implement an LSTM as the individual context encoder to output all speaker states for each time step. It exploits the current input utterance to update states only for the corresponding speaker. Specifically, for the tth utterance in a conversation, let $q = q(\mathbf{u}_t)$ denote the speaker of $\mathbf{u}_t$. The state $\mathbf{s}_{q,t}$ of an individual speaker q at timestep t in the conversation is updated by the following formula:

$$\mathbf{s}_{q,t} = LSTM_{q,t}(\mathbf{u}_t) \tag{2}$$

where $\mathbf{s}_{q,t}$ is specific to the speaker q and is updated by the current utterance $\mathbf{u}_t$ while excluding utterances from other speakers.

3.4 CRF Layer

Inspired by the emotional consistency of conversations, we consider ERC as a task of sequence tagging which is beneficial to consider the correlations of nearby tags and choose the globally best chain of tags for a given input sequence. For this reason, a CRF is employed in CESTa to yield final predictions with the aid of neighboring tags. In our scenario, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_T)$ represents an input sequence where $\mathbf{u}_t$ is the feature vector of the tth utterance, $\mathbf{y} = (y_1, \ldots, y_T)$ represents a generic sequence of tags for $\mathbf{U}$, $Y(\mathbf{U})$ represents all possible tag sequences for $\mathbf{U}$. The probability of $\mathbf{y}$ is generated by a softmax over all possible tag sequences:

$$p(\mathbf{y} \mid \mathbf{U}) = \frac{e^{s(\mathbf{U},\mathbf{y})}}{\sum_{\mathbf{y}' \in Y(\mathbf{U})} e^{s(\mathbf{U},\mathbf{y}')}} \tag{3}$$

where $s(\mathbf{U}, \mathbf{y})$ is the score for $\mathbf{y}$ which is given by the sum of two matrices: one $K \times K$ matrix of transition scores, one $T \times K$ matrix of scores comes from the concatenation of the global and individual context encoding, K is the number of distinct tags.

During training, we maximize the log-likelihood of correct tag sequences for a training set $\{(\mathbf{U}_i, \mathbf{y}_i)\}$, which is given by:

$$L = \sum_i \log(p(\mathbf{y} \mid \mathbf{U})) \tag{4}$$

Dataset	#Dialogues(Train/Val/Test)	#Utterances(Train/Val/Test)	#Classes
IEMOCAP	108/12/31	4810/1000/1523	6
DailyDialogue	11118/1000/1000	87170/8069/7740	7
MELD	1038/114/280	9989/1109/2610	7

Table 1: Statistics of training, validation and test datasets. For IEMOCAP, we use 10% of the training dialogues as the validation dataset. For DailyDialogue and MELD, we split train/val/test according to the same ratio provided by Zhong et al. (2019).

While decoding, we search for the tag sequence that obtains the maximum score, given by:

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in Y(\mathbf{U})} s\left(\mathbf{U}, \mathbf{y}\right) \tag{5}$$

Since we only model interactions of two successive tags, both the training and decoding can be solved efficiently by dynamic programming (Rabiner, 1989). In addition, it is favourable for improving results to apply a non-linear transformation to the concatenation of the global and individual context encoding before feeding it into the CRF layer (Lample et al., 2016). Accordingly, results with our method reported in Section 5 incorporate an extra hidden layer.

4 Experimental Setting

4.1 Datasets

For ease of comparison with state-of-the-art methods, we evaluate CESTa on three artificial conversation datasets: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018) and DailyDialogue (Li et al., 2017) rather than natural emotions corpus such as LEGO (Schmitt et al.; Ultes et al., 2015). IEMOCAP and MELD are both multimodal datasets with visual, acoustic and textual features, while DailyDialogue only contains textual features. For this work, we focus on emotion recognition in textual conversation. These three datasets are all split into training, validation and test datasets. The statistics are reported in Table 1.

IEMOCAP: This dataset contains five sessions, each of them was recorded from two actors. Training dataset is composed of dyadic conversations from session one to four. Annotations of utterances include six basic emotions, namely happy, sad, neutral, angry, excited and frustrated.

DailyDialogue: DailyDialogue is a human-written dyadic conversation dataset, reflecting daily communication way and covering various topics about human daily life. Emotion labels contains anger, disgust, fear, happiness, sadness, surprise and other. Since DailyDialogue does not provide speaker information, we treat utterance turns as speaker turns by default.

