information

MDPI

Article

Multiple Information-Aware Recurrent Reasoning Network for Joint Dialogue Act Recognition and Sentiment Classification

Shi Li and Xiaoting Chen *

College of Computer and Control Engineering, Northeast Forestry University, Harbin 150006, China
* Correspondence: chenxiaoting@nefu.edu.cn

Abstract: The task of joint dialogue act recognition (DAR) and sentiment classification (DSC) aims to predict both the act and sentiment labels of each utterance in a dialogue. Existing methods mainly focus on local or global semantic features of the dialogue from a single perspective, disregarding the impact of the other part. Therefore, we propose a multiple information-aware recurrent reasoning network (MIRER). Firstly, the sequence information is smoothly sent to multiple local information layers for fine-grained feature extraction through a BiLSTM-connected hybrid CNN group method. Secondly, to obtain global semantic features that are speaker-, context-, and temporal-sensitive, we design a speaker-aware temporal reasoning heterogeneous graph to characterize interactions between utterances spoken by different speakers, incorporating different types of nodes and meta-relations with node-edge-type-dependent parameters. We also design a dual-task temporal reasoning heterogeneous graph to realize the semantic-level and prediction-level self-interaction and interaction, and we constantly revise and improve the label in the process of dual-task recurrent reasoning. MIRER fully integrates context-level features, fine-grained features, and global semantic features, including speaker, context, and temporal sensitivity, to better simulate conversation scenarios. We validated the method on two public dialogue datasets, Mastodon and DailyDialog, and the experimental results show that MIRER outperforms various existing baseline models.

Keywords: multi-task learning; sentiment classification; dialogue act recognition; heterogeneous graph network

check for updates

Citation: Li, S.; Chen, X. Multiple Information-Aware Recurrent Reasoning Network for Joint Dialogue Act Recognition and Sentiment Classification. *Information* **2023**, *14*, 593. https://doi.org/10.3390/info14110593

Academic Editor: Arkaitz Zubiaga

Received: 15 September 2023
Revised: 17 October 2023
Accepted: 26 October 2023
Published: 1 November 2023

1. Introduction

Dialogue act recognition and dialogue sentiment classification are two related tasks that are essential for effectively understanding speakers' utterances in dialogue systems [1,2]. Researchers have observed that these tasks are closely related and mutually supportive when executed jointly. On the one hand, DAR provides useful clues for sentiment classification. In return, sentiment transformation can also be beneficial for predicting DAR.

DAR's goal is to predict the act label of each utterance in the dialogue, indicating the speaker's explicit intention [3]. Dialogue act labels of the utterances include inform, questions, directives, commissive, and so on; DSC aims to predict the sentiment label, indicating the speaker's implicit intention [4,5]. Dialogue sentiment labels can be divided into neutral, negative, and positive and can also be further subdivided into happiness, surprise, sadness, anger, disgust, fear, and so on. Take the conversation in the Mastodon dataset in Figure 1 as an example. To predict the sentiment of speaker u_b, in addition to its semantic information, its act label disagreement and the sentiment label negative of the previous utterance u_a can also provide helpful references. This is because the u_b act label of disagreement is opposite to the opinion of u_a, and thus the u_b sentiment label tends to be positive. Similarly, the opposite sentiment labels between u_a and u_b also help to infer the u_b act label of disagreement. There are three key factors in the dual-task reasoning process of this paper: (1) the semantics of utterances u_a and u_b, (2) the temporal relation between u_a and u_b, and (3) the labels of u_a and u_b used for another task.

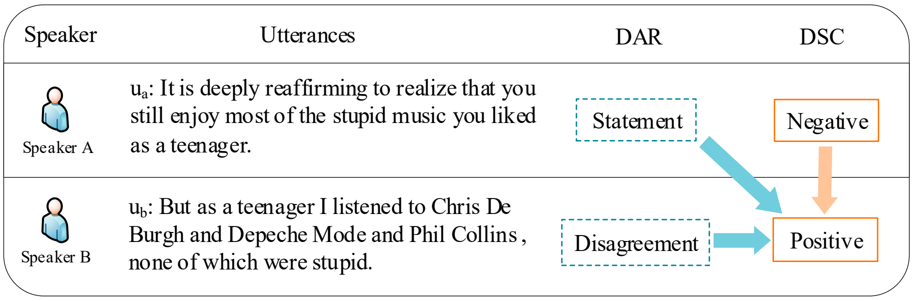
Speaker
Utterances
DAR
DSC
Speaker A
u_a: It is deeply reaffirming to realize that you still enjoy most of the stupid music you liked as a teenager.
Statement
Negative
Speaker B
u_b: But as a teenager I listened to Chris De Burgh and Depeche Mode and Phil Collins , none of which were stupid.
Disagreement
Positive

Figure 1. Sample conversations in the Mastodon dataset.

Although sequence-to-sequence (Seq2Seq) [6] models are effective at modeling sequential dialogues, they are limited when modeling graph-structured ones. To overcome this limitation, Qin et al. [7] proposed a graph-structured network Co-GAT to encode utterances based on the graph topology instead of their sequential appearances. The graph established in Co-GAT was homogeneous, where types of nodes and edges were not distinguished. Xing et al. [8] proposed using a heterogeneous graph relational graph convolutional network to jointly model DAR and DSC considering the interactions between multiple relationships (different types of edges) among dialogue utterances, where nodes represent only utterances. Most of the current mainstream models only consider global semantic information in the dialogue understanding module and fail to fully incorporate fine-grained-level context information. Intuitively, for the current utterance, its nearer dialogue neighbors are always more influential and informative than the more remote ones due to the closer replying relationships between them.

To this end, we propose a multiple information-aware recurrent reasoning network for joint DAR and DSC. Firstly, a hybrid CNN group with a BiLSTM connection is designed to extract features from sequence information by smoothly transmitting it to multiple levels of local information layers. Secondly, to extract rich global semantic features, we design a speaker-aware temporal dependencies heterogeneous graph transformer (SATD-HGT) to model the intra- and inter-speaker semantic information interaction by introducing temporal relations. Finally, in the recurrent reasoning mechanism, a dual-task reasoning temporal dependencies heterogeneous graph transformer (DRTD-HGT) is designed to model the utterance information containing time relationships within and between tasks. Using the framework shown in Figure 2, multiple semantic-level and prediction-level interactions are integrated to obtain complete deep semantic information. The goal of this paper is to predict the labels of two tasks. Due to the lack of real labels for prediction-level interactions, the recurrent dual-task reasoning mechanism is used to generate new predictions based on the estimated label distribution in the previous step as the current prediction clue. Thus, the label distribution of the two tasks gradually improves with the increasing number of steps. The contributions of this study are as follows: (1) Construction of heterogeneous graphs related to speakers and the dual task: we design different types of nodes and meta-relations with node-edge-type-dependent parameters to represent heterogeneous interactions within the dialogue graph structure. (2) Fusion of global and local information: we propose the SATD-HGT framework and the hybrid CNN group to obtain global and local semantic information, respectively. The experimental results show that the MIRER model integrates deep semantic information to help accurately understand the dialogue content and to better distinguish different dialogue acts or sentiment categories.

