CEM: Machine-Human Chatting Handoff via Causal-Enhance Module

Shanshan Zhong[1], **Jinghui Qin**[1], **Zhongzhan Huang**[1], **Daifeng Li**[2*]
[1] School of Computer Science and Engineering, Sun Yat-sen University
[2] School of Information Management, Sun Yat-sen University
{zhongshsh5,qinjingh,huangzhzh23}@mail2.sysu.edu.cn,
lidaifeng@mail.sysu.edu.cn

Abstract

Aiming to ensure chatbot quality by predicting chatbot failure and enabling human-agent collaboration, Machine-Human Chatting Handoff (MHCH) has attracted lots of attention from both industry and academia in recent years. However, most existing methods mainly focus on the dialogue context or assist with global satisfaction prediction based on multi-task learning, which ignore the grounded relationships among the causal variables, like the user state and labor cost. These variables are significantly associated with handoff decisions, resulting in prediction bias and cost increase. Therefore, we propose Causal-Enhance Module (CEM) by establishing the causal graph of MHCH based on these two variables, which is a simple yet effective module and can be easy to plug into the existing MHCH methods. For the impact of users, we use the user state to correct the prediction bias according to the causal relationship of multi-task. For the labor cost, we train an auxiliary cost simulator to calculate unbiased labor cost through counterfactual learning so that a model becomes cost-aware. Extensive experiments conducted on four real-world benchmarks demonstrate the effectiveness of CEM in generally improving the performance of existing MHCH methods without any elaborated model crafting.

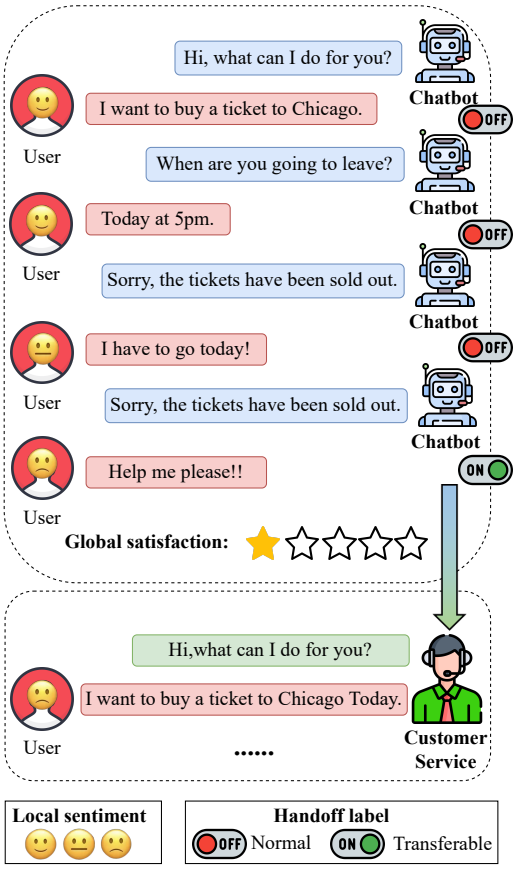

Figure 1: An example of MHCH. Handoff label includes two types "normal" & "transferable", which denotes whether the chatbot should be transferred to human service.

1 Introduction

In recent years, with the rapid development of deep learning (He et al., 2016; Ren et al., 2015), more and more service-oriented organizations have deployed chatbots to alleviate the problem of limited service resources. Although these chatbots can respond in real-time and save labor cost, they suffer from inappropriate responses and invalid conversations due to the limited quantity of available high-quality training data and the inherent biases (Xu et al., 2019; Liang et al., 2022) of neural networks. Moreover, human utterances sometimes are elusive since they are rich in acronyms, slang words, and even content without logic or grammar, which are too obscure for a chatbot to comprehend. To alleviate these drawbacks, researchers introduced a human-agent collaboration mechanism named Machine-Human Chatting Handoff (MHCH) to allow a human to take over the dialogue while a robot agent feels confused so that a dialogue can be continued to avoid a bad user experience and reduce the risk of customer churn (Liu et al., 2021a,b). As

*Corresponding author: lidaifeng@mail.sysu.edu.cn

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3242–3253
December 7-11, 2022

shown in Fig.1, when a chatbot tries to address the user's needs by giving an inappropriate response, the user will feel disappointed and give a low global satisfaction score for the current dialogue, which means a service failure and may lead to customer loss. If deploying with the MHCH mechanism, a human can take over the dialogue and give a satisfactory response to meet the user's needs, thus ensuring the user experience and service quality (Radziwill and Benton, 2017).