MELD: Multimodal Emotion Lines Dataset (MELD) is collected from TV-series *Friends* containing 1438 multi-party conversations. Each utterance is annotated with one of the seven emotion labels including happy/joy, anger, fear, disgust, sadness, surprise and neutral.

4.2 Baselines

CNN (Kim, 2014): A single-layer CNN which is identical to our utterance feature extraction network described in Section 3.1, which is the only baseline model without modeling contextual information.

CNN+cLSTM (Poria et al., 2017): Textual features of utterances are obtained by a CNN, over which a context LSTM (cLSTM) is applied to learn the contextual information.

DialogueRNN (Majumder et al., 2019): The RNN-based method that models both context and speaker information. After extracting textual features by a fine-tuned CNN, DialogueRNN applies global GRU and party GRU to the task of modeling speaker state and contextual information respectively.

DialogueGCN (Ghosal et al., 2019): Textual utterance features are extracted by a CNN in the same way as DialogueRNN does before they are fed into a bidirectional GRU to capture contextual information. After that, a graph convolutional network is applied to modeling speaker-level information. Contextual features and speaker-level features are concatenated and a similarity-based attention mechanism is used to obtain utterance representations for the final classification.

KET (Zhong et al., 2019): Enriched by the external commonsense knowledge, KET employs the

Models	IEMOCAP							DailyDialogue	MELD
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Avg.(w)	Avg.(micro)	Avg.(w)
CNN	35.34	53.66	51.61	62.17	50.66	55.56	51.28	49.27	55.86
CNN+cLSTM	33.90	69.76	48.40	57.55	62.37	57.64	56.04	51.84	56.87
DialogueRNN	37.94	78.08	58.95	**64.86**	68.11	58.85	62.26	51.64	57.07
DialogueGCN	42.75	**84.54**	63.54	64.19	63.08	**66.99**	64.18	-	58.10
KET	-	-	-	-	-	-	59.56	53.37	58.18
CESTa	**47.70**	80.82	**64.76**	63.41	**75.95**	62.65	**67.10**	**63.12**	**58.36**

Table 2: Comparisons with baselines and state-of-the-art methods. Best performances are highlighted in bold.

Transformer encoder to capture the contextual information and uses the Transformer decoder to predict the emotion tag for the target utterance.

4.3 Training Setup

All three datasets are preprocessed by lower-casing and tokenization[1]. In order to relieve the effect of out-of-vocabulary (OOV) words, we also impose a stemming procedure on these datasets.

GloVe vectors trained on Common Crawl 840B with 300 dimensions are used as fixed word embeddings. We use a 12-layers 4-heads Transformer encoder of which the inner-layer dimensionality is 2048 and the hidden size is 100. The number of hidden units of both context BiLSTM and speaker LSTM is 30. Along with a batch size of 64 and learning rate of 0.0005, the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.98$ and $\epsilon = 10^{-9}$ is used throughout the training process.

Note that due to utterances in the MELD dataset rarely contain emotion specific expressions, our model needs more expressive utterance features which can be extracted by a separate fine-tuned CNN. According to (Majumder et al., 2019; Ghosal et al., 2019), we train a CNN at utterance level with the emotion labels for MELD. As for datasets of IEMOCAP and DailyDialogue involving rich emotion representations in utterances, a CNN to extract textual features is trained jointly with the whole architecture of our model.

5 Results and Discussions

5.1 Comparison Results

We compare the performance of our model with baseline methods, as shown in Table 2. Note that statistics of conversation lengths which play an important role in ERC vary greatly between different datasets, as shown in Table 3, the performance of our model on different datasets changes accordingly, as what we analyze in the following.

Dataset	Max.	Min.	Avg.
IEMOCAP	110	8	50
DailyDialogue	35	2	8
MELD	33	1	10

Table 3: Statistics of conversation length of three datasets.

IEMOCAP: The weighted macro-F1 is used as the evaluation metric following (Majumder et al., 2019; Ghosal et al., 2019; Zhong et al., 2019). F1 scores of individual labels are also reported since the six emotion labels in IEMOCAP are unbalanced. As evidenced by Table 2, our model is around 3% better than DialogueGCN, 5% better than DialogueRNN and at least 7.5% better than all other baseline models.