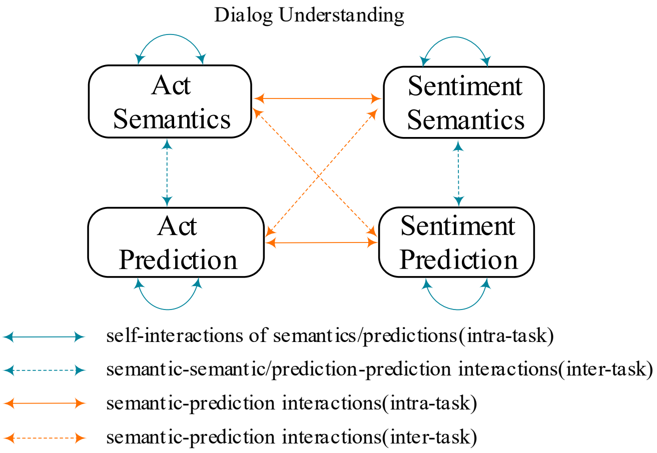
Dialog Understanding
Act Semantics
Sentiment Semantics
Act Prediction
Sentiment Prediction
self-interactions of semantics/predictions(intra-task)
semantic-semantic/prediction-prediction interactions(inter-task)
semantic-prediction interactions(intra-task)
semantic-prediction interactions(inter-task)

Figure 2. Framework for the interaction between DSC and DAR.

2. Related Work

Dialogue act recognition and sentiment classification are critical to building intelligent interactive dialogue systems and generating appropriate responses. DAR aims to detect the act category of each utterance during the dialogue process [9,10]. Existing research can be divided into three categories according to the perspective of information use: using only utterance information, independently using utterance and label information, and integrating label information into utterances [11–13]. DSC aims to detect the sentiment category of each utterance during the dialogue process [14]. Unlike traditional sentiment analysis, situation-level and speaker-level contexts play an essential role in identifying the sentiment content of utterances. Existing research often uses deep learning methods to obtain contextual features, which can be divided into sequence-based, graph-based, and knowledge-based methods [15–17].

Recent research further indicates that jointly modeling DAR and DSC can explore hidden cross-task interaction information in dialogue and better grasp the speaker’s intention. Kim et al. [4] proposed the IIIM model, which is equivalent to modeling two related tasks separately and cannot fully utilize dialogue text information. Qin et al. [7,18] proposed the DCR-Net and the Co-GAT models to fully consider cross-task information and dialogue context information. Lin et al. [19] proposed a cross-task collaborative graph attention network to encode cross-task information and contextual information. Li et al. [20] proposed using dynamic convolutional networks as the utterance encoder to capture dialogue context in multi-task learning. Xing et al. [8] proposed the DARER model to learn utterance representations of different speakers and tasks and estimate the label distributions. Recently, Xu et al. [21] proposed an ensemble model with two-stage contextual learning for joint DAR and SC, which introduces an EAGAT network to improve the performance of the classifier by using the confidence vector to selectively leverage the contextual information.

In this paper, we propose multiple information-aware recurrent reasoning networks in which the hybrid CNN group and SATD-HGT effectively extract fine-grained features and global semantic information. Furthermore, the designed RTD-HGT captures semantic-level and prediction-level interactions between the different tasks in the dialogue.

3. Methodology

The MIRER model is shown in Figure 3 and mainly consists of four modules: dialogue understanding, initial estimation, recurrent dual-task reasoning, and joint training. In the following parts of this section, we first describe the construction method of the heterogeneous graph in the dialogue and then introduce detailed information about each module in the MIRER framework.

