Integrating User History into Heterogeneous Graph for Dialogue Act Recognition

Dong Wang[1,2*], **Ziran Li**[1,2*], **Hai-Tao Zheng**[1,2†], **Ying Shen**[3†]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Tsinghua ShenZhen International Graduate School, Tsinghua University
[3]School of Intelligent Systems Engineering, Sun Yat-Sen University
{wangd18,lizr18}@mails.tsinghua.edu.cn
zheng.haitao@sz.tsinghua.edu.cn,sheny76@mail.sysu.edu.cn

Abstract

Dialogue Act Recognition (DAR) is a challenging problem in Natural Language Understanding, which aims to attach Dialogue Act (DA) labels to each utterance in a conversation. However, previous studies cannot fully recognize the specific expressions given by users due to the informality and diversity of natural language expressions. To solve this problem, we propose a Heterogeneous User History (HUH) graph convolution network, which utilizes the user's historical answers grouped by DA labels as additional clues to recognize the DA label of utterances. To handle the noise caused by introducing the user's historical answers, we design sets of denoising mechanisms, including a History Selection process, a Similarity Re-weighting process, and an Edge Re-weighting process. We evaluate the proposed method on two benchmark datasets MSDialog and MRDA. The experimental results verify the effectiveness of integrating user's historical answers, and show that our proposed model outperforms the state-of-the-art methods.

1 Introduction

Dialogue Acts Recognition (DAR) is an important but challenging task in Natural Language Understanding (NLU), which aims to attach Dialogue Act (DA) labels to each utterance in a conversation and recognize the speaker's intention. Automatic DAR can be applied to many applications such as question answering, speech recognition and dialogue systems (Higashinaka et al., 2014; Khanpour et al., 2016). In this work, different from recognizing a single DA, we focus on the task of recognizing multiple DA in a multi-party conversation (i.e., a forum). The latter is difficult but is more common in the real world. Figure 1 shows an example of multiple DA recognition in a tech forum.

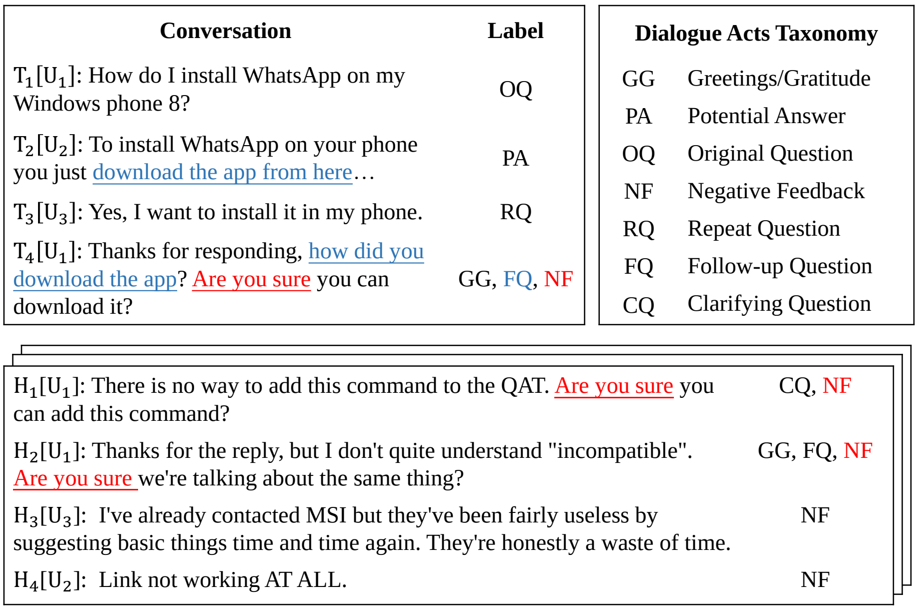
Previous studies have proposed deep learning models, which approach DAR as a multi-classification problem (Ji et al., 2016; Lee and Dernoncourt, 2016) or a sequence labeling problem (Kumar et al., 2018; Chen et al., 2018; Raheja and Tetreault, 2019; Li et al., 2019). Most of these approaches assume that utterances are sequentially organized, ignoring the rich interaction between multiple users in a conversation. To alleviate such a problem, some recent studies exploit graph-structured networks (Ghosal et al., 2019; Hu et al., 2019), which leverage speaker interaction of the interlocutors to model conversational context. However, due to the informality and diversity of natural language expressions, the same intention has a very rich form of expression. In forums where many users participate, it is more common for different users to express their intentions using personalized expressions. When encountering the unobvious intention or uncommon expression pattern, existing deep learning methods that generalize the utterance feature into a low dimensional vector may degrade the performance of DA recognition.