A high-quality MHCH service should consider multiple factors, such as dialogue context, local sentiments, global satisfaction, user state, labor cost, etc. However, most existing MHCH methods are mainly concerned with the dialogue context (Liu et al., 2021a) or assisting with global satisfaction prediction under the multi-task learning setting (Liu et al., 2021b), ignoring the grounded relationships among the other causal variables of MHCH, like the user state and human cost.

To address above issues and improve the performance of MHCH, we propose a general Causal-Enhance Module (CEM), which can be plugged into existing MHCH networks (Liu et al., 2021a,b), to incorporate the considerations of other potential causal variables of MHCH. Specifically, we first analyzes MHCH task based on causal graph by mining all potential causal variables and deduce that user states and labor cost are the other two causal variables that should be considered for high-quality customer service. Then, to incorporate the consideration of user state, we train a user state network mainly driven by local sentiments to maintain the changes of user state during the dialogue and adjust the handoff predictions by correcting the prediction bias according to the causal relationship between user states and handoff decisions. To consider the labor cost of customer service and reduce it as much as possible while maintaining the same service quality, we construct a counterfactual-based cost simulator to regress the cost of a dialogue as an auxiliary task which can make the MHCH backbone become cost-aware and minimize the labor cost as much as possible.

The contributions of our CEM can be summarized as follows:

- We conduct causal analysis based on causal graph for MHCH and identify the other two causal variables: user state and human cost, which should be considered to build high-quality MHCH service.
- To consider the impact of user state, the user state is applied to correct the handoff prediction bias according to the causal relationship between user states and handoff decisions.
- To minimize the labor cost of customer service while maintaining the same service quality, we construct a counterfactual-based cost simulator to regress the cost of a dialogue as an auxiliary task, which can make the MHCH backbone become cost-aware.

We release our code to help other researchers to reproduce the results of CEM [1].

2 Related Work

Machine-Human Chatting Handoff. The research on MHCH is originated in 2018. Using the idea of reinforcement learning, Huang et al. (2018) proposed a dialogue robot to choose an assistant. Rajendran et al. (2019) utilize a reinforcement learning framework to maximize success rate and minimize human workload. Liu et al. (2021a,b) regraded the MHCH as a classification problem and focused on identifying which sentence should be transferred to the human service.

Causal inference and counterfactual learning. For structural causal models (Halpern et al., 2005), related studies (Heskes, 2013; Claassen et al., 2014; Xia et al., 2021) utilize graph neural networks for directed acyclic graph structure learning. For Rubin causal models, Rubin (2006) and Bengio et al. (2019) use neural networks to approximate the propensity scores, matching weights, etc., which can satisfy the covariate balancing (Kallus, 2020; Kuang et al., 2017); The representation learning (Huang et al., 2020b; Liang et al., 2020) can also be used to matched the covariate balance between the test group and the reference group (Shalit et al., 2017; Louizos et al., 2017; Lu et al., 2020). Several studies (Yoon et al., 2018; Yuan et al., 2019; Liu et al., 2020) uses counterfactual methods based on the generative models over the observed distributions for causal inference.

Multi-task learning in dialogue systems. Xu et al. (2020) uses multi-task learning for auxiliary pre-training tasks of dialogue data. Qin et al. (2020) combines dialogue behavior recognition and sentiment classification. Ide and Kawahara (2021) proposes a model which includes generation and classification tasks.