To explain the gap in performances, one major reason is that some models like CNN, CNN+cLSTM and KET neglect the speaker-level information modeling so that models will treat utterances equally from different speakers, leading to certain loss in performance. Besides, considering that the average conversation length in IEMOCAP is 50 and the maximum length exceeds 100, the Transformer is capable of better capturing long-range dependency compared to RNNs-based context encoders like LSTM or GRU. Moreover, our model utilizes CRF to exploit the influence that past and future tags have on the current tag, which is not taken into account by any of existing models. We surmise that the CRF layer takes the emotional consistency into consideration when classifying similar emotions, such as ”happy” and ”excited”, hence CESTa is aware of the similarity between them and

[1] https://www.tensorflow.org/datasets/api_docs/python/tfds/features/text/Tokenizer

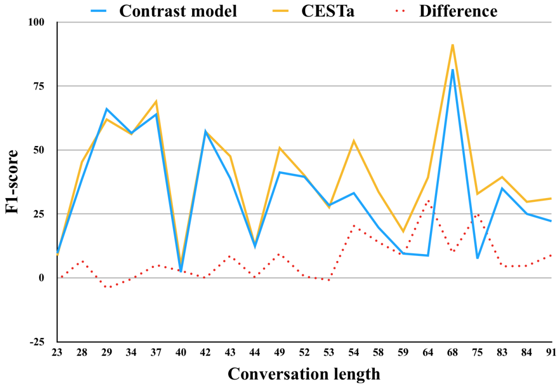

Figure 3: The performance of different models on conversations with different length. Yellow solid: our CESTa. Blue solid: the contrast model with only LSTM-based global context encoder. Red dotted: the difference between CESTa and the contrast model.

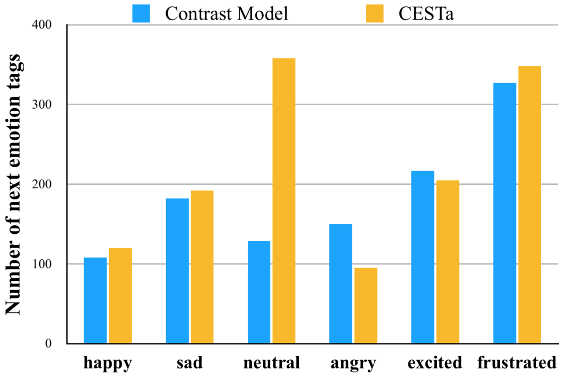

Figure 4: Statistics of pairs consisting of two identical tags which are consecutive in the conversation given by different models. Yellow: our CESTa. Blue: the contrast model without the CRF layer.

outperforms other models on these emotions.

DailyDialogue: In this dataset, the majority class(neutral) accounts for more than 80% in the test dataset. We use the micro-averaged F1 excluding the neutral class as the evaluation metric due to the imbalanced data distribution. DailyDialogue contains lots of short dyadic conversations of which average length is 8, this leads to frequent speaker turnovers. In this case, modeling speaker-level information with speaker encoder releases more ability in improving the performance. According to Li et al. (2017), DailyDialogue contains rich emotions so that our model can learn more expressive representations for utterances. Furthermore, DailyDialogue reflects human communication style, which means a definite emotional consistency can be utilized by the CRF layer in CESTa. This explains the reason of our model outperforming baselines by a large margin.

MELD: On MELD, we follow the same metric used on IEMOCAP. The performance differences between baseline models and our CESTa is not as contrasting as they are on IEMOCAP and DailyDialogue. This is mostly because of the nature of MELD. In MELD, there are many conversations containing more than 5 speakers while the average conversation length is only 10 and the minimum length is only 1. For short conversations, the advantage of the Transformer which is superior to RNNs in capturing the long-range inter-speaker dependency is not obvious. In general, majority of the speakers attending the conversation in MELD only utter a small amount of utterances. This leads the difficulty of modeling the self dependency. Additionally, utterances in MELD suffer a shortage of emotion specific expressions, this further increases the difficulty for emotion modeling. Nevertheless, CESTa achieves better results than baselines. We attribute this to the CRF layer which has an insight into the emotional consistency.

5.2 Model Analysis

Analysis of Transformer Enhancing: We evaluate the effect of Transformer enhancing on conversations with different lengths. On the test dataset of IEMOCAP, conversations are grouped by length and fed into two models: one is our CESTa with the Transformer-enhanced global context encoder, another is the contrast model that only uses LSTM-based global context encoder. The average F1 score of different groups are shown in Figure 3.

It is easy to observe that both context encoders have similar effect on relatively short conversations. However, the advantage of Transformer enhancing are more obvious as the length of conversation exceeds 54. This confirms the contribution of Transformer to the modeling of long-range contextual information.