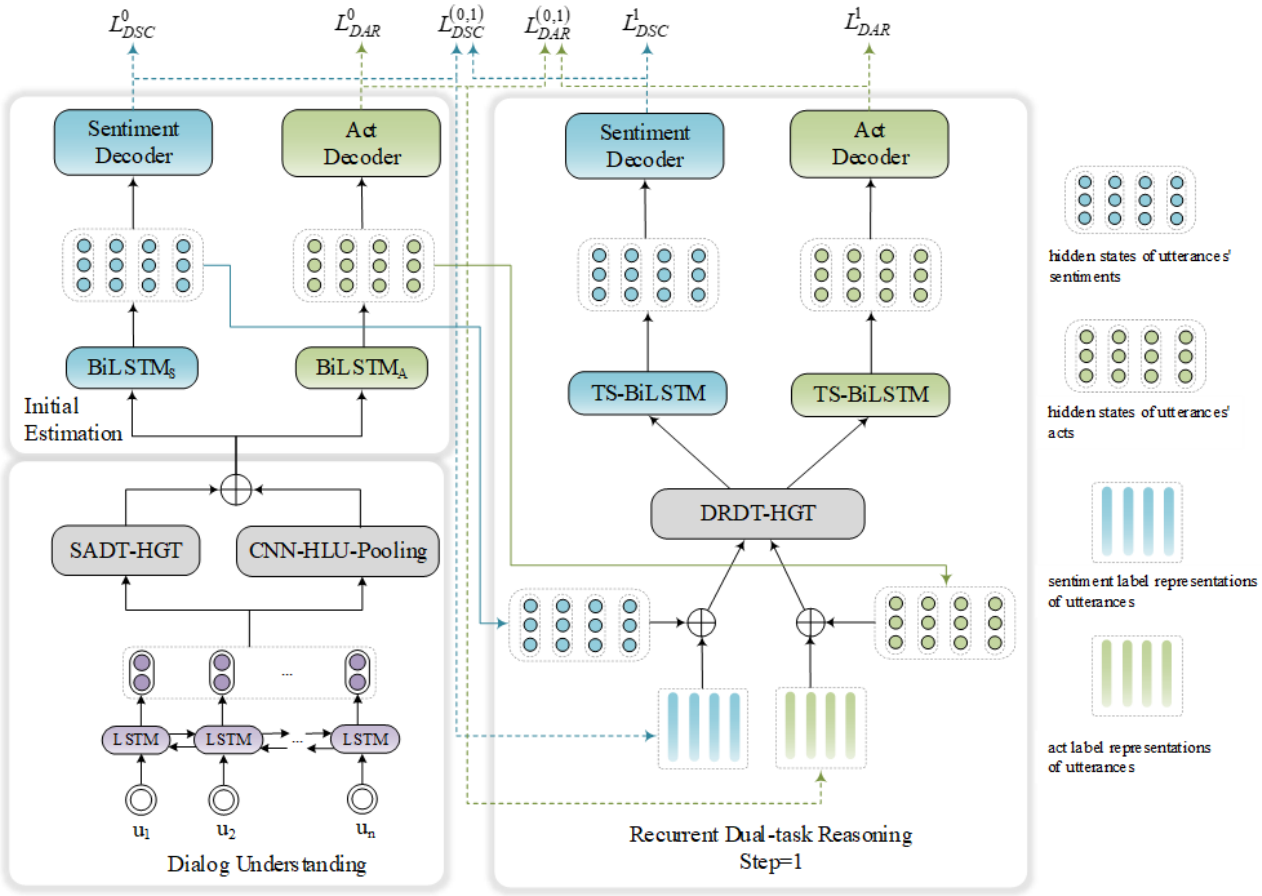

Figure 3. The network architecture of the proposed MIRER model.

3.1. Graph Construction

Inspired by [22–24], this paper models the information propagation network in dialogue as a heterogeneous graph. It designs a speaker-aware temporal graph and dual-task reasoning temporal graph to aggregate information from source nodes and obtains contextual representations of target nodes. Its meta-relations are represented as $\langle \tau(u_s), \varnothing(r), \tau(u_t) \rangle$ for the edge $r = (u_s, u_t)$ connecting neighbor utterance u_s to the current utterance node u_t.

3.1.1. Speaker-Aware Temporal Graph

We design a speaker-aware temporal graph to model the intra- and inter-speaker semantic interactions. Formally, this heterogeneous graph can be represented as $G = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$, where the node set $\mathcal{V}$ consists of utterances spoken by speakers in the conversation, represented as $V = (u_1, \ldots, u_N)$, and $\mathcal{E}$ is the set of edges, where each edge represents the relation between nodes, i.e., the information aggregation from u_i to u_j under the relation $r_{ij} \in \mathcal{R}$. Two types of speaker nodes model complex interactions between the same and different speakers, and eight meta-relations with node-edge-type-dependent parameters represent heterogeneous interactions within the conversation graph structure: temporal relations between both the intra- and inter-speaker. Table 1 provides definitions for all relation types in $\mathcal{R}$, and $I_s(i)$ and $I_s(j)$ indicate that source utterance node i and target utterance node j are from Speaker A and Speaker B, respectively, and $pos(i, j)$ represents the relative position of u_i and u_j. Specifically, the ">" symbol means that the source utterance node is spoken before the target utterance node. The "<" symbol means that the source utterance node is spoken after the target utterance node. The "=" symbol means that the source utterance node and the target utterance node occur at the same time, meaning that either the source node and the target node are the same utterance or the source node and the target node belong to different utterances but are spoken at the same time. As utterances in a conversation are arranged in chronological order, the relative position of two utterance nodes is their temporal relation. As shown in Figure 4, we use utterance nodes spoken by different speakers as target nodes to aggregate its neighbor information. The neighbors of the current utterance based on the relation triplet are the entire set of utterances belonging to all speakers in the conversation, including the utterance itself.

Table 1. All relation types in the speaker-aware temporal graph.

	Intra-Speaker				Inter-Speaker			
r_{ij}	1	2	3	4	5	6	7	8
$pos(i,j)$	$>$	$\leq$	$>$	$\leq$	$>$	$\leq$	$>$	$\leq$
$I_s(i)$	A	A	B	B	A	A	B	B
$I_s(j)$	A	A	B	B	B	B	A	A

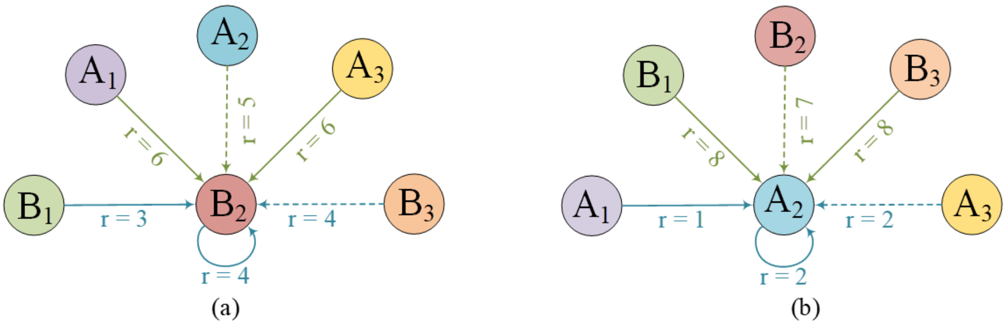

Figure 4. An example of a speaker-aware temporal graph. (**a**) Aggregate neighbor information of target node B_2. (**b**) Aggregate neighbor information of target node A_2.

3.1.2. Dual-Task Reasoning Temporal Graph

The objective of this paper is to identify the sentiment label and the act label of each utterance in a conversation, which correspond to the DAR task and DSC task, respectively, in the dual-task reasoning process. An utterance corresponds to two nodes: the sentiment node and act node. We design a dual-task reasoning temporal graph to model an intra-task and inter-task semantic interaction. Intra-task refers to the interaction between sentiment tasks of utterance when both the target node and source node are sentiment nodes. The same is true for act tasks. Inter-task refers to the interaction between sentiment and act tasks of utterance when the target node is a sentiment node and the source node is an act node, and vice versa. Formally, the heterogeneous graph consists of 2N dual nodes: N sentiment nodes and N act nodes. Two types of utterance nodes model complex intra-task and inter-task interactions, and twelve meta-relations with node-edge-type-dependent parameters represent heterogeneous interactions within the dialogue graph: temporal relations of both the intra- and inter-task. Table 2 lists the definitions of all relation types, where $I_t(i)$ indicates that utterance node i belongs to the sentiment node (S) in the DSC or the act node (D) in the DAR. An example of DTRT-HGT is shown in Figure 5, where the neighbors of the current utterance based on the relation triplet are the entire set of utterances belonging to all tasks in the conversation, including the utterance itself.

Table 2. All relation types in a dual-task reasoning temporal graph.