[1] https://github.com/Qrange-group/CEM

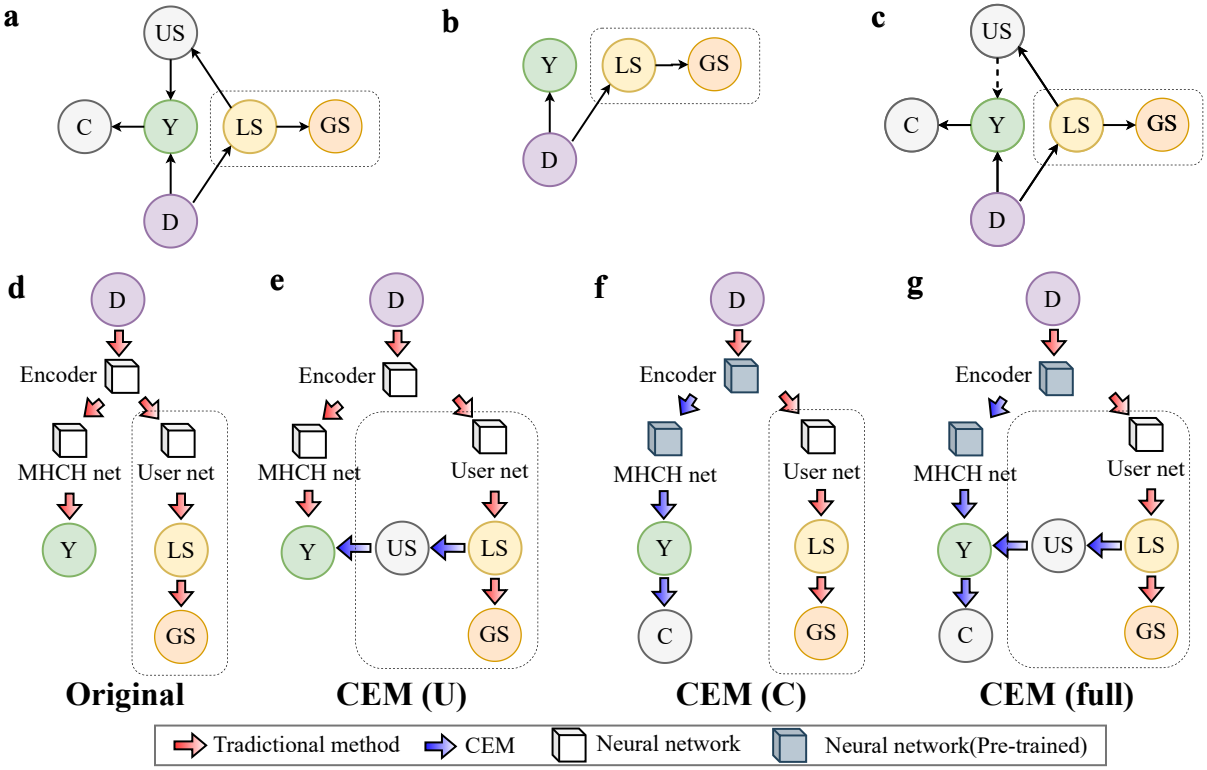

Figure 2: Causal graphs and model structures. D: Dialog, Y: prediction of MHCH, LS: local sentiment, GS: global satisfaction, US: user state, C: labor cost. The solid lines represent causality, the dashed line is adjustment, and the dotted line outlines the part that MHCH classification models doesn't have. **a, b, c** are the causal graphs of MHCH, traditional multi-task methods and CEM based on multi-task learning, respectively. **d** is the original model structure based on **b** (Song et al., 2019; Liu et al., 2021a,b). **e**, **f** and **g** are the model structures enhanced by CEM (U), CEM (C) and CEM (full).

3 Preliminary

A given dialogue $D = [u_1, u_2, \ldots, u_L]$ contains L utterances and have a label sequence $Y^h = [y_1^h, \ldots, y_L^h]$, where y_t^h is the handoff label of u_t,$1 \leq t \leq L$. The handoff labels Γ have two kinds of labels, i.e.,"normal" and "transferable", where "normal" means that the utterance is no need to transfer, and "transferable" means that the utterance needs to be transferred to the manual service. The dialogue D also have a global satisfaction label {"satisfactory", "neutral", "dissatisfied"}. Then, the local sentiment of each utterance u_t is measured by an open-source tool SnowNLP, which includes three labels {"positive", "neutral", "negative"} .

