DCR-Net: A Deep Co-Interactive Relation Network for Joint Dialog Act Recognition and Sentiment Classification

Libo Qin, Wanxiang Che,* Yangming Li, Mingheng Ni, Ting Liu
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, Harbin, China
{lbqin, car, yangmingli, mhni, tliu}@ir.hit.edu.cn

Abstract

In dialog system, dialog act recognition and sentiment classification are two correlative tasks to capture speakers' intentions, where dialog act and sentiment can indicate the explicit and the implicit intentions separately (Kim and Kim 2018). Most of the existing systems either treat them as separate tasks or just jointly model the two tasks by sharing parameters in an implicit way without explicitly modeling mutual interaction and relation. To address this problem, we propose a Deep Co-Interactive Relation Network (DCR-Net) to explicitly consider the cross-impact and model the interaction between the two tasks by introducing a co-interactive relation layer. In addition, the proposed relation layer can be stacked to gradually capture mutual knowledge with multiple steps of interaction. Especially, we thoroughly study different relation layers and their effects. Experimental results on two public datasets (Mastodon and Dailydialog) show that our model outperforms the state-of-the-art joint model by 4.3% and 3.4% in terms of F1 score on dialog act recognition task, 5.7% and 12.4% on sentiment classification respectively. Comprehensive analysis empirically verifies the effectiveness of explicitly modeling the relation between the two tasks and the multi-steps interaction mechanism. Finally, we employ the Bidirectional Encoder Representation from Transformer (BERT) in our framework, which can further boost our performance in both tasks.

Introduction

A dialog system should correctly understand speakers' utterances and respond in natural language. Dialog act recognition (DAR) and sentiment classification are two correlative tasks to realize the former. The goal of DAR is to attach semantic labels to each utterance in a dialog and identify the underlying intentions (Kim, Seon, and Seo 2011). Meanwhile, sentiment classification can detect the sentiments which are implicated in utterances and can help to capture speakers' intentions (Kim and Kim 2018).

Intuitively, the two tasks are closely related and the information of one task can be utilized in the other task. For example, as illustrated in Figure 1, when predicting `User B` sentiment label, it's more likely to be `Negative` in the case of known `Agreement` DA label, since `Agreement` means the current utterance agrees with previous `User A` utterance and hence `User B` sentiment label tends to be the same with the `User A` response sentiment `Negative`. Similarly, knowing the sentiment information also contributes to the current DA prediction. Hence, it's promising to take the cross-impact between the two tasks into account. In recent years, Cerisara et al., 2018 has explored the multi-task framework to model the correlation between sentiment classification and dialog act recognition. Unfortunately, their work does not achieve the promising performance, even underperforms some works which consider them as separate tasks. In this paper, we argue that this modeling method with no explicit interaction between the two tasks is not effective enough for transferring knowledge across the two tasks and has following weaknesses: (1) A simple multi-task learning framework just implicitly considers mutual connection between two tasks by sharing latent representations, which cannot achieve desirable results (Chen et al. 2019). (2) With the shared latent representations, it is hard to explicitly control knowledge transfer for both tasks, resulting in lack of interpretability.

Speaker	Utterance	DA Label	Sentiment Label
`User A`	they are as tired of social media as i am .	`Statement`	`Negative`
`User B`	yes ! i don't get it . everyone i talk to about facebook--everyone - - hates it ,but none of them will take action .	`Agreement`	`Negative`

Figure 1: A snippet of a dialog sample from the Mastodon Corpus and each utterance has a corresponding DA label and a sentiment label. (DA represents Dialog Act)

To address the aforementioned issues, we propose a **D**eep **C**o-Interactive **R**elation **N**etwork (**DCR-Net**) for joint dialog act recognition and sentiment classification, which can explicitly model relation and interaction between two tasks with a *co-interactive relation layer*. In practice, we first adopt a shared hierarchical encoder with utterance-level self-attention mechanism to obtain the shared representations of dialog act and sentiment among utterances. The shared representations are then fed into the *co-interactive*

*Corresponding author

relation layer to get fusion of dialog act and sentiment representations and we call the process of fusion as one step of interaction. With the *co-interactive relation layer*, we can directly control knowledge transfer for both tasks, which makes our framework more interpretable. Besides, the *relation layer* can be stacked to form a hierarchy that enables multi-step interactions between the two tasks, which can further better capture mutual knowledge. The underlying motivation is that if a model extracts mutual knowledge in one step of interaction, then by stacking multiple such steps, the model can gradually accumulate useful information and finally capture the semantic relation between the two tasks (Tao et al. 2019). Specifically, we explore several *relation layers* including: 1) *Concatenation* that concats the representation of dialog act and sentiment. 2) *Multilayer Perceptron (MLP)* that uses the *MLP* to learn the rich representation which contains both dialog act and sentiment information. 3) *Co-Attention* that uses the co-attention mechanism (Xiong, Merity, and Socher 2016) to capture mutually important information to contribute to the two tasks (sentiment to act and act to sentiment). Finally, the final integrated outputs are then fed to separate decoders for dialog act and sentiment prediction respectively.