	Intra-Task						Inter-Task					
r_{ij}	1	2	3	4	5	6	7	8	9	10	11	12
$pos(i,j)$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
$I_t(i)$	S	S	S	A	A	A	S	S	S	A	A	A
$I_t(j)$	S	S	S	A	A	A	A	A	A	S	S	S

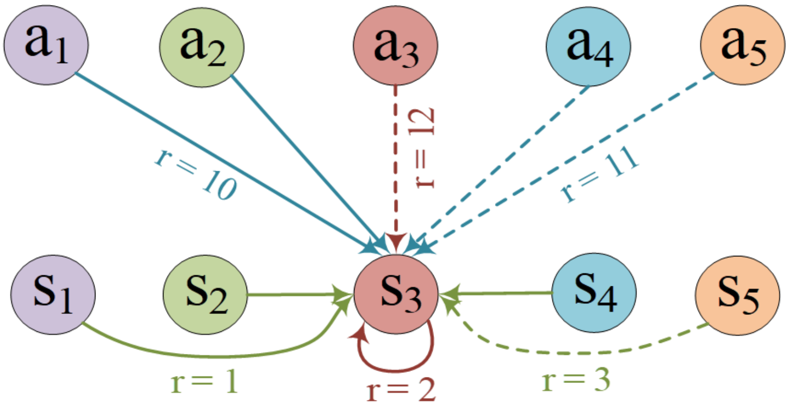

Figure 5. An example of a dual-task reasoning temporal graph.

3.2. Dialogue Understanding

3.2.1. Context-Level Feature Extraction

We use BiLSTM [25] as the utterance encoder applied to the word embeddings of the utterance u_i to capture the intra-sentence dependencies and temporal relationships between words, producing a series of hidden states. Then, we input $H_{u,i}$ into a max pooling layer to obtain the representation of each u_i, generating the initial utterance representation $H_{u,i} = \left(h^0_{u,i}, \ldots, h^{l_i}_{u,i}\right)$, where l_i is the length of u_i. In addition to BiLSTM, we also study the effects of different pre-trained language models (PLMs) as utterance encoders in Section 4.3.

3.2.2. Fine-Grained-Level Feature Extraction

The initial utterance representation obtained only through context-level feature extraction is not enough; we also need to extract fine-grained-level features to get more complete information. Traditional CNN methods tend to over-extract at the fine-grained level, disrupt existing sequence information, and compromise features from the contextual level. Therefore, we employ three smooth hybrid CNN groups to connect features at the contextual level and fine-grained level, reducing conflicts between them. The hybrid CNN group is composed of Conv1D, a hyperbolic linear unit (HLU) [26] activation function, and average pooling:

$$H_c = \mathrm{Conv1D}(H_i) \tag{1}$$

$$H_p = aveage - pooling(H_c) \tag{2}$$

$$H_f = \mathrm{HLU}(H_p) \tag{3}$$

where H_c, H_p, and H_f are the outputs of the convolution layer, pooling layer, and after the HLU activation function, respectively. During the feature extraction process, one-dimensional convolution has a smoother effect compared to high-dimensional convolution, resulting in less damage to sequence information. The HLU function is chosen as the activation function because its average output is close to 0, which reduces the offset between natural gradients and normal gradients and makes the convolutional process smoother. The HLU function equation is:

$$f(x) = \begin{cases} x, & x > 0 \\ ax/(1-x), & x < 0 \end{cases} \tag{4}$$

3.2.3. Speaker-Aware Temporal Dependencies Heterogeneous Graph Transformer

We use HGT as a graph neural network encoder and apply it to the speaker-aware temporal graph. As shown in Figure 6, there are three basic operations in SATD-HGT: Attention is used to estimate the importance of each neighboring node, Message is used to extract features of each neighboring node, and Aggregate uses attention weights to aggregate neighbor information. The process of aggregating all neighboring utterance nodes u_s of the target utterance node u_t in the graph transformer module can be simply represented as:

$$H^{l}_{u_t} \leftarrow \underset{\forall u_s\in N(u_t),\forall r\in R(u_t,u_s)}{Aggregate}\left(Attention(u_s,r,u_t)\cdot Message(u_s,r,u_t)\right) \tag{5}$$

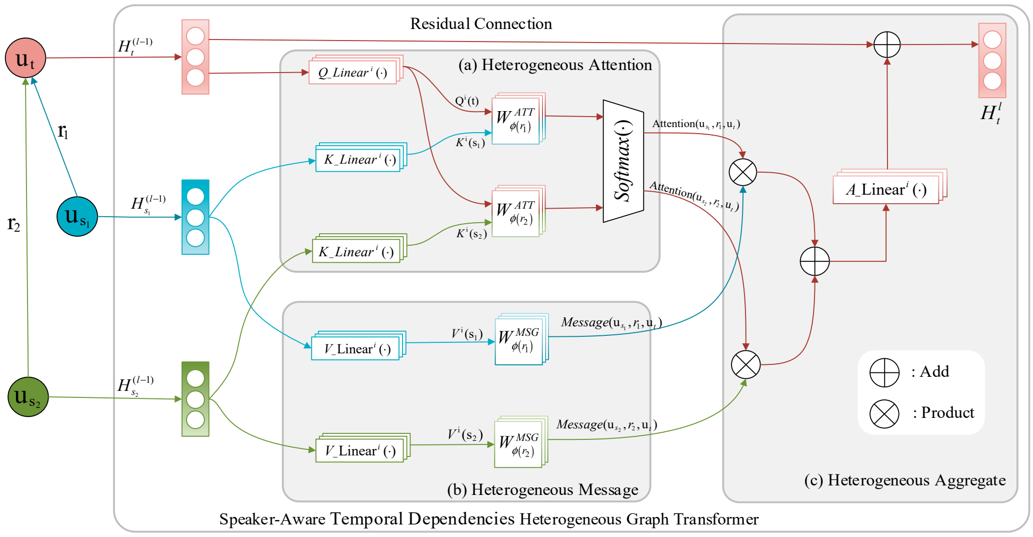

Figure 6. The overall architecture of a speaker-aware heterogeneous graph transformer.

Heterogeneous Attention. We map the current utterance u_s to a key vector $K^{i}_{u_s}$ and its neighborhood utterance u_t into a query vector $Q^{i}_{u_t}$. For each node pair $r = (u_s, u_t)$, the multi-head attention is defined as follows:

$$\mathrm{HAttention}\,(u_s,r,u_t) = \underset{\forall u_s\in N(u_t)}{\mathrm{Softmax}}\left(\Vert_{i\in[1,h]}\, HATT_{head}{}^{i}(u_s,r,u_t)\right) \tag{6}$$

$$HATT_{head^{i}}(u_s,r,u_t) = \left(K^{i}(u_s)W^{ATT}_{\phi(r)}Q^{i}\left(u_t\right)^{T}\right)\cdot\frac{\mu_{\langle\tau(u_s),\phi(r),\tau(u_t)\rangle}}{\sqrt{d}} \tag{7}$$

$$\begin{aligned} K^{i}_{u_s} &= K - Linear^{i}_{\tau(u_s)}\left(H^{l-1}_{u_s}\right)\\ Q^{i}_{u_t} &= Q - Linear^{i}_{\tau(u_t)}\left(H^{l-1}_{u_t}\right) \end{aligned} \tag{8}$$

where $N(u_t)$ denotes the neighborhood of current utterance u_t and h is the number of attention heads. A distinct edge-based matrix $W^{ATT}_{\varnothing(r)}$ is utilized in case there are multiple types of edges between the same node type pair, while μ is a prior tensor indicating the general significance of each meta relation triplet. Note that the softmax process is to make the sum of attention vectors of all neighborhood utterances equal to 1.