4 Methodology

In this section, we analyse the impact of variables on MHCH from a fundamental view of causality. Then we present our CEM framework that eliminates the bad effect of ignored causal variables.

4.1 Causal analysis of MHCH

Causal graph is a directed acyclic graph where a node denotes a variable and an edge denotes a causal relation between two nodes (Pearl, 2009). It is widely used to describe the process of data, which can guide the design of predictive models (Zhang et al., 2021). Fig.2(a) shows the causal graph of MHCH. The rationality of this causal graph is explained as follows:

- D denote the dialogue $D = [u_1, \ldots, u_L]$.
- Y = $[p_1, p_2, ..., p_L]$ is the prediction of MHCH, where $p_t, 1 \leq t \leq L$ is the probability of that the handoff label of u_t is "transferable".
- LS is the local sentiments of each utterance in a dialogue.
- GS represents the user's subjective evaluation of the current dialogue.
- US is a state for a given dialogue. Unlike GS, it is a variable that describes the objective

state of the user. We can model US through local sentiments.

- C is the labor cost caused by the wrong prediction of MHCH.
- Edge $D \rightarrow Y$: The MHCH can judge when to transfer to manual service according to the dialogue content. Therefore, the dialogue can affect the prediction of MHCH.
- Edge $Y \rightarrow C$: The labor cost depends on the prediction of MHCH. If we do not need to transfer to human service, there will not be labor cost.
- Edges $D \rightarrow LS \rightarrow GS$: The dialogue quality of chatbot will affect users' sentiment, and then affect users' evaluation of the services.
- Edge $LS \rightarrow US$: The user state can be modeled from local sentiments.
- Edge $US \rightarrow Y$: In Fig.1, the user state can affect MHCH to judge whether the service should be transferred to a manual service.

However, instead of Fig.2(a), the existing common solutions, which are mainly based on multi-task methods, e.g., service satisfaction analysis (SSA) (Song et al., 2019), adopt the causal graph as Fig.2(b), which models the relationship of $D \rightarrow LS \rightarrow GS$. Specifically, they consider two neural networks for a multi-task of SSA and MHCH, i.e., train a user network (UN) for SSA and a MHCH network for MHCH as shown in Fig.2(d). Since there is an encoder network that share weights between the two tasks to integrate information, the local sentiment can assist MHCH network by sharing dialogue features. Although such modeling is simple and has good performance on MHCH tasks, it is established through a simplified causal graph without considering the factors of user and cost, so it can not completely show the overall picture as shown in Fig.2(a). Therefore, we design a new causal graph as seen in Fig.2(c) to consider further factors, e.g. user state and labor cost to bridge the MHCH network and UN. Based on the new causal graph, a novel CEM (full) model (Fig.2(g)) as well as its variants CEM (U) and CEM (C) will be introduced in the following sections.

4.2 User State

As shown in Fig.2(e), we can use local sentiment, which is the output of UN, to restore user state. Since user state has a strong correlation with relative time (Ding and Li, 2005), we can measure the user state of u_t by Eq.(1) with local sentiment.

$$US_t = \beta_t \text{UN}(D), \tag{1}$$

where $\text{UN}(D) \in R^{L\times 3}$ is the local sentiment from UN when given D as input. And the weight $\beta_t \in R^L$ is

$$\beta_t = softmax([1, \ldots, t-1, t, 0, \ldots, 0]), \tag{2}$$

Soft Adjustment. In Fig.2(a), if we establish the causal relationship between user state and MHCH task directly, D will become a confounder to Y due to the intervention of user state. To solve this problem, a simple way is to ignore the causal relationship $US \rightarrow Y$.

However, the utterance sometime can not affect directly whether it is necessary to transfer to human service since the complexity of the language. And the user state restored by the local sentiment can help decision of MHCH. For example,in Fig.1, the information of the sentiment is more important than those of utterance. Therefore, we can use another strategy to use user state for adjustment the decision of MHCH network. In particular, we can mask the neutral local sentiment since the neutral sentiments are confusing and therefore not highly recognizable, which means that neutral sentiment's impact on MHCH Song et al. (2019) task is lower.