We conduct experiments on two real-world benchmarks including Mastodon dataset (Cerisara et al. 2018) and Dailydialog dataset (Li et al. 2017). The experimental results show that our system achieves significant and consistent improvement as compared to all baseline methods and achieves the state-of-the-art performance. Finally, Bidirectional Encoder Representation from Transformer (Devlin et al., 2019, BERT), a pre-trained model, is used to further boost the performance.

To summarize, the contributions of this work are as follows:

- We propose a deep co-interactive relation network for joint dialog act recognition and sentiment classification , which can explicitly control the cross knowledge transfer for both tasks and make our framework more interpretable.
- Our *relation layer* can be stacked to form a hierarchy for multi-step interactions between the two tasks, which can gradually capture mutual relation and better transfer knowledge.
- We thoroughly study different relation layers and present extensive experiments demonstrating the benefit of our proposed framework. Experiments on two publicly available datasets show substantial improvement and our framework achieves the state-of-the-art performance.
- Finally, we analyze the effect of incorporating BERT in our framework. With BERT, our framework reaches a new state-of-the-art level.

Problem Formulation

In this section, we describe the formulation definition for dialog act recognition and sentiment classification in dialog.

- **Dialog Act Recognition** Given a dialog $C = (u_1, u_2..., u_T)$ consisting of a sequence of T utterances, dialog act recognition can be seen as a utterance-level sequence classification problem to decide the corresponding utterance dialog act label $(y_1^d, y_2^d, ..., y_T^d)$ for each utterance in dialog.
- **Sentiment Classification in Dialog** Sentiment classification in dialog can also be treated as an utterance-level sequence classification task that maps the utterance sequence $(u_1, u_2..., u_T)$ to the corresponding utterance sequence sentiment label $(y_1^s, y_2^s, ..., y_T^s)$.

Our Approach

In this section, we describe the architecture of DCR-Net; see the top part (a) of Figure. 2 for its overview. DCR-Net mainly consists of three components: a shared hierarchical encoder, a stack of *co-interactive relation layers* that repeatedly fuse dialog act and sentiment representations to explicitly model the relation and interaction between the two tasks, and two separate decoders for dialog act and sentiment prediction. In the following sections, the details of our framework are given.

Hierarchical Encoder

In our framework, dialog act recognition and sentiment classification share one hierarchical encoder that consists of a bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber 1997), which captures temporal relationships within the words, followed by a utterance-level self-attention layer to consider the dialog contextual information.

Utterance Encoder with BiLSTM Given a dialog $C = (u_1, ..., u_T)$ consists of a sequence of T utterances and $u_t = (w_t^1, ..., w_t^{K_t})$ which consists of a sequence of K_t words, we first adopt the BiLSTM to encode each utterance $u_t \in C$ to produce a series of hidden states $(\mathbf{h}_t^1, ..., \mathbf{h}_t^{K_t})$, and we define $\mathbf{h}_t^i$ as follows:

$$\mathbf{h}_t^i = \text{concat}\left(\overrightarrow{\mathbf{h}}_t^i, \overleftarrow{\mathbf{h}}_t^i\right), \tag{1}$$

where $\text{concat}(\cdot,\cdot)$ is an operation for concatenating two vectors, and $\overrightarrow{\mathbf{h}}_t^i$ and $\overleftarrow{\mathbf{h}}_t^i$ are the i-th hidden state of the forward LSTM and backward LSTM for w_t^i respectively.

Then, we regard the last hidden state $\mathbf{h}_t^{K_t}$ as the utterance u_t representation. Hence, the sequence of T utterances in C can be represented as $\mathbf{H} = (\mathbf{h}_1^{K_1}, \ldots, \mathbf{h}_T^{K_t})$.

Utterance-Level Self-Attention Self-attention is an effective method of leveraging context-aware features over variable-length sequences for natural language processing tasks (Yin et al. 2017; Tan et al. 2018). In our case, we use self-attention mechanism to capture dialog-level contextual information for each utterance. In this paper, we adopt the self-attention formulation by Vaswani et al. (2017) . We first map the matrix of input vectors $\mathbf{H} \in \mathbb{R}^{T \times d}$ (d represents the mapped dimension) to queries ($\mathbf{Q}$), keys ($\mathbf{K}$) and values ($\mathbf{V}$) matrices by different linear projections:

$$\begin{bmatrix} \mathbf{K} \\ \mathbf{Q} \\ \mathbf{V} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_k\mathbf{H} \\ \mathbf{W}_q\mathbf{H} \\ \mathbf{W}_v\mathbf{H} \end{bmatrix}. \tag{2}$$