Heterogeneous Message. Similarly, we would like to incorporate the meta relations of edges into the message-passing process to distinguish the differences between nodes and edges of different types. For each node pair $r = (u_s, u_t)$, its multi-head message is calculated by:

$$HMessage(u_s,r,u_t) = \underset{i\in[1,h]}{\Vert}\; HMSG_{head^{i}}(u_s,r,u_t) \tag{9}$$

To get the i-th head message $HMSG_{\mathrm{head}^{i}}(u_s,r,u_t)$, we first apply a linear projection $V - Linear^{i}_{\tau(u_s)}$ to project the $\tau(u_s)$-type source node u_s into the i-th message vector and then incorporate the edge dependency with a matrix $W^{MSG}_{\phi(r)}$. Finally, we concatenate all h message heads and get $HMessage(u_s,r,u_t)$ for each node pair.

Heterogeneous Aggregation. The final step is to aggregate heterogeneous multi-head attention and messages of utterances, thereby aggregating the information from neighbors into the target nodes. We can use the attention vector as the weight to average the corresponding messages from the source nodes and get the updated vector as:

$$\overset{\sim l}{H}_{u_t} = \underset{\forall u_s \in N(u_t)}{\oplus} \left(\text{HAttention}(u_s, r, u_t) \cdot HMessage(u_s, r, u_t)\right) \tag{10}$$

where $\oplus$ denotes the overlay operation. To map the target node u_t vector back to its type-specific distribution, the residual connection is used to generate the final updated embeddings:

$$H^l_{u_t} = \theta \cdot A - Linear_{\tau(u_t)}\left(\overset{\sim l}{H}_{u_t}\right) + (1-\theta) \cdot H^{l-1}_{u_t} \tag{11}$$

where θ is a trainable parameter. By stacking L layers, we get the final embedding for the target node u_t, i.e., $H_{u_t} = H^L_{u_t}$. Now, we obtain utterance representations incorporating global and local information:

$$H = H_f + H_{u_t} \tag{12}$$

3.3. Initial Estimation

To enhance task specificity by amplifying the differences in information between different tasks, two independent BiLSTMs are applied on H to obtain the hidden states of utterances for sentiments and acts separately: $H^0_s = \text{BiLSTM}_{DSC}(H)$ and $H^0_a = \text{BiLSTM}_{DAR}(H)$, where $H^0_s = \left\{h^0_{s,1}, \ldots, h^0_{s,N}\right\}$ and $H^0_a = \left\{h^0_{a,1}, \ldots, h^0_{a,N}\right\}$. Then, H^0_s and H^0_a are fed into the sentiment decoder and act decoder, respectively, to generate the initial label estimation distribution:

$$Y^0_{s,i} = \text{softmax}\left(W^s h^0_{s,i} + b_s\right) \tag{13}$$

$$Y^0_{a,i} = \text{softmax}\left(W^a h^0_{a,i} + b_a\right) \tag{14}$$

3.4. Recurrent Dual-Task Reasoning

At step t, the recurrent dual-task reasoning module receives two inputs: (1) the hidden states H^{t-1}_s and H^{t-1}_a of the two tasks and (2) the label distributions Y^{t-1}_s and Y^{t-1}_s of the two tasks.

3.4.1. Prediction Labels

The label information should be represented in vector form and participate in calculations to achieve prediction-level interactions. Using Y^{t-1}_s and Y^{t-1}_a multiplied by the sentiment label embedding matrix M^e_s and act label embedding matrix M^e_a, respectively, we obtain the sentiment label representation $E^t_S = \left\{e^0_{s,1}, \ldots, e^0_{s,N}\right\}$ and act label representation $E^t_A = \left\{e^0_{a,1}, \ldots, e^0_{a,N}\right\}$. The computation of the sentiment and act label for each utterance is as follows:

$$e^t_{s,i} = \sum_{j=1}^{N_S} y^{(j,t-1)}_{s,i} \cdot v^j_s \tag{15}$$

$$e^t_{a,i} = \sum_{j'=1}^{N_A} y^{(j',t-1)}_{a,i} \cdot v^{j'}_a \tag{16}$$

where v^j_s and $v^{j'}_a$ are the label embeddings of sentiment class j and act class j', respectively.

3.4.2. Dual-Task Reasoning Temporal Dependencies Heterogeneous Graph Transformer

To achieve self-interaction and cross-interaction between semantic information and predictive labels, for each node in DRTD-HGT, the corresponding utterance label embeddings of the two tasks are added to its hidden state:

$$\begin{aligned} h_{s,i}^{t} &= h_{s,i}^{t-1} + e_{s,i}^{t} + e_{a,i}^{t} \\ h_{a,i}^{t} &= h_{a,i}^{t-1} + e_{s,i}^{t} + e_{s,i}^{t} \end{aligned} \tag{17}$$

Thus, the representation of each node contains task-specific semantic features and predictive label information for the two tasks, which are then merged into the relationship inference process to achieve semantics-level and prediction-level interactions. The obtained H_s^t and H_a^t both have N vectors separately corresponding to N sentiment nodes and N act nodes on DRTD-HGT, which are then input into the dual-task relationship reasoning of DRTD-HGT. Specifically, the node update process of DRTD-HGT is similar to that of SATD-HGT and can be expressed by the formula:

$$h_{u_t}^{t} = \text{A-Linear}_{\tau(u_t)}\left(\sigma\left(h_i^t\right)\right) + h_i^{t-1} \tag{18}$$

3.4.3. Output Layer

For each task, a task-specific BiLSTM (TS-BiLSTM) is used to generate a series of new hidden states that are more specific to the task.

$$\begin{aligned} H_s^t &= \text{TS} - BiLSTM_S\left(\overline{H}_s^t\right) \\ H_a^t &= \text{TS} - BiLSTM_A\left(\overline{H}_a^t\right) \end{aligned} \tag{19}$$

Then, H_s^t and H_a^t are sent to the sentiment decoder and act decoder to obtain Y_s^t and Y_a^t, respectively.