Intuitively, the user's historical expressions extracted according to the DA category can be used as a DA-specific clue to help the encoding process recognize the user's inexplicable or uncommon expressions. For example, as shown in Figure 1, we intent to recognize three DA labels for T_4. Although we can easily label T_4 as *Greetings/Gratitude (GG)* simply based on the utterance features (i.e. "*Thanks*"),

[*] indicates equal contribution.

[†] Corresponding authors.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4211–4221
Barcelona, Spain (Online), December 8-13, 2020



Conversation	Label
$T_1[U_1]$: How do I install WhatsApp on my Windows phone 8?	OQ
$T_2[U_2]$: To install WhatsApp on your phone you just download the app from here…	PA
$T_3[U_3]$: Yes, I want to install it in my phone.	RQ
$T_4[U_1]$: Thanks for responding, how did you download the app? Are you sure you can download it?	GG, FQ, NF

Dialogue Acts Taxonomy	
GG	Greetings/Gratitude
PA	Potential Answer
OQ	Original Question
NF	Negative Feedback
RQ	Repeat Question
FQ	Follow-up Question
CQ	Clarifying Question

$H_1[U_1]$: There is no way to add this command to the QAT. Are you sure you can add this command?	CQ, NF
$H_2[U_1]$: Thanks for the reply, but I don't quite understand "incompatible". Are you sure we're talking about the same thing?	GG, FQ, NF
$H_3[U_3]$: I've already contacted MSI but they've been fairly useless by suggesting basic things time and time again. They're honestly a waste of time.	NF
$H_4[U_2]$: Link not working AT ALL.	NF

Figure 1: A tech forum conversation and sampled user's historical answers labeled as *NF*. $T_i[U_j]$: the i-th turn of conversation posted by user U_j. $H_i[U_j]$: the i-th historical answer of user U_j.

to detect the *Follow-up Question (FQ)* label, additional information from T_2 need to be considered. Besides, T_4 seems to be a question and there is no explicit *Negative Feedback (NF)*. However, from the user's historical answers that reflect NF label such as H_1, H_2, we find that the phrase "*are you sure*" indicates a negative feedback that the user doubts about something.

Inspired by this, we propose a Heterogeneous User History (HUH) graph convolution networks, which integrates utterance, conversation, and user's historical answers into multiple DA recognition. The proposed DA recognition process can be divided into two stages. In the 1st-phase, we first extract the hidden features for each utterance by stacking a convolutional neural network (CNN) utterance encoder and a Bi-directional Long Short-Term Memory (BiLSTM) utterance context encoder, and we use the hidden features to predict the initial score of DA labels for each utterance. In the 2nd-phase, we use the user's historical answers as additional information to construct a Heterogeneous User History (HUH) graph, and we use a Relation-weighted graph Convolutional Network (RGCN) (Schlichtkrull et al., 2018) to learn this graph and model the interaction between utterance and user's historical answers before recognizing user's intentions.

Despite the benefits of user's historical answers, improper use may inevitably bring a lot of noise. We thus further design sets of denoising mechanisms. Firstly, we use a History Selection process with similarity measures to filter out the irrelevant user's historical answers. Then, we adopt the learned initial score of DA labels from the 1st-phase to re-weight the similarity matrix between the user's historical answers and conversation, and re-weight the edges in the proposed HUH graph, thereby reducing the noise caused by introducing supplementary information.

The main contributions of our work can be summarized as follows: 1) We propose a novel Heterogeneous User History (HUH) graph convolution network, which models the interaction between users and integrates utterance, conversation and user's historical answers for recognizing user's intent. To the best of our knowledge, we are the first to integrate user's historical answers into heterogeneous graph on DAR. 2) To alleviate the noise issue caused by introducing the user's historical answers, we design sets of denoising mechanisms, including a History Selection process, a Similarity Re-weighting process, and an Edge Re-weighting process. 3) We evaluate our model on two benchmark datasets MSDialog and MRDA. Compared to the state-of-the-art DAR methods, our model achieves better performance. The experimental results verify the effectiveness of incorporating the user's historical answers.

2 Related Work

The goal of the DAR is to assign DA labels to each utterance in a conversation. Early studies on DAR are mostly based on general statistical machine learning methods and approach this task as a multi-class classification problem or a sequence labeling problem, such as Hidden Markov Model (HMM) (Stolcke et al., 2000), Support Vector Machines (SVM) (Surendran and Levow, 2006) and Bayesian Network (Keizer et al., 2002). Recent studies on DAR have proposed deep learning models and have obtained promising results. Deep learning approaches typically model the interaction between adjacent utterances (Ji et al., 2016; Lee and Dernoncourt, 2016). Some researchers capture the dependencies among both utterances and labels with Conditional Random Field (CRF) (Kumar et al., 2018; Chen et al., 2018; Raheja and Tetreault, 2019; Li et al., 2019). Furthermore, Colombo et al. (2020) leverage a sequence to sequence approach to model both the conversation and the global tag dependencies. Besides, some researches explore joint models to solve DAR and sentiment classification simultaneously in a unified framework (Cerisara et al., 2018; Kim and Kim, 2018; Qin et al., 2020). However, these methods assume that utterances are sequentially organized, ignoring the rich interaction process between users in a conversation.