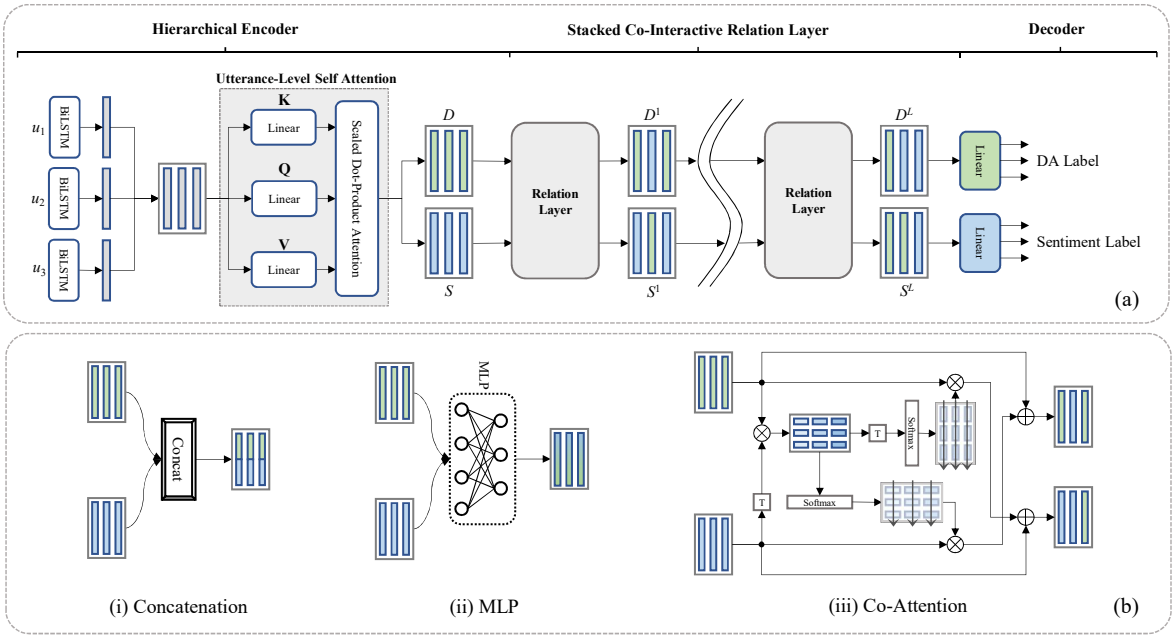

Figure 2: The top part (a) illustrates the overflow of our framework and the bottom part (b) represents different relation layers.

The attention weight is then computed by dot product between $\mathbf{Q}$, $\mathbf{K}$ and the self-attention output $\mathbf{C} \in \mathbb{R}^{T \times d}$ is a weighted sum of values $\mathbf{V}$:

$$\mathbf{C} = \text{ Attention } (\mathbf{Q},\ \mathbf{K}, \mathbf{V}\ ) = \text{Softmax}\left(\frac{\mathbf{Q}\,\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V}, \quad (3)$$

where we can see $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_T)$ as the sequence utterances representations and each utterance representation captures the whole dialogue history information. d_k represents the dimension of keys.

Now, we obtain the initial shared representations of sequence utterances dialog act $\mathbf{D} = (\mathbf{c}_1, ..., \mathbf{c}_T)$ and sentiment representations $\mathbf{S} = (\mathbf{c}_1, ..., \mathbf{c}_T)$.

Stacked Co-Interactive Relation Layer

We now describe the proposed *co-interactive relation layer*; see the bottom part (b) of Figure.2. In our paper, we use the *co-interactive relation layer* to explicitly model the relation and interaction between dialog act recognition and sentiment classification. It takes the dialog act $\mathbf{D}$ and sentiment representations $\mathbf{S}$ as inputs and then outputs their updated versions which consider cross-impact on two tasks. In particular, it can be stacked to perform multi-step interaction for better capturing mutual knowledge and relation. In our framework, we explore several types of relation layers, which can either be used individually or combined together. Formally, given the l^{th} layer inputs $\mathbf{D}^l = (\mathbf{d}_1^l, ..., \mathbf{d}_T^l) \in \mathbb{R}^{T \times d}$ and $\mathbf{S}^l = (\mathbf{s}_1^l, ..., \mathbf{s}_T^l) \in \mathbb{R}^{T \times d}$, we can adopt the following strategies to integrate the mutual knowledge between the two tasks. Before fusing information, we first apply a BiLSTM and MLP over act information and sentiment information separately to make them more task-specific, which can be written as $\mathbf{S}^{l'} = \text{MLP}\ (\mathbf{S}^l\ )$ and $\mathbf{D}^{l'} = \text{BiLSTM}\ (\mathbf{D}^l\ )$.

Concatenation Concatenation is an simple and effective method to combine two information (Wu et al. 2018). Hence, we concatenate the l^{th} layer of dialog act and sentiment representations as the updated representations.

$$\mathbf{D}^{l+1} = \text{Concat}(\mathbf{S}^{l'}, \mathbf{D}^{l'}), \quad (4)$$

$$\mathbf{S}^{l+1} = \text{Concat}(\mathbf{S}^{l'}, \mathbf{D}^{l'}). \quad (5)$$

MLP *Multilayer Perceptron (MLP)* can automatically abstract the integrated representation (Nguyen and Okatani 2018). Here, we add an MLP layer on the concatenation output to further learn the relation between two tasks and capture the mutual information, which can be formulated as follows:

$$\mathbf{D}^{l+1} = \text{MLP}(\text{Concat}(\mathbf{S}^{l'}, \mathbf{D}^{l'})), \quad (6)$$

$$\mathbf{S}^{l+1} = \text{MLP}(\text{Concat}(\mathbf{S}^{l'}, \mathbf{D}^{l'})). \quad (7)$$