3.5. Joint Training

The MIRER model uses cross-entropy loss function $\mathcal{L}_*^{t-1}$ to control the accuracy of the predicted labels generated for the sentiment and act at step $t-1$, which provides useful label information for t-step reasoning. Hinge loss $\mathcal{L}_*^{(t,t-1)}$ is used to encourage the two tasks to learn more beneficial knowledge from each other during the dual-task recurrent reasoning process. As the number of steps t increases, the estimated label distribution can be gradually improved. The dialogue sentiment classification objection is formulated as:

$$L_{DSC}^{t-1} = \sum_{i=1}^{N} \sum_{j=1}^{N_S} \hat{y}_{s,i}^{j} log\left(y_{s,i}^{j}\right) \tag{20}$$

$$L_{DSC}^{(t,t-1)} = \sum_{i=1}^{N} \sum_{j=1}^{N_S} \hat{y}_{s,i}^{j} max\left(0, 1 - y_{s,i}^{t-1} \cdot y_{s,i}^{t}\right) \tag{21}$$

$$L_{DSC} = \sum_{t=0}^{T-1} L_{DSC}^{t-1} + \alpha * \sum_{t=1}^{T-1} L_{DSC}^{(t,t-1)} \tag{22}$$

where $\mathcal{L}_{DSC}$ is the weighted sum of $\mathcal{L}_*^{t-1}$ and $\mathcal{L}_*^{(t,t-1)}$, with a hyperparameter θ balancing the two kinds of punishments. The cross-entropy loss of the label distributions generated at the final step t is as follows:

$$L_{DSC}^{t-1} = \sum_{i=1}^{N} \sum_{j=1}^{N_S} \hat{y}_{s,i} \log\left(y_{s,i}^{t}\right) \tag{23}$$

The total loss of the DSC task is obtained as follows:

$$L_S = L_{DSC} + L_{DSC}^t \tag{24}$$

Similarly, the total loss of the DAR task ($\mathcal{L}_A$) can be derived similarly to Equations (19)–(23). The final training joint objective of MIRER is the sum of the total losses of DSC and DAR:

$$L = L_S + L_A \tag{25}$$

4. Experiments

4.1. Datasets and Settings

To validate the effectiveness of MIRER, we conducted experiments on two publicly available datasets, Mastodon [6] and Dailydialog [27], which have sentiment and act labels. The Mastodon dataset consists of 269 dialogue segments in the training set and 266 dialogue segments in the test set. The sentiment labels are positive, negative, and neutral. Dialogue act labels of the utterances are annotated with one of fifteen labels: disagreement, agreement, statement, suggest, request, thanking, sympathy, exclamation, and so on. As there is no official validation set, this paper follows the same partitioning method as Qin et al. [7], with 243 dialogue segments used for training, 26 dialogue segments for validation, and 266 dialogue segments for testing. The Dailydialog dataset is consistent with the official partitioning in the original dataset, with 11,118 dialogue segments in the training set, 1000 dialogue segments in the validation set, and 1000 dialogue segments in the test set. The sentiment labels of utterances are annotated with one of seven labels: neutral, happiness, surprise, sadness, anger, disgust, and fear. Dialogue act labels contain inform, questions, directives, and commissive.

This experiment uses the Adam optimizer for training and 300-dimensional Glove vectors for word embedding. The learning rate was set to 1×10^{-3}, and the weight decay was set to 1×10^{-8}. For the Mastodon and DailyDialog datasets, the number of epochs was set to 100 and 50, respectively. The number of steps in the dual-task recurrent reasoning was set to 3 and 1, respectively. The hidden state (label embedding) dimensions were set to 128 and 256, and the dropout rates were set to 0.2 and 0.3 to alleviate overfitting.

To verify the validity of the MIRER model, we compared it to the following baseline models, which can be divided into three categories. The first group is a dialogue act recognition method modeled separately, including HEC [12] and CASA [3]. The second group is a dialogue sentiment classification method with separate modeling, including DialogueRNN [15] and DialogueGCN [28]. The third group is a joint approach to modeling dialogue act recognition and sentiment classification tasks, including JointDAS [6], IIIM [4], DCR-Net [18], BCDCN [20], Co-GAT [7], and TSCL [21].

4.2. Main Results

Similar to previous works, macro-averaged precision (P), recall (R), and F1 were used as evaluation metrics for the DSC and DAR tasks on the Dailydialog dataset, while the weighted-average F1 scores were used as the evaluation metric for the DAR task and neutral sentiment labels were ignored for the DSC task on the Mastodon dataset. Table 3 presents the experimental results of MIRER compared to baseline models.

Except for JointDAS and IIIM, the performance of the joint task model was consistently better than that of the baseline model trained on individual tasks on two datasets, indicating the necessity of joint training for DAR and DSC. In the joint learning of the two tasks, our MIRER achieved better results on evaluation metrics on two datasets. Specifically, on the Mastodon dataset, MIRER achieve a 4.9% improvement in the F1 score for the DSC task and a 2.1% improvement in the F1 score for the DAR task, and similar improvement trends were observed on the Dailydialog dataset. The satisfactory results are attributed to the following. (1) Our framework not only considers context-level features in the dialogue but also integrates fine-grained features through a hybrid CNN group and obtains global semantic features based on a heterogeneous graph network, thereby obtaining deep dialogue utterance information. (2) Intuitively, prediction can provide feedback for semantics, and semantics can rethink and help to reverse the prediction. In the dual-task recurrent reasoning, the estimated label distribution of the previous step is used as a prediction clue for the current step to generate new predictions, improving the accuracy of dialogue act recognition and sentiment classification by continuously correcting the label information. (3) We construct heterogeneous graphs related to speakers and dual tasks to model the complex interactions within and between speakers as well as within and between tasks. We introduce multiple types of meta-relations to model different edges. With these node-edge-type-dependent structures and parameters, MIRER can better utilize the structural knowledge of the dialogue for node representation compared to traditional homogeneous graphs.

Table 3. Experiment results.