Some recent researches design graph-structured networks to model speaker interaction in a conversation. Hu et al. (2019) first propose a graph-structured network (GSN) to model graph-structured dialogues for response generation. Ghosal et al. (2019) leverage self and inter-speaker interaction of the interlocutors to model conversational context for emotion recognition. Although these methods can play a role, they do not make full use of the user information in the conversation. The recent study (Wen et al., 2018) encodes the user's historical answers to boost community question answering (CQA) task. But it only considers a single user and ignores the noise issue caused by introducing multiple users' historical answers. In this paper, we propose a Heterogeneous User History (HUH) graph convolution networks, which integrates utterance, conversation, and user's historical answers into multiple DA recognition. The experimental results verify the effectiveness of integrating user's historical answers, and show that our proposed model outperforms the state-of-the-art methods.

3 Methodology

Before describing our proposed model, we first introduce the basic mathematical notions and terminologies for the problem of DAR. The task of DAR takes a conversation C as input, which contains a sequence of utterances $\{U_t\}_{t=1}^{N}$. For each utterance U_t (the t-th utterance) in a conversation, we predict a subset of DA labels $y_t = \{y_1^t, y_2^t, ..., y_S^t\}$ that describes the functionality of the utterance from a candidate set of DA labels $D = \{d_1, d_2, ..., d_S\}$. And $y_j^t = \{0, 1\}$ indicates whether the t-th utterance is labeled with DA label d_j. For each DA label, we use the user's historical answers belonging to that DA label to construct a label node. Specifically, for the j-th DA label d_j, we retrieval a series of user's historical answers belonging to this label. We select top-K user's historical answers that are most relevant to the conversation, and sum them by similarity weight to generate a label node e_j corresponding to d_j. Here, K is a hyperparameter.

The overall architecture of our proposed model HUH is shown in Figure 2. HUH mainly contains two phases: 1) The 1st-phase, as shown in the left part, aims to predict the initial score of DA labels for each utterance as a guide to the 2nd-phase process. 2) The 2nd-phase, as shown in the right part, constructs a Heterogeneous User History graph to integrate utterance, conversation, and user's historical answers into multiple DA recognition. Additionally, the proposed denoising mechanisms are presented at the corresponding phase in Figure 2. In the following sections, the details of our framework are given.

3.1 1st-phase: Encoding Utterances

In the 1st-phase, we predict the initial score of DA labels for each utterance. For each word in an utterance, we firstly convert them into pre-trained word-level embeddings with Glove (Pennington et al., 2014) as initialization, and get an utterance representation (the t-th utterance) as $U_t = \{w_1^t, w_2^t, ..., w_L^t\}$. And then we use a CNN (Kim, 2014) followed by max-pooling to extract utterance features as:

$$\hat{u}_t = \mathrm{CNN}(U_t). \tag{1}$$

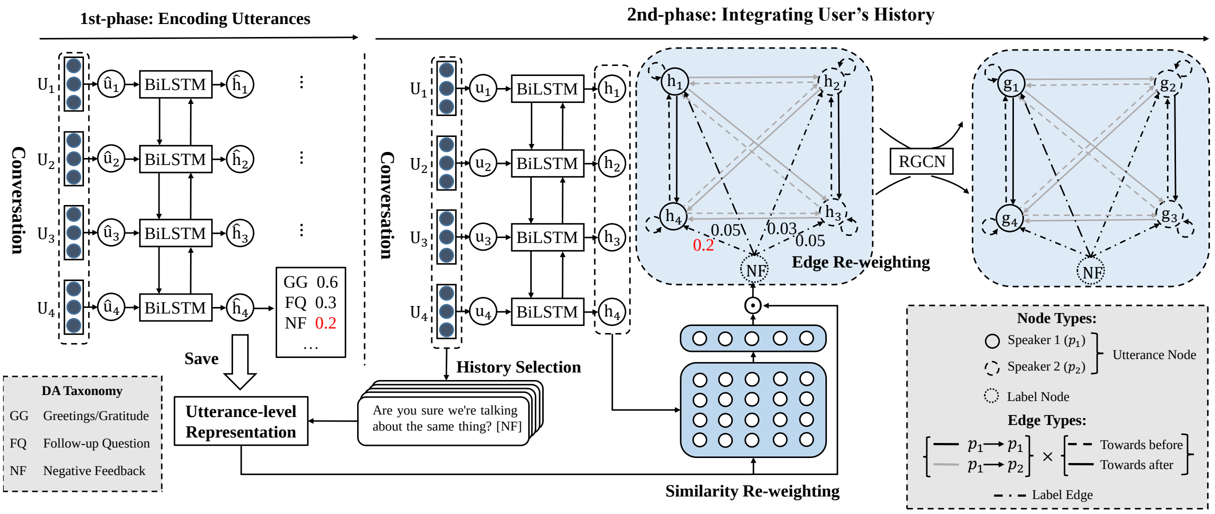


Figure 2: The overall architecture of our proposed Heterogeneous User History graph convolution network.

Based on the local utterance features extracted from CNN, a BiLSTM is applied to gather features from the context:

$$\overrightarrow{\hat{h}_t} = \overrightarrow{\mathrm{LSTM}}(\hat{u}_t, \overrightarrow{\hat{h}_{t-1}}), \tag{2}$$

$$\overleftarrow{\hat{h}_t} = \overleftarrow{\mathrm{LSTM}}(\hat{u}_t, \overleftarrow{\hat{h}_{t+1}}), \tag{3}$$

$$\hat{h}_t = \mathrm{concat}(\overrightarrow{\hat{h}_t}, \overleftarrow{\hat{h}_t}), \tag{4}$$

where $\hat{h}_t \in \mathcal{R}^{2d_h}$ is a sequential context-aware utterance representation for the t-th utterance and d_h is the hidden size of BiLSTM. During the training phase, we save the $\hat{h}_t$ as utterance-level user's historical answers representation and provide it to the 2nd-phase.

With the extracted local textual features $\hat{u}_t$ and context-aware features $\hat{h}_t$, we predict the initial score of DA labels for each utterance:

$$\hat{p}_t = W_\alpha[\hat{u}_t, \hat{h}_t] + b_\alpha, \tag{5}$$

$$\hat{y}_t = \mathrm{sigmoid}(\hat{p}_t), \tag{6}$$

where $[\hat{u}_t, \hat{h}_t]$ represents the concatenated result and W_α,b_α are weight matrices to be learned. $\hat{y}_t \in \mathbb{R}^S$ represents the initial score of DA labels for the t-th utterance and S is the number of the DA labels. In the 1st-phase, we can also calculate the DA labels prediction loss here, denoted as $loss_{1p}$. Though $loss_{1p}$ will not be used as the final prediction, it is also a good auxiliary loss for training 1st-phase.

3.2 2nd-phase: Integrating User's History

To capture the interaction between users and integrate the user's historical answers into the utterance encoding properly, we present a novel Heterogeneous User History graph convolution network. We denote our Heterogeneous User History graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$, where $\mathcal{V}$ stands for node representations and $\mathcal{E}$ represent edges between nodes, $\mathcal{R}$ and $\mathcal{W}$ are the type and weight of the edges.

3.2.1 Graph Node

There are two kinds of nodes in our heterogeneous graph: Utterance Node and Label Node.

Utterance Node To represent utterance nodes, we share the same encoder (parameter sharing) with the 1st-phase to get the sequentially encoded feature vector h_i for all $i \in [1, 2, ..., N]$, where N is the number of utterances in the conversation.

Label Node For each DA label, there is a corresponding label node. And we use the user's historical answers to generate the representation of the label node. At first, we retrieve all the historical answers of the users enrolled in the conversation from the training set, and group the answers according to the DA labels. And then, for the j-th DA label, we select top-K user's historical answers (denoted as $H(j)$) most relevant to the current conversation and convert them into utterance-level representation learned from the 1st-phase. We sum these user's historical answers by weight and generate corresponding label node as e_j for all $j \in [1, 2, ..., S]$:

$$e_j = \sum_{\hat{h}_k \in H(j)} \alpha_k \hat{h}_k, \tag{7}$$

where S is the number of DA labels, $\hat{h}_k$ is the utterance-level representation as **Formula (4)** learned from the 1st-phase and the weights we use here are simple initialized as $\alpha_k = \frac{1}{K}$.