Co-Attention Co-Attention is a very effective method to grasp the mutually important information on both correlated tasks (Xiong, Merity, and Socher 2016). Here, we extend the basic co-attention mechanism to utterance-level co-attention. It can produce the updated dialog act representations considering sentiment information, and the updated sentiment representations incorporating act knowledge. By doing this, we can transfer mutually relevant knowledge for the two tasks. The process can defined as follows:

$$\mathbf{D}^{l+1} = \mathbf{D}^{l'} + \text{Softmax}(\mathbf{D}^{l'}((\mathbf{S}^{l'})^{\top}))\mathbf{S}^{l'}, \quad (8)$$

$$\mathbf{S}^{l+1} = \mathbf{S}^{l'} + \text{Softmax}(\mathbf{S}^{l'}((\mathbf{D}^{l'})^{\top}))\mathbf{D}^{l'}, \quad (9)$$

where $\mathbf{D}^{l+1} = (\mathbf{d}_1^{l+1}, ..., \mathbf{d}_T^{l+1})$ and $\mathbf{S}^{l+1} = (\mathbf{s}_1^{l+1}, ..., \mathbf{s}_T^{l+1})$ are the l^{th} layer updated representations.

Decoder for Dialog Act Recognition and Sentiment Classification

After multi-step interaction with stacked co-interactive relation layer, we can get the outputs $\mathbf{D}^L = (\mathbf{d}_1^L, ..., \mathbf{d}_T^L)$ and $\mathbf{S}^L = (\mathbf{s}_1^L, ..., \mathbf{s}_T^L)$ of the last relation layer. We then adopt separate decoder to perform dialog act and sentiment prediction, which can be denoted as follows:

$$\mathbf{y}_t^d = \text{softmax}(\mathbf{W}^d \mathbf{d}_t^L + \mathbf{b}_d), \tag{10}$$

$$\mathbf{y}_t^s = \text{softmax}(\mathbf{W}^s \mathbf{s}_t^L + \mathbf{b}_s), \tag{11}$$

where $\mathbf{y}_t^d$ and $\mathbf{y}_t^s$ are the predicted distribution for dialog act and sentiment respectively; $\mathbf{W}^d$ and $\mathbf{W}^s$ are transformation matrices; $\mathbf{b}_d$ and $\mathbf{b}_s$ are bias vectors; L is the number of stacked relation layers in our framework.

Joint Training

The dialog act recognition objection is formulated as:

$$\mathcal{L}_1 = -\sum_{i=1}^{T} \hat{\mathbf{y}}_i^d \log\left(\mathbf{y}_i^d\right). \tag{12}$$

Similarly, the sentiment classification objection is defined as:

$$\mathcal{L}_2 = -\sum_{i=1}^{T} \hat{\mathbf{y}}_i^s \log\left(\mathbf{y}_i^s\right), \tag{13}$$

where $\hat{\mathbf{y}}_i^d$ and $\hat{\mathbf{y}}_i^s$ are gold utterance act label and gold sentiment label separately.

To obtain dialog act recognition and sentiment classification jointly, we follow Qin et al. (2019) to obtain the final joint objective:

$$\mathcal{L}_\theta = \mathcal{L}_1 + \mathcal{L}_2. \tag{14}$$

Experiments

Dataset

We evaluate the performance of our model on two publicly available dialogue datasets, Mastodon (Cerisara et al. 2018) and Dailydialog (Li et al. 2017).

Mastodon The Mastodon dataset[1] consists of 269 dialogues for a total of 1075 utterances in training dataset and the test dataset is a corpus of 266 dialogues for a total of 1142 utterances. The vocabulary size is 5330. We follow the same partition as Cerisara et al. (2018).

DailyDialog For Dailydialog dataset,[2] we adopt the standard split from the original dataset (Li et al. 2017), employing 11,118 dialogues for training, 1,000 for validating, and 1,000 for testing.

Experimental Settings

In our experiment setting, dimensionality of the embedding and all hidden units is selected from $\{100, 128, 256, 512, 600, 700, 800, 1024\}$. We do not use any pre-trained embedding and all word embeddings are trained from scratch. L2 regularization used on our model is 1×10^{-8} and dropout ratio adopted is selected from $\{0.1, 0.2, 0.25, 0.3, 0.4, 0.5\}$. In addition, we add a residual connection in self-attention and relation layer for reducing overfitting. We use Adam (Kingma and Ba 2014) to optimize the parameters in our model and adopt the suggested hyper-parameters for optimization. We set the stacked number of relation layer as 3. For all experiments, we pick the model which works best on dev set, and then evaluate it on test set.