Models	Mastodon						DailyDialog					
	SC			DAR			SC			DAR		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
HEC	-	-	-	56.1	55.7	56.5	-	-	-	77.8	76.5	77.8
CASA	-	-	-	56.4	57.1	55.7	-	-	-	78.0	76.5	77.9
DialogueRNN	41.5	42.8	40.5	-	-	-	40.3	37.7	44.5	-	-	-
DialogueGCN	42.4	43.4	41.4	-	-	-	43.1	44.5	41.8	-	-	-
JointDAS	36.1	41.6	37.6	55.6	51.9	53.2	35.4	28.8	31.2	76.2	74.5	75.1
IIIM	38.7	40.1	39.4	56.3	52.2	54.3	38.9	28.5	33.0	76.5	74.9	75.7
DCR-Net	43.2	47.3	45.1	60.3	56.9	58.6	56.0	40.1	45.4	79.1	79.0	79.1
BCDCN	38.2	62.0	45.9	57.3	61.7	59.4	55.2	45.7	48.6	80.0	80.6	80.3
Co-GAT	44.0	53.2	48.1	60.4	60.6	60.5	65.9	45.3	51.0	81.0	78.1	79.4
TSCL	46.1	58.7	51.6	61.2	61.6	60.8	56.6	49.2	51.9	78.8	79.8	79.3
MIRER	55.6	57.7	56.5	64.1	61.5	62.9	60.9	47.3	53.2	80.8	80.5	80.7

4.3. Ablation Study

We conducted ablation experiments to verify the effectiveness of each component in MIRER, as shown in Table 4. (1) Removing label embedding leads to the inability to achieve prediction-level interaction, and the sharp decline in results proves that the method of using label information to achieve prediction-level interaction effectively improves dual-task reasoning by capturing explicit dependency relations. (2) SATD-HGT captures intra- and inter-speaker semantic state transitions, providing global semantic information for the dual task. Without it, some potential features would be lost, resulting in a decrease in results. (3) CNNs can help solve the problem of incomplete integration of local context information in dialogue. Combining it with SATD-HGT helps provide deep semantic information for both tasks. (4) The results after removal show that DTRD-HGT is the core of the MIRER model, playing a crucial role in dual-task reasoning for both semantic-level and prediction-level interaction information.

Table 4. Results of ablation experiments on F1 score.

Variants	Mastodon		DailyDialog	
	DSC	DAR	DSC	DAR
MIRER	56.5	62.9	53.2	80.7
w/o Label Embeddings	53.6	61.2	50.6	78.9
w/o SATD-HGT	54.8	61.7	50.4	79.7
w/o CNNs	55.3	60.9	50.9	79.5
w/o DTRD-HGT	54.2	60.4	49.7	79.8

This once again demonstrates the effectiveness of the proposed model in our paper. To further explore the effect of joint pre-training language models BERT [29], RoBERTa [30], and XLNet [31], we combined the PLM encoders with our model by replacing the BiLSTM utterance encoder in MIRER while keeping other components the same. We conducted experiments on the Mastodon dataset, and the results are shown in Table 5. The PLM encoders have strong conversational context understanding ability and to some extent can deal with non-standard language, colloquial expressions, and other text noise. We found that even without the interaction between utterances and dual-task mutual learning, they can still obtain good results. In contrast, Co-GAT only models semantic-level interactions, and its advantages are attenuated by PLMs. Therefore, using PLMs as utterance encoders, Co-GAT results in act recognition and sentiment classification that are less improved than MIRER models. Stacking MIRER on the PLM encoder further significantly improves on F1 because we not only capture context-level features, fine-grained features, and global semantic features but also realize self-interaction and mutual interaction between semantic and predictive levels and integrate temporal relations, which complements the high-quality semantic information captured by the PLM encoder. The validity of the proposed model is proven again.

Table 5. Results based on different PLM encoders.

Models		Mastodon					
		DSC			DAR		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
BERT	+Linear	61.8	61.1	60.6	70.2	67.5	68.8
	+Co-GAT	66.1	58.1	61.5	70.7	67.6	69.1
	+MIRER	65.1	66.3	65.7	72.9	71.7	72.3
RoBERTa	+Linear	59.7	54.4	55.7	61.4	61.8	61.6
	+Co-GAT	64.3	58.8	61.3	67.5	64.8	66.1
	+MIRER	62.5	64.6	63.5	69.4	67.8	68.6
XLNet	+Linear	56.6	60.9	58.7	63.4	61.8	62.6
	+Co-GAT	66.1	65.8	65.9	69.2	66.0	67.5
	+MIRER	67.2	68.3	67.7	70.9	68.5	69.7

4.4. Error Analysis

By examining the sentiment classes and act classes present in the two datasets, we found that a significant proportion of neutral samples is prevalent in these datasets, which leads to data imbalance. During the training process, the model struggles to effectively understand the characteristics of other genuine sentiments within the samples. Consequently, the experimental results of the MIRER model on two datasets yield subpar recall rates. We can observe from Table 6 that the sentiment classes with limited available data, such as "fear" and "disgust" in the DailyDialog dataset, are notably poor in the initial step of sentiment prediction. However, as the step number increases, we observed a gradual reduction in prediction errors, indicating an improvement in the model's predictive performance. For example, in the 10th utterance said by speaker A, the sentiment of this utterance is difficult to judge only based on its text, and the prediction of the first step is wrong. However, the prediction of the second step is modified to be correct, which is due to the context utterances and act label information.

Table 6. Examples of initial and final sentiment prediction in the MIRER model.

Speaker	Utterance	Step = 1	Step = 2	Gold
A	That dress is very pretty. Why don't you like it?	happiness	happiness	happiness
B	It's too loud.	neutral	neutral	neutral
A	We have been looking around for many hours. What on earth are you looking for?	neutral	angry	angry
B	Well, you know, those styles or colors don't suit me.	neutral	neutral	neutral
A	What style do you want?	neutral	neutral	neutral
B	I want to buy a V-neck checked sweater, and it should be tight.	neutral	neutral	neutral
A	Oh, I see. How about the color?	neutral	neutral	neutral
B	Quiet color.	neutral	neutral	neutral
A	I know a shop selling this kind of sweater.	neutral	neutral	neutral
B	Really? Let's go there.	neutral	surprise	surprise

As shown in Table 7, the act label is incorrectly inferred to be “answer”, and the label is further corrected if we consider the intra-task and cross-task dependencies and temporal relationships while combining the interaction at the prediction level. Moreover, when the distribution of “negative” sentiment labels is large, the model is more likely to predict act labels such as “disagree” or “symmetry”. Similarly, the larger distribution of act labels such as “thinking”, “agreement”, and “suggestion” made it easier for the model to predict “positive” sentiment labels. In conclusion, dual-task learning effectively utilizes the explicit correlation between labels, improves label prediction through mutual learning, and improves the performance of the model. This approach also improves the interpretability of the correlation, consistent with human cognition. Overall, this error study emphasizes the importance of addressing misclassifications, especially sentiment classes with a small amount data available and too many neutral labels. Our recurrent reasoning improvement mechanism exhibits promise in enhancing the accuracy of sentiment predictions, further underscoring the potential of our MIRER model in real-world emotion-recognition applications.

Table 7. Examples of initial and final act prediction in the MIRER model.