3.2.2 Graph Edge

We define the following types of edges between pairs of nodes to encode various structural information in our graph:

Speaker Edge Every speaker in the conversation is effected by himself and other speakers, resulting in two different edge types: *one to one's self* and *one to others*. In addition, the impact between utterances depends on the relative position that occurs in the conversation: *before* or *after*. As a result, there are a total of $2 \times 2 = 4$ different speaker edge types and there can be N^2 speaker edges in a conversation, where N is the number of utterances in the conversation. For each speaker edge, we use a similarity-based attention module to obtain edge weight:

$$\alpha^s_{i,j} = \text{softmax}(h_i^T W_e [h_1, ...h_N]), j = 1, ..., N, \tag{8}$$

where utterance node $h_i, i \in [1, 2, ...N]$ which has incoming edges with utterance nodes $h_1, ..., h_N$ receives a total weight contribution of 1.

Label Edge We use a label edge to connect the label node with the utterance node. As an illustration, we have a label node e_j and an utterance node h_i, and we construct a directed label edge from e_j to h_i as $e_j \rightarrow h_i$. Each label node $e_j, j \in [1, ..., S]$ is connected to all utterance nodes $h_i, i \in [1, ..., N]$ and there can be $N \times S = NS$ label edges in a conversation. We initial the weight of these label edges as:

$$\alpha^l_{i,j} = \frac{1}{S}, j = 1, ..., S, \tag{9}$$

where S is the number of DA labels and we give each label node the same weight.

3.2.3 Message Passing

To consider various relationships between nodes, we use a relation specific message passing strategy inspired by RGCN (Ghosal et al., 2019), which can be formulated as:

$$z_i^{(l+1)} = \sigma(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{\alpha_{i,j}}{c_{i,r}} W_r^{(l)} z_j^{(l)} + \alpha_{i,i} W_0^{(l)} z_i^{(l)}), \tag{10}$$

where z_j represents a graph node representation, $\alpha_{i,j}$ and $\alpha_{i,i}$ are edge weights, $\mathcal{N}_i^r$ denotes the set of neighbor indices of node i under relation $r \in \mathcal{R}$. $c_{i,r}$ is a problem-specific normalization constant that can either be learned or chosen in advance (such as $c_{i,r} = |\mathcal{N}_i^r|$). σ is an activation function (ReLU), $W_r^{(l)}$ and $W_0^{(l)}$ are learnable parameters.

3.3 Denoising Mechanisms

Inevitably, introducing users' historical answers from conversations on different topics will bring a lot of noise. And it is difficult to extract useful information from user's historical answer by simple average like **Formula (7)** and **Formula (9)**. To alleviate the noise issue and make full use of the user's historical answers, we design sets of denoising mechanisms, including a History Selection process to filter out the historical answers that are less relevant to the conversation, a Similarity Re-weighting process to re-weight the similarity matrix between historical answers and conversation, and an Edge Re-weighting process to re-weight the label edge weights.

History Selection To filter out the irrelevant history from a large number of user's historical answers, we propose a coarse-grained History Selection process. In this process, we retrieve all the historical answers of the users enrolled in the conversation from the training set, and group the answers according to DA labels. For the i-th historical answers in the j-th DA label (denoted as U_i^j), we convert each word in U_i^j into pre-trained word embeddings and apply max-pooling on the word-level to generate utterance-level representation $u_i^j \in \mathbb{R}^d$:

$$u_i^j = \max_{1<l<L_u} U_i^j, \tag{11}$$

where d is the dimension of word embeddings, L_u is the length of the user's historical answer. And then, for current conversation C, we convert each word in C into pre-trained word embeddings and apply max-pooling on the word-level and utterance-level to generate conversation-level representation $c \in \mathbb{R}^d$:

$$c = \max_{1<t<N, 1<l<L_c} C, \tag{12}$$

where L_c is the length of the utterance in the conversation and N is the number of utterances in the conversation. And then we calculate the cosine similarity between user's historical answer U_i^j and convsersation C and select top-K historical answers (denoted as $H(j)$) most relevant to current conversation:

$$\text{similarity}(U_i^j, C) = \frac{u_i^j \cdot c}{||u_i^j||||c||}. \tag{13}$$

Similarity Re-weighting The History Selection is a coarse-grained selecting process, which cannot accurately measure the relevance between the user's historical answers and the current conversation. Therefore, we design a Similarity Re-weighting process to re-weight the similarity matrix. Considering a conversation with N utterances $C = \{h_1, h_2, ..., h_N\}$ and user's historical answers for the j-th DA label with K utterances $H(j) = \{\hat{h}_1^j, \hat{h}_2^j, ..., \hat{h}_K^j\}$, we calculate a similarity matrix $M \in \mathbb{R}^{N \times K}$:

$$M_{tk} = f(h_t, \hat{h}_k^j), \tag{14}$$

where f is a similarity function, and the function we choose here is:

$$f(x, y) = x^T W_s y, \tag{15}$$

where W_s is a parameter to be learned. And then, we use the initial score of DA labels $\hat{y} \in \mathbb{R}^{N \times S}$ as **Formula (6)** learned from the 1st-phase to re-weight the similarity matrix. Firstly, we transpose $\hat{y}$ and select the j-th row of the $\hat{y}^T$ as the attention of each utterance under the j-th DA label. Next, we use the attention $\hat{y}_j' \in \mathbb{R}^{N \times 1}$ to re-weight the similarity matrix M as $M' \in \mathbb{R}^{N \times K}$and we apply a column-wise max-pooling to get the weight $m' \in \mathbb{R}^K$. Finally, we sum the user's historical answers by weight and get the representation of label node $e_j \in \mathbb{R}^{2d_h}$, which replaces the original **Formula (7)**:

$$\hat{y}_j' = \hat{y}^T[j], \tag{16}$$

$$M' = \hat{y}_j' M, \tag{17}$$

$$m' = \max_{1<t<N} M', \tag{18}$$

$$e_j = \sum_{\hat{h}_k^j \in H(j)} m_k' \hat{h}_k^j, \tag{19}$$

where $[j]$ represents selecting the j-th row.

Edge Re-weighting Each utterance in a conversation is related to only a few DA labels and not to most others, so irrelevant label nodes will bring the noise to the utterance. The main idea of this process is to re-weight the edges between the utterance node and the label node, so that the weight of the irrelevant label nodes is lower. We use prediction result $\hat{p} \in \mathbb{R}^{N \times S}$ as **Formula (5)** learned from the 1st-phase to re-weight the label edge weights as:

$$\begin{aligned} \alpha_{i,j}^l &= \text{softmax}(\hat{p}_i) = \text{softmax}([\hat{p}_{i,1}, ..., \hat{p}_{i,S}]), \\ i &= 1, ..., N, j = 1, ..., S \end{aligned} \tag{20}$$

which replaces the original **Formula (9)**.

3.4 Dialogue Act Recognition

The local utterance feature vector u_t (from utterance encoder), contextually encoded feature vector h_t (from utterance context encoder) and g_t (from Heterogeneous User History graph) are concatenated and fed into a fully-connected network to obtain the final prediction results:

$$y_t = \text{sigmoid}(W_\alpha[u_t, h_t, g_t] + b_\alpha), \tag{21}$$

where W_α is also used in the 1st-phase, we reload the W_α from the 1st-phase for further training. To make the dimensions consistent, we set the g_t as zeros in the 1st-phase. Our Dialogue Act Recognition is a multi-label classification problem, so we use *sigmoid* as activation and use *Binary Cross-Entropy (BCE)* as loss function.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the performance of our model on two benchmark datasets used in several prior studies for the DAR task. MSDialog dataset (Qu et al., 2018; Yang et al., 2018) is a labeled dialog dataset of question answering (QA) interactions between information seekers and providers from an online forum on Microsoft products. The dataset contains more than 2,000 multi-turn QA dialogs with 10,020 utterances that are annotated with a subset of 12 user intent on the utterance level. The ICSI Meeting Recorder Dialogue Act Corpus (Janin et al., 2003; Shriberg et al., 2004; Ang et al., 2005) (MRDA) contains 72 hours of naturally occurring multi-party meetings that were first converted into 75 word level conversations. The original MRDA tag set had 11 general tags and 39 specific tags. Following previous work (Qu et al., 2019) on multi-label classification, we adopt label-based accuracy (i.e., Hamming score) and micro-F1 score as our main evaluation metrics.

4.2 Implementation Details

In our experiments, we split training/validation/testing datasets following Yu et al. (2019) for MSDialog and Lee and Dernoncourt (2016) for MRDA. For two datasets, we first strip punctuation, and then we convert the characters into lower-case and tokenize the texts with NLTK [1]. Pre-trained GloVE embeddings of 100 dimensions are adopted as word-level embeddings. Out-of-vocabulary words are set by randomly sampling values from the standard normal distribution. The max length of utterance is set to 800 in MSDialog and 80 in MRDA. All the hyper-parameters have been optimized on the validation set using accuracy. For CNN, we use filters of size 3, 4 and 5 with 200 feature maps in each dataset. The hidden size of LSTM and GCN is set to 400 in MSDialog and 200 in MRDA. We use Adam optimizer for optimization with learning rate 1e-3. The hyperparameter K is set to 10 in MSDialog and 90 in MRDA. Note that MRDA conversations are much longer compared to MSDialogue (1000 vs 10), we split the MRDA conversations into smaller parts containing a maximum of 90 utterances.