Baselines

We first make a comparison with the state-of-the-art dialog act recognition models: HEC, CRF-ASN and CASA, and then we compare our model with some state-of-the-art sentiment classification models: VDCNN, Region.emb, DRNN and DialogueRNN. Finally, we compare our framework with the existing state-of-the-art joint models including: JointDAS and IIIM. We briefly describe these baseline models below: 1) **HEC** (Kumar et al. 2018): This work uses a hierarchical Bi-LSTM-CRF (Bi-directional Long Short Term Memory with CRF) model for dialog act recognition, which can capture both kinds of dependencies including word-level and utterance-level. 2) **CRF-ASN** (Chen et al. 2018): This model proposes a crf-attentive structured network for dialog act recognition, which can dynamically separate the utterances into cliques. 3) **CASA** (Raheja and Tetreault 2019): This work leverages a context-aware self-attention mechanism coupled with a hierarchical deep neural network and achieves state-of-the-art performance. 4) **VDCNN** (Conneau et al. 2017): This work proposes a deep CNN with 29 convolutional layers for text classification. 5) **Region.emb** (Qiao et al. 2018): This work proposes a new method of region embedding for text classification, which can effectively learn and utilize task-specific distributed representations of n-grams. 6) **DRNN** (Wang 2018): This work proposes a disconnected recurrent neural network for text classification which can incorporate position-invariance into RNN. 7) **DialogueRNN** (Majumder et al. 2019): DialogueRNN is a RNN-based neural architecture for emotion detection in a conversation, which can keep track of the individual party states throughout the conversation and uses this information. 8) **JointDAS** (Cerisara et al. 2018): This model uses a multitask modeling framework for joint dialog act recognition and sentiment classification, which models relation and interaction between two tasks by sharing parameters. 9) **IIIM** (Kim and Kim 2018): This work proposes an integrated neural network model which simultaneously identifies speech acts, predicators, and sentiments of dialogue utterances.

For *HEC, CRF-ASN, CASA* and *IIM* we re-implemented the models. For *VDCNN, Region.emb*, *DRNN* and *DialogueRNN*, we adopted the open-sourced code[3] to get the results. For *JointDAS*, we adopted the reported results from Cerisara et al. (2018) and run their open-source code on Dailydialog dataset to obtain results. For *IIIM*, we re-implemented the model and obtained results on the same

[1] https://github.com/cerisara/DialogSentimentMastodon

[2] http://yanran.li/dailydialog

[3] https://github.com/Tencent/NeuralNLP-NeuralClassifier and https://github.com/senticnet/conv-emotion

Model	Mastodon						Dailydialog					
	SC			DAR			SC			DAR		
	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)
HEC (Kumar et al. 2018)	-	-	-	56.1	55.7	56.5	-	-	-	77.8	76.5	77.8
CRF-ASN (Chen et al. 2018)	-	-	-	55.1	53.9	56.5	-	-	-	76.0	75.6	78.2
CASA (Raheja and Tetreault 2019)	-	-	-	56.4	57.1	55.7	-	-	-	78.0	76.5	77.9
VDCNN (Conneau et al. 2017)	39.6	31.6	44.0	-	-	-	39.7	35.6	55.2	-	-	-
Region.emb (Qiao et al. 2018)	40.3	33.6	42.8	-	-	-	41.0	36.6	56.4	-	-	-
DRNN (Wang 2018)	37.9	34.3	39.7	-	-	-	41.1	37.0	56.4	-	-	-
DialogueRNN (Majumder et al. 2019)	41.5	42.8	40.5	-	-	-	40.3	37.7	44.5	-	-	-
JointDAS (Cerisara et al. 2018)	37.6	41.6	36.1	53.2	51.9	55.6	31.2	28.8	35.4	75.1	74.5	76.2
IIIM (Kim and Kim 2018)	39.4	40.1	38.7	54.3	52.2	56.3	33.0	28.5	38.9	75.7	74.9	76.5
DCR-Net + Concat	42.1	41.3	42.9	57.1	56.9	57.2	41.2	37.4	57.4	78.2	77.6	78.7
DCR-Net + MLP	42.3	43.7	45.4	57.2	56.7	57.7	42.7	37.5	**58.8**	79.1	78.5	79.2
DCR-Net + Co-Attention	***45.1**	***47.3**	***43.2**	***58.6**	***56.9**	***60.3**	***45.4**	***40.1**	56.0	***79.1**	***79.0**	***79.1**
DCR-Net + Co-Attention + BERT	55.1	56.5	56.5	67.1	65.2	69.2	48.9	46.9	63.2	80.0	79.9	80.2

Table 1: Comparison of our model with baselines on Mastodon and Dailydialog test datasets. SC represents Sentiment Classification and DAR represents Dialog Act Recognition. The numbers with * indicate that the improvement of our model over all baselines is statistically significant with $p < 0.01$ under t-test.

datasets.[4] For all BERT-based experiments, we just replace our utterance encoder LSTM with BERT base model.[5]

Overall Results

On Dailydialog dataset, following Kim and Kim (2018), we adopt macro-average Precision, Recall and F1 for both sentiment classification and dialog act recognition. On Mastodon dataset, following Cerisara et al. (2018), we ignore the neural sentiment label and adopt the average of the dialog-act specific F1 scores weighted by the prevalence of each dialog act for dar. The experimental result is shown in Table 1. From the result, we can observe that:

1. We obtain large improvements compared with prior joint models. In Mastodon dataset, compared with *IIIM* model, our framework with Co-Attention achieves 5.7% improvement on F1 score on sentiment classification task and 4.3% improvement on F1 score on dialog act recognition task. In Dailydialog dataset, we achieve 12.4% improvement on F1 score on sentiment classification task and 3.4% improvement on F1 score on dialog act recognition task. It is worth noting that the prior joint models have modeled the relation between two tasks implicitly by sharing parameters. This result demonstrates the effectiveness of explicitly modeling the interaction between the two tasks and both tasks can boost performance from this mechanism.
2. Our framework with Co-Attention outperforms the state-of-the-art dialog act recognition models and sentiment classification models in all metrics in two datasets. It illustrates the advantages and effectiveness of our proposed joint model where the information of one task can be effectively utilized in the other task.
3. The *MLP* relation layer outperforms the *concatenation*, which shows that the *MLP* can further learn the deep implicit relation between two tasks and improve the performance. Especially, we can see that the *Co-Attention* relation layer gains the best performance among three relation layers on F1 scores on all datasets. We attribute this to the fact that the *Co-Attention* operation can automatically detect the mutually important information to each other and better interact with the two tasks.
4. From the last block of Table 1, the BERT-based model performs remarkably well on both two datasets and achieves a new state-of-the-art performance, which indicates the effectiveness of a strong pre-trained model in two tasks. We attribute this to the fact that pre-trained models can provide rich semantic features, which can help to improve the performance.

Unless otherwise stated, we only apply the Co-Attention relation layer in the following experiments.

Analysis

Although achieving good performance, we would like to know the reason for the improvement. In this section, we study our model from several directions. We first conduct several ablation experiments to analyze the effect of different components in our framework. Next, we give a quantitative analysis to study how our proposed framework improves performance. Finally, we provide a co-attention visualization to better understand how relation layer affects and contributes to the performance.

Ablation In this section, we perform several ablation experiments in our framework on two datasets and the results are shown in Table 2 . The results demonstrate the effectiveness of different components of our framework to the final performance. We give a detailed analysis in the following:

- **w/o relation layer:** In this settings, we conduct experiments on the multi-task framework where dialog act recognition and sentiment classification promote each other only by sharing parameters of the encoder, which similar to Cerisara et al. (2018). From the result, we can see that 4.8% drop in terms of F1 scores in sentiment classification while 2.6% drops in dialog act recognition in

[4] All experiments are conducted on the public datasets provided by Cerisara et al. (2018) and the dataset does not annotate the predictors. For direct comparison, we re-implemented the models excepting predicting the predicators and obtained the results on the same dataset.

[5] The BERT model is fine-tuned with our framework.

Model	Mastodon		Dailydialog	
	SC (F1)	DAR (F1)	SC (F1)	DAR (F1)
Full Model	45.1	58.6	45.4	79.1
w/o relation layer	40.3	55.2	38.0	78.4
w/o stacked relation layer	42.5	57.4	42.1	78.5
w/o self-attention	43.2	57.3	42.1	77.2
+CNN	43.9	58.2	43.1	78.4

Table 2: Ablation study on Mastodon and Dailydialog test datasets.

Mastodon dataset. In Dailydialog dataset, we can also observe the same trends that the F1 score drops a lot. This demonstrates that explicitly modeling the strong relations between two tasks with relation layer can benefit them effectively.

- **w/o stacked relation layer:** Here, we set the number of the stacked relation layer as 1 in our framework. From the result, we can see that performance drops in all metrics. It indicates that stacked structure with multiple steps of interaction does better model the semantic relation.
- **w/o self-attention:** In this setting, we remove our self-attention layer and there is no hierarchical architecture to capture dialog-level context information. The results show a significant drop in performance, indicating that capturing the dialog-level context information by the hierarchical encoder is effective and important for dialog act recognition and sentiment classification. In addition, we replace our self-attention with CNN (Kim 2014) which can also model the dialog context information. The result is shown in the last row of Table 2. We can see that CNN outperforms *w/o self-attention* version and underperforms our full model, which further demonstrates the effectiveness of the dialog context information and self-attention mechanism.

Quantitative Analysis In our DCR-Net model, we adopt the relation layer to model the interaction and relation between two tasks explicitly. To better understand our model, we compare the DA and sentiment performance between DCR-Net model and baseline without relation layer, as shown in Figure 3 and Figure 4.

We choose several DA types with a large performance boost which are shown in Figure 3. From the results, we can see that our model yields significant improvements on the act type `Exclamation`, `Thanking`, `Agreement`, `Explicit Performative`. We attribute the improvements to the fact that those acts are strong correlative with sentiment and our model can provide sentiment information explicitly for DAR rather than in an implicit method by sharing parameters. Take the fourth utterance in Figure 5 for example, providing the current utterance `Negative` sentiment information explicitly and previous utterance sentiment `Negative` label can contribute to DA `Agreement` prediction, which demonstrates the effectiveness of our proposed framework. In addition, from Figure 4, we can observe that our model outperforms baseline in both positive and negative sentiment label. We think that our relation layer can explicitly capture DA information which benefits sentiment classification task.