Moreover, the dimension of US is three, which does not match the two-dimension output Y of the MHCH network. While masking the neutral sentiments, we can propose de-neutral soft adjustment shown in Fig.2(e), whose specific operation is as follows:

$$y_D^h = softmax(\text{Mask}_n(US_D) \odot \text{MHCH}(D)), \tag{3}$$

where Mask_n is the masking operator for neutral sentiment.MHCH is the model using to modeling the causal relationship of $D \rightarrow Y$. y_D^h is the predicted result of D. $\odot$ represents a product operation at the element level, which makes the probability of "positive" times the probability of "normal" and makes the probability of "negative" times the probability of "transferable". This adjustment can modify the normal probability with positive sentiment and the transferable probability with negative sentiment.

Table 1: Comparison of classification performance (%) in Clothing and Makeup1. Bold indicates the best result, underlined indicates the second best result. CEM-DAMI (C) refers to the model without the part in dashed box in Fig.2(f). We don't use full CEM to enhance DAMI for the reason that CEM cannot restore US from pure MHCH models.

Models	Clothing					Makeup1				
	F1	Mac.F1	GT-I	GT-II	GT-III	F1	Mac.F1	GT-I	GT-II	GT-III
HRN (Lin et al., 2015)	57.3	73.5	62.3	71.8	76.5	58.2	74.1	62.3	72.5	78.1
HAN (Yang et al., 2016)	58.2	74.1	62.9	72.2	76.7	60.1	75.3	65.4	74.9	79.9
BERT (Devlin et al., 2018)	56.0	72.9	59.3	68.1	73.1	57.0	73.3	61.5	71.0	76.5
CRF-ASN (Chen et al., 2018)	57.6	73.4	61.5	72.6	78.0	56.8	73.6	63.7	74.2	79.8
HBLSTM-CRF (Kumar et al., 2018)	59.0	74.4	63.6	73.7	78.8	60.1	75.4	67.0	76.3	81.2
DialogueRNN (Majumder et al., 2019)	59.0	74.3	63.1	73.8	79.0	61.3	76.1	66.3	76.0	81.2
CASA (Raheja and Tetreault, 2019)	59.7	74.7	64.8	74.9	79.7	60.4	75.7	67.8	77.0	81.8
LSTMLCA (Dai et al., 2020)	61.8	76.1	66.4	76.3	81.1	62.1	76.6	67.8	76.9	81.7
CESTa (Wang et al., 2020)	60.5	75.2	64.0	74.6	79.6	60.2	75.2	65.2	75.9	81.5
DAMI (Liu et al., 2021a)	67.3	**79.7**	**70.3**	**79.1**	**83.9**	67.1	79.5	67.8	76.9	82.1
CEM-DAMI (C)	**67.5**	**79.7**	69.7	77.6	81.5	**67.5**	**79.9**	**70.4**	**78.1**	**82.2**

4.3 Cost Simulator

Using US can adjust MHCH models, eliminate the bias caused by not considering user factors, and obtain Y that is closer to the ground truth. Based on this, we conduct counterfactual modeling of labor cost. Labor cost can be divided into two types, one is effective cost and the other is invalid cost. The effective cost refers to incurring the cost for the utterances that must be transferred to human, and the invalid cost refers to incurring the cost for the utterances that do not need to transfer to human, i.e., the cost caused by the wrong prediction. Note that we can not change the effective cost, and we can only reduce invalid cost. The network which models the causal relationship of $D \rightarrow Y \rightarrow C$ is named as cost simulator. We first define the cost simulator as follows:

$$\begin{cases} Y^h \sim P_{Y^h}(Y^h|D) \\ C = F_c(Y^h, D), \end{cases} \tag{4}$$

where Y^h represents the ground truth of the MHCH task and C denotes the labor cost of D. P_{Y^h} is the probability calculation function for MHCH. F_c represents the cost calculation function. P_{Y^h} can be defined as follows:

$$P_{Y^h}(Y^h|D) = \prod_{t=1}^{L} P(\hat{y}_t^h = y_t^h|u_t), \tag{5}$$

where $\hat{y}_t^h$ is the prediction of MHCH network for u_t, and y_t^h is the ground truth of MHCH network for u_t. Then let ζ represent the upper limit of the cost of one utterance in human service. Since we only need to estimate the relative cost, which means that it does not need to obtain a specific and accurate estimated value. Therefore, ζ is set to 1 by default. F_c can be defined as follows:

$$F_c(Y^h, D) = \sum_{t=1}^{L} \zeta \cdot P(\hat{y}_t^h = 1|u_t). \tag{6}$$

In datasets, if the label corresponding to the data u_t is "transferable", $P\left(y_t^h = 1 \middle| u_t\right) = 1$, otherwise $P\left(y_t^h = 1 \middle| u_t\right) = 0$. Next, we pretrain the cost simulator based on Eq.(10), so that the predicted cost measured by the output of the MHCH network is close to the real labor cost.

$$\mathcal{L}_{c_{pre}} = MSE\left(\hat{C} - C\right), \tag{7}$$

where $\hat{C}$ represents the cost predicted by the simulator, and C represents the ground truth of cost. After supervised pre-training, the cost simulator can become cost-aware and can give a counterfactual cost.

On the trained cost simulator, we begin to train the MHCH model, and calculate the counterfactual cost of each D through F_c. Since we want to make labor cost as low as possible, the loss function $\mathcal{L}_c$ of the counterfactual cost simulator is defined as:

$$\begin{aligned} \mathcal{L}_c &= \sum_{i=1}^{|\Psi|} \hat{C} \\ &= \frac{1}{L} \sum_{i=1}^{|\Psi|} \sum_{t=1}^{L} \zeta \cdot P(y_{i,t}^h = 1|u_{i,t}), \end{aligned} \tag{8}$$

where $|\Psi|$ is the dataset size. Overall, the total loss $\mathcal{L}(\Theta)$ of CEM for the multi-task MHCH is:

$$\mathcal{L}(\Theta) = \mathcal{L}_h + \eta_s \cdot \mathcal{L}_s + \eta_c \cdot \mathcal{L}_c + \delta||\Theta||_2^2, \tag{9}$$

Table 2: Overall statistics of the datasets.

Statistics items	Clothing	Makeup1	Clothes	Makeup2
# (Dialogues)	3500	4000	10000	3540
# (Dissatisfied dialogues)	-	-	2302	1180
# (Neutral dialogues)	-	-	6399	1180
# (Satisfactory dialogues)	-	-	1299	1180
# (Transferable utterances)	6713	7446	16921	7668
# (Normal utterances)	28901	32488	237891	86778
Avg # (Utterances per dialogues)	10.18	9.98	25.48	26.68

where L_h and L_s are the loss function of MHCH task and SSA task, which are based on previous studies (Liu et al., 2021b). η_c and $\eta_s \in R^+$ are the weight parameters for multi tasks. the ℓ_2 regularization $\delta||\Theta||_2^2$ is used to mitigate model overfitting.

5 Experiments

5.1 Dataset and Experimental Settings

We evaluate our approach on four datasets including Clothing(Liu et al., 2021a), Makeup1(Liu et al., 2021a), Clothes(Liu et al., 2021b), and Makeup2(Liu et al., 2021b). The statistics of the data are shown in Table 2. To verify the effectiveness of CEM and fairly compare with the baselines on the same datasets, we evaluate the performance of our CEM on the classification model DAMI (Liu et al., 2021a) by using Clothing and Makeup1 while testing the performance of CEM on the multi-task model RSSN (Liu et al., 2021b) with Clothes and Makeup2, for the reason that these models are state-of-the-art. Because DAMI only models the relationship of $D \rightarrow Y$, not $D \rightarrow LS \rightarrow GS$, it is not possible to use DAMI to get US, so we only add cost adjustment on DAMI named as CEM-DAMI (C).