4.3 Baselines

For a comprehensive evaluation of our proposed model HUH, we compare our model with the following baseline methods: 1) **HEC** (Kumar et al., 2018) builds a hierarchical BiLSTM-CRF model for DAR, which learns representations at multiple levels. 2) **CNN-CR** (Qu et al., 2019) designs a CNN model that incorporates context information with a window size of 3. 3) **CASA** (Raheja and Tetreault, 2019) proposes a context-aware self-attention mechanism coupled with a hierarchical recurrent neural network for DAR. 4) **GA-Seq** (Colombo et al., 2020) leverages a sequence to sequence approach to improving the modeling of tag sequentiality. 5) **CRNN** (Yu et al., 2019) is an adapted Convolutional Recurrent Neural Network (CRNN) that models the interactions between utterances of long-range context. 6) **DialogueGCN** (Ghosal et al., 2019) proposes a graph-based model that leverages self and inter-speaker interaction of the interlocutors to model conversational context for emotion recognition. 7) **BERT** (Devlin et al., 2019) is a pre-trained language model which has been applied to many NLU applications. We encode utterance through BERT alongwith a feedforward network for classification.

[1] http://www.nltk.org/

Methods	MSDialog		MRDA	
	Acc	F1	Acc	F1
HEC (Kumar et al., 2018)	63.8	69.5	67.3	67.5
CNN-CR (Qu et al., 2019)	63.5†	70.3†	67.6	67.4
CASA (Raheja and Tetreault, 2019)	66.4	72.1	67.9	68.2
GA-Seq (Colombo et al., 2020)	66.4	72.4	68.5	68.9
CRNN (Yu et al., 2019)	68.2†	73.4†	68.4	68.5
DialogueGCN (Ghosal et al., 2019)	68.5	73.9	68.5	68.9
BERT (Devlin et al., 2019)	70.1	75.8	70.0	70.5
HUH	69.6	75.3	70.1	70.7
HUH + BERT	**71.8**	**77.2**	**70.9**	**71.1**
w/o graph	68.2	73.4	68.8	68.9
w/o user's history	68.5	73.7	69.1	69.7
w/o denoise	68.4	73.5	69.2	69.9

Table 1: Performance on MSDialog and MRDA. Acc is a shorthand for Accuracy. † indicates that the results are reported from the CRNN.

4.4 Experimental Results

Table 1 shows the results for all models. The results show that our HUH model outperforms all the compared baselines and achieves the best performance on both datasets. The experiments reveal some interesting points: 1) Graph-based models (DialogueGCN and HUH) perform better than sequence-based methods such as CASA and CRNN. Since the graph-based models consider interaction among speakers in the conversations, they can explicitly model inter-speaker dependency and recognize long-range dialogue acts better. 2) The major difference between our proposed model and the strong baseline DialogueGCN lies in three aspects: First, we integrate the user's historical answers into our model, rather than just focusing on the utterances in the conversation. Second, we divide the recognition process into two organically combined phases and use the initial score of DA labels to guide the Heterogeneous User History graph-based prediction. Third, we employ sets of denoising mechanisms to filter out the irrelevant content from the user's historical answers. These three modifications improve the performance significantly. 3) Compared to the large pretrained model BERT with 110M parameters, our model achieves comparable performance while the model size is much smaller (20M). What's more, by replacing the utterance encoder in our model with BERT and fine-tuning on the target datasets, a further improvement is gained and our model achieve the state-of-the-art performance.

4.5 Ablation Study

To analyze the effectiveness of different factors of our HUH, we report the ablation test in terms of: 1) **w/o graph**: We use the initial score of DA labels from the 1st-phase as the final result. 2) **w/o user's history** we discard the label nodes in our proposed Heterogeneous User History graph. 3) **w/o denoise** We replace the History Selection process with random selection and discard the Similarity Re-weighting process and Edge Re-weighting process. The results are listed in Table 1. Compared our proposed HUH with **w/o user's history**, we can conclude that adding the user's historical answers can improve the performance of the DAR task. However, such historical answers that contain irrelevant information might sometimes play as a negative role, which makes **w/o denoise** less performing than HUH. Our model outperforms **w/o graph** by a great margin, demonstrating the contribution of the graph structure and user's history in our model.

4.6 Quantitative Analysis

Table 2 shows the f1 score of HUH, DialogueGCN and CRNN for DA labels on the MSDialog dataset. We can observe that our model achieves the best performance on all DA labels. Besides, the three models achieve satisfying performance on *OQ*, *PA* and *GG*, but not on *RQ*, *NF* and *JK*. We analyze the main reason for the poor performance of the latter is the vague expression of sentence patterns. We observe that HUH and DialogueGCN perform much better than CRNN for the Repeat Question (RQ) label, this mainly because our model and DialogueGCN utilize the graph-structured model to capture

DA	Taxonomy	HUH	DialogueGCN	CRNN
OQ	Original Question	**96.9**	96.9	96.9
PA	Potential Answer	**92.1**	90.6	87.7
GG	Greetings/Gratitude	**83.0**	82.4	82.5
FD	Further Details	**64.2**	63.4	64.1
IR	Information Request	**62.1**	61.3	61.9
RQ	Repeat Question	**60.8**	52.1	28.1
NF	Negative Feedback	**41.9**	21.8	28.8
JK	Junk	**41.4**	35.7	34.5

Table 2: The f1 score of HUH, DialogueGCN and CRNN for DA labels on MSDialog dataset.