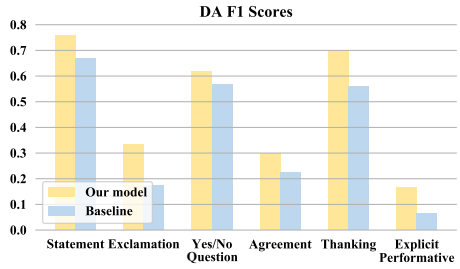

Figure 3: Quantitative analysis on different types of DA between our model with baseline.

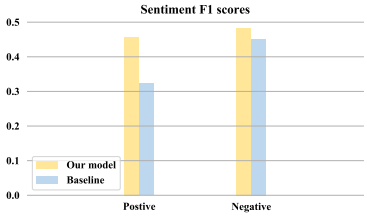

Figure 4: Quantitative analysis on different types of sentiments between our model with baseline.

Visualization of Co-Attention

In this section, with the attempt to better understand what the model has learnt, we visualized the co-attention distribution among utterances in dialogues. In particular, we visualized the attention distribution of the fifth utterance corresponding to other utterances with the number of stacked relation layers varying from 1 to 6.

From Figure 5, we can observe: (1) the act-to-sentiment attention distribution score in the fourth utterance is larger than other utterances. This is due to that the fifth utterance is more related to the fourth utterance and the `Agreement` DA represents that current utterance agrees with the fourth utterance statement. Similarly, we can see that sentiment-to-act attention in the fourth utterance distribution score is also the largest compared to other utterances. Those results demonstrate that our framework can correctly capture mutually important knowledge. (2) Using deeper layers could generally lead to better performance, especially when the number of stacked layers is less than four. It is because the stacked relation layer can better model the relation between two tasks and learn mutual knowledge. When the number of stacked layers exceeds three, the experimental performance goes worse. We suggest that the reason might lie in the gradient vanishing or overfitting problem as the whole network goes deeper.

Related Work

In this section, we will introduce the related work about dialog act recognition, sentiment classification and the joint model for the two tasks.

Speaker	Utterance	Dialog Act	Sentiment
1. User A	deleted my facebook account a few days ago and i never felt so free in my entire life .	Statement	Positive
2. User B	now i just have to encourage my closest friends to do the same .	Offer	Positive
3. User A	it shouldn't be that hard .	Statement	Positive
4. User B	they are as tired of social media as i am .	Statement	Negative
5. User A	yes ! i don't get it . everyone i talk to about facebook--everyone - - hates it , but none of them will take action .	Agreement	Negative
6. User B	they are as tired of social media as i am .	Statement	Negative

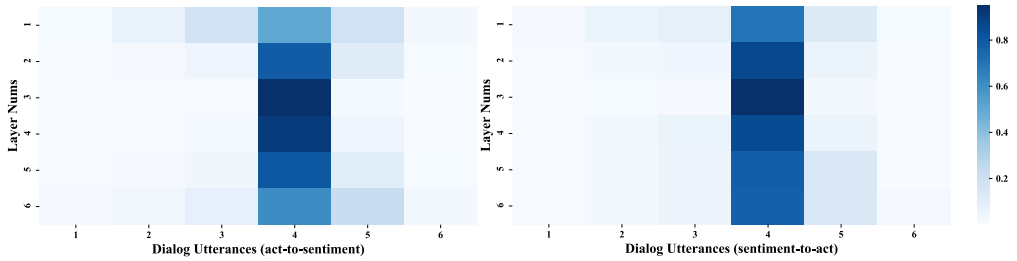

Figure 5: Co-Attention distribution score from the fifth utterance to the whole dialog utterances. The top is corresponding dialog context. The bottom left is the act-to-sentiment attention and the right is the sentiment-to-act attention.

Dialog Act Recognition

Recently, more and more neural networks have been propose to solve the DAR. Kalchbrenner and Blunsom (2013) propose the hierarchical CNN to model the utterance sequence for DA classification. Lee and Dernoncourt (2016) propose a model based on CNNs and RNNs which incorporated the previous utterance as context to classify the current DA and show the promising performance. Ji, Haffari, and Eisenstein (2016) propose the latent variable recurrent neural network for jointly modeling sequences of words and discourse relations between adjacent sentences. Furthermore, many work (Liu et al. 2017; Kumar et al. 2018; Chen et al. 2018) explore different architectures to incorporate the context information for DAR. Raheja and Tetreault (2019) propose the token-level self-attention mechanism for DAR and achieved state-of-the-art performance.

Sentiment Classification

Sentiment classification in dialog system can be seen as the sentence-level sequence classification problem. One series of works are based on CNN (Zhang, Zhao, and LeCun 2015; Conneau et al. 2017; Johnson and Zhang 2017) to capture the local correlation and position-invariance. Another series of works adopt RNN based models (Tang, Qin, and Liu 2015; Yang et al. 2016; Xu et al. 2016) to leverage temporal features and contextual information to perform sentence classification. Besides, Some works (Xiao and Cho 2016; Shi et al. 2016; Wang 2018) attempt to combine the advantages of CNN and RNN for sentence classification.