5.2 Evaluation Metrics

Following prior works (Liu et al., 2021a,b), we adopt F1, Macro F1 (Mac.F1) and GT-T (Golden Transfer within Tolerance) as accuracy metrics for evaluating the MHCH task. GT-T takes into account the tolerance property of the MHCH task through a tolerance range T, which allows for "biased" predictions. The T can be ranged from 1 to 3 corresponding to GT-I, GT-II, and GT-III.

Furthermore, to verify that CEM can effectively control the cost, we compare the labor cost of different models. It is obvious that the higher the accuracy, the lower the labor cost. To eliminate the impact of model accuracy, we only compare the invalid cost which is more meaningful than the full cost. Therefore, we compute the invalid cost as follows:

$$\text{IC} = \frac{\sum_{i=1}^{|\Psi|}\sum_{t=1}^{L}(\hat{y}_{i,t}^h != y_{i,t}^h,\ \hat{y}_{i,t}^h == \text{"transferable"})}{\sum_{i=1}^{|\Psi|}\sum_{t=1}^{L}(\hat{y}_{i,t}^h != y_{i,t}^h)} \quad (10)$$

where IC is the abbreviation of invalid cost.

5.3 Implementation Details

We use TensorFlow[2] to implement our method with one RTX2080 GPU card. Back-propagation is used to compute gradients and the Adam optimizer (Kingma and Ba, 2014) is used for parameter updates. The dimension of word embedding is set as 200. The total vocabulary size of datasets is 48.5K. Other trainable model parameters are initialized by sampling values from initializer. Hyperparameters of CEM and baselines are tuned on the validation set. η_s is set as 0.3. The sizes of model units are based on the baselines setting and remain the same in the comparison experiments. The L_2 regularization weight is 10^{-4} in DAMI and 3×10^{-5} in RSSN. The batch size is set as 32. The number of epochs is set as 30 in DAMI and CEM-DAMI(C), and set as 80 in RSSN, CEM-RSSN(C), CEM-RSSN(U) and CEM-RSSN. Finally, we train the models with a learning rate of 7.5×10^{-3} in DAMI and 1.5×10^{-3} in RSSN. Following the data processing setting (Liu et al., 2021a), the datasets are divided into training sets, validation sets, and test sets with a ratio of 8:1:1.

5.4 Results on Clothing and Makeup1

The experimental results of models on Clothing and Makeup1 are shown in Table 1. In Clothing, CEM-DAMI outperforms most baselines, and is slightly weaker than DAMI on GT-T metrics. In Makeup1, CEM-DAMI is the best performing model. This experimental result shows that incorporating cost into models does not reduce model accuracy.

IC of DAMI with the adjustment of CEM is significantly reduced in both Clothing and Makeup1 as shown in Table 4. We can conclude based on Table 1 and Table 4 that CEM-DAMI(C) can achieve competitive performance with lower labor cost, which means that our cost simulator can reduce labor cost while maintaining the model performance.

5.5 Results on Clothes and Makeup2

The experimental results of the methods on Clothes and Makeup2 are shown in Table 3. The cost results are shown in Figure 3. From Table 3, we can

[2]https://www.tensorflow.org/

Table 3: Comparison of classification performance (%) in Clothes and Makeup2. Bold indicates the best result, underlined indicates the second best result. CEM-RSSN (U) refers to the model in Fig.2(e), CEM-RSSN (C) refers to the model in Fig.2(f), and CEM-RSSN (full) refers to the model in Fig.2(g).