Speaker	Utterance	Step = 1	Step = 2	Gold
A	My face?	question	question	question
B	Ugly?	question	question	question
A	It’s more likely than you think.	answer	statement	statement
B	Very wrong.	disagreement	disagreement	disagreement

5. Conclusions

In this paper, we propose a multiple information-aware recurrent reasoning network to jointly model DAR and DSC. First, we use BiLSTM as an utterance encoder to extract context-level feature vectors independent of dialogue context. Second, to capture deep semantic information more comprehensively, we use a hybrid CNN group to smoothly send sequence information to multiple local information layers for fine-grained feature extraction. This BiLSTM-connected hybrid CNN group method can not only keep the integrity of context sequence information but also extract more fine-grained information. In addition, we design a speaker-aware temporal reasoning graph to capture global semantic features that are speaker-, context-, and temporal-sensitive. To realize self-interaction and mutual interaction between the semantic level and prediction level in the process of dual-task recurrent reasoning, we design a dual-task temporal reasoning graph and apply HGT as the encoder of the heterogeneous graph neural network. Specifically, we design different types of nodes and meta relationships with node-edge-type-dependent parameters for these two conversational heterogeneous graphs to characterize heterogeneous interactions in the graphs. Finally, task-specific hidden states generated by applying two task-specific BiLSTMs also play an important role in MIRER and have some sequence-tag-aware reasoning ability. Experiments verify the effectiveness of the proposed model and achieve advanced performance beyond the existing baseline. We also analyze the benefits of incorporating pre-trained language models into federated models and find that this combined approach is very beneficial for improving model performance.

Author Contributions: Conceptualization and methodology, S.L. and X.C.; software, validation, and formal analysis, S.L. and X.C.; writing—original draft preparation, X.C.; writing—review and editing, S.L. and X.C.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The codes used in this manuscript are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ghosal, D.; Majumder, N.; Mihalcea, R.; Poria, S. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Morristown, NJ, USA, 2021; pp. 1435–1449.
2. Enayet, A.; Sukthankar, G. Improving the generalizability of collaborative dialogue analysis with multi feature embeddings. *arXiv* **2023**, arXiv:2302.04716.
3. Raheja, V.; Tetreault, J. Dialogue act classification with context-aware self-attention. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 3–5 June 2019; pp. 3727–3733.
4. Kim, M.; Kim, H. Integrated neural network model for identifying speech acts, predictors, and sentiments of dialogue utterances. *Pattern Recognit. Lett.* **2018**, *101*, 1–5. [CrossRef]
5. Bibi, M.; Abbasi, W.A.; Aziz, W.; Khalil, S.; Uddin, M.; Iwendi, C.; Gadekallu, T.R. A novel unsupervised ensemble framework using concept based linguistic methods and machine learning for Twitter sentiment analysis. *Pattern Recognit. Lett.* **2022**, *158*, 80–86. [CrossRef]
6. Cerisara, C.; Jafaritazehjani, S.; Oluokun, A.; Le, H. Multi-task dialog act and sentiment recognition on mastodon. In Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Association for Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 745–754.
7. Qin, L.; Li, Z.; Che, W.; Ni, M.; Liu, T. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 18 May 2021; pp. 13709–13717.
8. Xing, B.; Tsang, I. Darer: Dual-task temporal relational recurrent reasoning network for joint dialog sentiment classification and act recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*; Association for Computational Linguistics: Morristown, NJ, USA, 2022; pp. 3611–3621.
9. Ko, Y. New feature weighting approaches for speech-act classification. *Pattern Recognit. Lett.* **2015**, *51*, 107–111. [CrossRef]
10. Kang, S.; Ko, Y.; Seo, J. Hierarchicalspeech-actclassificationfordiscourse analysis. *Pattern Recognit. Lett.* **2013**, *34*, 1119–1124. [CrossRef]
11. Liu, Y.; Han, K.; Tan, Z.; Lei, Y. Using context information for dialog act classification in dnn framework. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2170–2178.
12. Kumar, H.; Agarwal, A.; Dasgupta, R.; Joshi, S. Dialogue act sequence labeling using hierarchical encoder with crf. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 3440–3447.
13. Li, R.; Lin, C.; Collinson, X.; Li, M.; Chen, G. A dual-attention hierarchical recurrent neural network for dialogue act classification. In Proceedings of the 23rd Conference on Computational Natural Language Learning, Hong Kong, China, 3–4 November 2019; pp. 383–392.
14. Sunitha, D.; Patra, R.K.; Babu, N.V.; Suresh, A.; Gupta, S.C. Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in india and european countries. *Pattern Recognit. Lett.* **2022**, *158*, 164–170. [CrossRef] [PubMed]
15. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6818–6825.
16. Ma, H.; Wang, J.; Lin, H.; Pan, X.; Zhang, Y.; Yang, Z. A multi-view network for real-time emotion recognition in conversations. *Knowl.-Based Syst.* **2022**, *236*, 107751. [CrossRef]
17. Li, J.; Lin, Z.; Fu, P.; Wang, W. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*; Association for Computational Linguistics: Morristown, NJ, USA, 2021; pp. 1204–1214.
18. Qin, L.; Che, W.; Li, Y.; Liu, T. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 8665–8672.
19. Lin, H.; Liu, J.; Zheng, Z.; Hu, R.; Luo, Y. Multi-task network for joint dialog act recognition and sentiment classification. *Comput. Eng. Appl.* **2023**, *59*, 104–111.
20. Li, J.; Fei, H.; Ji, D. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In Proceedings of the 28th International Conference on Computational Linguistics, online, 8–13 December 2020; pp. 616–626.
21. Xu, Y.; Yao, E.; Liu, C.; Liu, Q.; Xu, M. A novel ensemble model with two-stage learning for joint dialog act recognition and sentiment classification. *Pattern Recognit. Lett.* **2023**, *165*, 77–83. [CrossRef]
22. Hu, Z.; Dong, Y.; Wang, K.; Sun, Y. Heterogeneous graph transformer. In Proceedings of the Web Conference, New York, NY, USA, 25–29 April 2022; pp. 2704–2710.
23. Roy, S.; Min, K.; Tripathi, S.; Guha, T.; Majumdar, S. Learning Spatial-Temporal Graphs for Active Speaker Detection. *arXiv* **2021**, arXiv:2112.01479.
24. Li, J.; Liu, M.; Wang, Y.; Zhang, D.; Qin, B. A speaker-aware multiparty dialogue discourse parser with heterogeneous graph neural network. *Cogn. Syst. Res.* **2023**, *79*, 15–23. [CrossRef]

25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
26. Li, J.; Xu, H.; Deng, J.; Sun, X. Hyperbolic linear units for deep convolutional neural networks. In Proceedings of the 2016 International Joint Conference on Neural Networks, Vancouver, BC, Canada, 24–29 July 2016; pp. 353–359.
27. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 1 December 2017; pp. 986–995.
28. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 154–164.
29. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
30. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized Bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
31. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.G.; Salakhutdinov, R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advance in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2019; pp. 5754–5764.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.