Joint Model

Considering the correlation between dialog act recognition and sentiment classification, joint models are proposed to solve two tasks simultaneously in a unified framework. Cerisara et al. (2018) explore the multi-task framework to model the correlation between the two tasks. Compared with their model, we propose a relation layer to explicitly model the correlation between dialog act recognition and sentiment classification while they model in an implicit way simply by sharing parameters. Specifically, our relation layer can be stacked to capture mutual knowledge sufficiently. Kim and Kim (2018) propose an integrated neural network model for identifying dialog act, predicators, and sentiments of dialogue utterances. Their framework classifies the current dialog act only considering the last time dialog act results, which can not make full use of context information, while we adopt the hierarchical encoder with utterance-level self-attention to leverage context information. In addition, their model does not model the sentiment information for dialog act while our framework considers interaction and mutual relation between two tasks.

Conclusion

This paper focuses on explicitly establishing the bi-directional interrelated connections for dialog act recognition and sentiment information. We propose a deep relation network to jointly model the interaction and relation between the two tasks, which adopts a stacked co-interactive relation layer to incorporate mutual knowledge explicitly. In addition, we explore three different relation layers and make a thorough study on their effects on the two tasks. Experiments on two datasets show the effectiveness of the proposed models and achieve the state-of-the-art performance. Extensive analysis further confirms the correlation between two tasks and reveals that modeling the relation explicitly can boost their performance. Besides, we analyze the effect of incorporating strong pre-trained BERT model in our joint model. With BERT, the result reaches a new state-of-the-art level.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Natural Science Foundation of China (NSFC) via grant 61976072, 61632011 and 61772153.

References

Cerisara, C.; Jafaritazehjani, S.; Oluokun, A.; and Le, H. T. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proc. of COLING*.

Chen, Z.; Yang, R.; Zhao, Z.; Cai, D.; and He, X. 2018. Dialogue act recognition via crf-attentive structured network. In *Proc. of SIGIR*.

Chen, Z.; Wang, X.; Xie, X.; Wu, T.; Bu, G.; Wang, Y.; and Chen, E. 2019. Co-attentive multi-task learning for explainable recommendation. In *Proc. of IJCAI*.

Conneau, A.; Schwenk, H.; Barrault, L.; and Lecun, Y. 2017. Very deep convolutional networks for text classification. In *Proc. of ACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*.

Ji, Y.; Haffari, G.; and Eisenstein, J. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913*.

Johnson, R., and Zhang, T. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proc. of ACL*.

Kalchbrenner, N., and Blunsom, P. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

Kim, M., and Kim, H. 2018. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. *Pattern Recognition Letters*.

Kim, H.-S.; Seon, C.-N.; and Seo, J.-Y. 2011. Review of korean speech act classification: machine learning methods. *Journal of Computing Science and Engineering*.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*.

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

Kumar, H.; Agarwal, A.; Dasgupta, R.; and Joshi, S. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proc. of AAAI*.

Lee, J. Y., and Dernoncourt, F. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proc. of IJCNLP*.

Liu, Y.; Han, K.; Tan, Z.; and Lei, Y. 2017. Using context information for dialog act classification in dnn framework. In *Proc. of EMNLP*.

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

Nguyen, D.-K., and Okatani, T. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proc. of CVPR*.

Qiao, C.; Huang, B.; Niu, G.; Li, D.; Dong, D.; He, W.; Yu, D.; and Wu, H. 2018. A new method of region embedding for text classification. In *Proc. of ICLR*.

Qin, L.; Che, W.; Li, Y.; Wen, H.; and Liu, T. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proc. of EMNLP*.

Raheja, V., and Tetreault, J. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proc. of NAACL*.

Shi, Y.; Yao, K.; Tian, L.; and Jiang, D. 2016. Deep LSTM based feature mapping for query classification. In *Proc. of NAACL*.

Tan, Z.; Wang, M.; Xie, J.; Chen, Y.; and Shi, X. 2018. Deep semantic role labeling with self-attention. In *Proc. of AAAI*.

Tang, D.; Qin, B.; and Liu, T. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proc. of ACL*.

Tao, C.; Wu, W.; Xu, C.; Hu, W.; Zhao, D.; and Yan, R. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proc. of ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. of NIPS*. Curran Associates, Inc.

Wang, B. 2018. Disconnected recurrent neural networks for text categorization. In *Proc. of ACL*.

Wu, Z.; Dai, X.-Y.; Yin, C.; Huang, S.; and Chen, J. 2018. Improving review representations with user attention and product attention for sentiment classification. In *Proc. of AAAI*.

Xiao, Y., and Cho, K. 2016. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *Proc. of ICML*.

Xu, J.; Chen, D.; Qiu, X.; and Huang, X. 2016. Cached long short-term memory neural networks for document-level sentiment classification. In *Proc. of EMNLP*.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proc. of NAACL*.

Yin, Q.; Zhang, Y.; Zhang, W.; and Liu, T. 2017. Chinese zero pronoun resolution with deep memory network. In *Proc. of EMNLP*, 1309–1318.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Proc. of NIPS*, 649–657.