Utterance	Golden	HUH	DialogueGCN	CRNN
U1: My webcam does not work with Skype on my PC.	OQ	OQ	OQ	OQ
U2: Please, run the DirectX diagnostics tool and save the results to a file.	PA	PA	PA	PA
U3: Hi, I can not access camera for skype for bussiness, please attach my screen shot.	RQ	RQ	RQ	PA
U4: How I can suggest a street to Bing Maps? Thank you.	GG OQ	GG OQ	GG OQ	GG OQ
U5: Believe me, I still can not delete my street to Bing. Bing, spend years and nothing changes!	FD NF	FD NF	FD RQ	FD

Table 3: Some examples from MSDialog. The errors are in yellow.

the interaction between users, and the RQ label needs to consider the long-range context from different users in the conversation. Furthermore, we observe that our model outperforms the other methods on NF by a noticeable margin. A possible reason is that the negative feedback is an unobvious intention, which makes the model difficult to recognize from scratch. By introducing the user's historical answers as clues, our model can extract the hidden feature in the utterance more easily and improve the performance on the labels that are less evident.

To further analyze the performance of different models, some examples from MSDialog are demonstrated in Table 3. We find that a model needs to find the semantic relevance between "webcam work with skype" and "access camera for skype" to recognize the *Repeat Question (RQ)* label in the *U3*. While previous methods of sequentially modeling conversations (such as CRNN) lack the ability to capture long-range interaction between utterance. In addition, we can observe that the previous methods cannot recognize the Negative Feedback (NF) contained in *U5*. This might due to the reason that the negative feedback expressed by "spend years and nothing changes" is unobvious. Compared with DialogueGCN and CRNN, our method can recognize the NF label accurately.

5 Conclusion

In this paper, we focus on the task of multiple DA recognition in a multi-party conversation. We propose a Heterogeneous User History (HUH) graph convolution network model to learn utterance, conversation and user's historical answers. To handle the noise caused by introducing the user's historical answers, we design sets of denoising mechanisms, including a History Selection process, a Similarity Re-weighting process and an Edge Re-weighting process. We evaluate the proposed method on two benchmark datasets MSDialog and MRDA. The experimental results verify the effectiveness of integrating user's historical answers, and show that our proposed model outperforms the state-of-the-art methods.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No. 61773229), Shenzhen Giiso Information Technology Co. Ltd., the Basic Research Fund of Shenzhen City (Grand No. JCYJ20190813165003837), Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202032), Overseas Cooperation Research Fund of Graduate School at Shenzhen, Tsinghua University (Grant No. HW2018002), and the Shenzhen General Research Project (No. JCYJ20190808182805919).

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005*, pages 1061–1064.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 745–754.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 225–234.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 7594–7601.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019*, pages 4171–4186.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 154–164.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 928–939.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5010–5016.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003*, pages 364–367.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342.

Simon Keizer, Rieks op den Akker, and Anton Nijholt. 2002. Dialogue act recognition with bayesian networks for dutch dialogues. In *Proceedings of the SIGDIAL 2002 Workshop, The 3rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 88–94.

Hamed Khanpour, Nishitha Guntakandla, and Rodney D. Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 2012–2021.

Minkyoung Kim and Harksoo Kim. 2018. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. *Pattern Recognit. Lett.*, 101:1–5.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1746–1751.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 3440–3447.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019*, pages 383–392.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543.

Libo Qin, Wanxiang Che, Yangming Li, Minheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 8665–8672.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 989–992.

Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019*, pages 25–33.

Vipul Raheja and Joel R. Tetreault. 2019. 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, NAACL-HLT 2019*, pages 3727–3733.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018*, pages 593–607.

Elizabeth Shriberg, Rajdip Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the SIGDIAL 2004 Workshop, The 5th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 97–100.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden markov models. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing*.

Jiahui Wen, Jingwei Ma, Yiliu Feng, and Mingyang Zhong. 2018. Hybrid attentive answer selection in CQA with deep users modelling. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 2556–2563.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 245–254.

Yue Yu, Siyao Peng, and Grace Hui Yang. 2019. Modeling long-range context for concurrent dialogue acts recognition. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pages 2277–2280.