Analysis of Emotional Consistency: We experiment on the test dataset of IEMOCAP to check the fitting of emotional consistency. We compare two models: one is our CESTa with the CRF layer, another is the contrast model that uses a softmax layer instead of CRF for classification. Statistics are given by Figure 4.

For most emotion tags, CESTa demonstrates a more obvious emotional consistency, that is, the same tag are more likely appear adjacently in given

Transformer	LSTM-based Global Context Encoder	Individual Context Encoder	CRF	IEMOCAP	DailyDialogue
No	Yes	Yes	Yes	64.25	60.28
Yes	No	Yes	Yes	64.86	59.13
Yes	Yes	No	Yes	62.35	57.10
Yes	Yes	Yes	No	65.31	60.17
Yes	Yes	Yes	Yes	67.10	63.12

Table 4: Ablation results on IEMOCAP and DailyDialogue.

conversations. We assume CESTa has learnt the emotional consistency very well and thus achieves a better performance. For the tag of "angry" and "excited", CESTa reflects less emotional consistency than the contrast model. However, we find that the quantitative distribution of next tags of "angry" and "excited" given by CESTa is closer to the ground truth than the contrast model. This trade-off between the emotional consistency and evaluation of performance is worth to further study.

Ablation Study: We conduct ablation study to investigate the necessities of the Transformer enhancing, global context encoder, individual context encoder and the CRF layer. The study is performed on IEMOCAP and DailyDialogue by removing one component at a time. Results are given in Table 4.

The results align with our analysis as the four components all improve performance by varying extents. The individual context encoder contributes most of the improvements against the baseline on both datasets. This shows the individual context encoder can capture emotional inertia for each speaker.

For IEMOCAP, the Transformer enhancing brings CESTa almost 3% increase of performance, which is the second biggest increase only after the increase 4.75% brought by the individual context encoder. For DailyDialogue, the dataset of which conversations are generally short, the Transformer enhancing leads to the minimum growth of performance. This demonstrates the importance of the Transformer enhancing for processing long conversations.

For both datasets, the performance falls by 2.24% and 3.99% respectively if we remove the LSTM-based global context encoder while keeping only the Transformer encoder. This demonstrates the importance of sequential contextual information captured by LSTM. Also, the CRF layer contributes 1.69% and 2.95% respectively to our model performance on IEMOCAP and DailyDialogue by optimizing globally with past and future emotion tags which contain information of emotional consistency.

Together these results provide important insights into what really counts in ERC. First, long-range sequential global context encoder is essential for emotion recognition in conversation. Modeling adequate contextual information enables the model to know the background of the current utterance. Besides, with the help of individual context encoder, emotion inertia can be learned by our model to seize the personality of the current speaker. Finally yet importantly, emotion tags flowing throughout a conversation to some extent have coherence naturally, which makes it meaningful to exploit the influence that past and future emotion tags have on the current tag with CRF.

6 Conclusion

We have introduced a new method, CESTa, to model ERC task as sequence tagging. Based on the contextualized utterance representations, it leverages past and future emotion tags to jointly decode the best tag sequence for the entire conversation at once. We conduct numerous experiments on three benchmark datasets. Through ablation studies, we have confirmed modeling the emotional consistency via CRF and enhancing the context encoder via the Transformer are beneficial to our model. Experimental results show that CESTa leads to a further performance improvement against strong baselines and achieves new state-of-the-art results.

Future works will focus on the representation of emotional consistency for each interlocutor in the conversation. We also plan to incorporate multimodal information into CESTa and evaluate it on more natural conversation datasets. Since CESTa needs to use emotion information of the whole dialogue, we will study its performance on the online dialogue system which has no access to the information of future emotions.

References

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2016. Quasi-recurrent neural networks. *arXiv preprint arXiv:1611.01576*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852.

Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Chul Min Lee, Shrikanth S Narayanan, et al. 2005. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303.

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

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

Costanza Navarretta. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 469–474.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *CoNLL-2014*, page 78.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *arXiv preprint arXiv:1905.02947*.

Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system.

Stefan Ultes, María Jesús Platero Sánchez, Alexander Schmitt, and Wolfgang Minker. 2015. Analysis of an extended interaction quality corpus. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 41–52. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chenguang Wang, Mu Li, and Alexander J. Smola. 2019. Language models with transformers. *CoRR*, abs/1904.09408.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.