Models	Clothes					Makeup2				
	F1	Mac.F1	GT-I	GT-II	GT-III	F1	Mac.F1	GT-I	GT-II	GT-III
HAN (Yang et al., 2016)	59.8	78.7	71.7	73.1	74.0	54.3	75.4	68.5	70.1	71.3
BERT+LSTM (Devlin et al., 2018)	60.4	78.9	73.4	74.9	75.9	42.2	84.2	72.9	66.4	77.6
HEC (Kumar et al., 2018)	59.8	78.7	71.2	72.3	73.0	57.1	76.8	68.0	69.5	70.5
DialogueRNN (Majumder et al., 2019)	60.8	79.2	73.1	74.6	75.6	58.3	77.4	68.8	70.5	71.6
CASA (Raheja and Tetreault, 2019)	62.0	79.8	73.6	75.0	75.9	58.4	77.5	70.6	72.7	73.9
LSTMLCA (Dai et al., 2020)	62.6	80.1	72.4	73.9	74.8	57.4	77	70.2	71.7	72.6
CESTa (Wang et al., 2020)	60.6	79.1	73.4	74.8	75.6	59.3	78.0	69.6	71.2	72.2
DAMI (Liu et al., 2021a)	66.7	82.2	74.2	75.9	77.1	61.1	79.0	73.3	74.4	75.2
MT-ES (Ma et al., 2018)	61.7	79.7	74.6	75.9	76.8	57.1	76.9	69.9	71.7	72.8
JointBiLSTM (Bodigutla et al., 2020)	62.0	79.9	75.0	76.1	76.9	59.3	78.0	70.1	72.0	73.1
DCR-Net (Qin et al., 2020)	62.1	79.9	71.4	72.8	73.7	58.8	77.7	70.0	72.1	73.4
RSSN (Liu et al., 2021b)	67.1	82.5	75.8	77.1	78	63.5	80.2	74.9	76.5	77.7
CEM-RSSN (U)	66.6	82.4	**79.8**	80.5	81.0	**67.6**	**82.8**	**80.2**	**81.1**	81.8
CEM-RSSN (C)	66.1	82.1	78.6	79.6	80.2	65.2	81.6	77.3	78.2	78.9
CEM-RSSN (full)	**67.9**	**82.9**	79.6	**80.7**	**81.4**	64.8	80.9	79.4	**81.1**	**82.3**

Table 4: Comparison of IC (%) in Clothing and Makeup1.

Models	Clothing	Makeup1
DAMI	59.0	56.6
CEM-DAMI (C)	53.4	54.2

observe that our CEM can effectively improve the performance of RSSN, especially on GT-I, GT-II, and GT-III.

5.6 Performance Analysis

Ablation Test. We investigate the effects of user state and cost simulator through ablation experiments. According to the experimental results shown in Table 3, CEM-RSSN (U) can effectively improve model performance by modeling and tracking user state which is highly-correlated with user's tolerances for invalid responses. Besides, similar with the results on Clothing and Makeup1, CEM-RSSN (C) still can achieve competitive performance even the cost simulator is not designed to improve the accuracy of the model.

Analysis of Generality. In order to evaluate the generality of the proposed CEM, we conduct additional experiments for HAN and BERT with CEM(C) on Clothes and Makeup2. Compared to HAN, the average improvements of CEM-HAN(C) are 2.03%, 0.83%, and 4.15% in terms of F1, Mac. F1 and IC respectively; Compared to BERT, the average improvements of CEM-BERT(C) are 2.11%, 1.32%, and 5.74% respectively in terms of F1, Mac. F1 and IC respectively. Experimental results further illustrate the proposed CEM can be easily plugged into different models and get performance improvements.

Analysis of IC metrics. The total error predictions of RSSN and CEM-RSSN(full) are 937 and 900 on Clothes respectively, the error predictions corresponding to IC among the total are 358 and 326, and the error predictions corresponding to (1-IC) are 579 and 574. Similarly, the total error predictions, error predictions for IC, and (1-IC) of RSSN are 487, 221, and 266 on Makeup2. These values of CEM-RSSN are 466, 205, and 261. From these results, we can conclude that although IC declines, the error predictions corresponding to (1-IC) will not be larger since our CEM can also reduce total error predictions. This indicates that the error predictions corresponding to both IC and (1-IC) are all decreasing, but the decrease of IC is larger. Therefore, the increase of (1-IC) is a relative value compared to the decrease of IC, which indicates that the service quality of chatbots can be guaranteed, and declining IC will not impact user experience.

Analysis of Cost. We also compare the cost of RSSN and CEM-RSSN through multiple experiments, and perform two sided t-test to verify the whether the significant difference between RSSN and CEM-RSSN over metrics. The results shown in Fig.3 mean that CEM can significantly reduce labor cost while improving model performance.

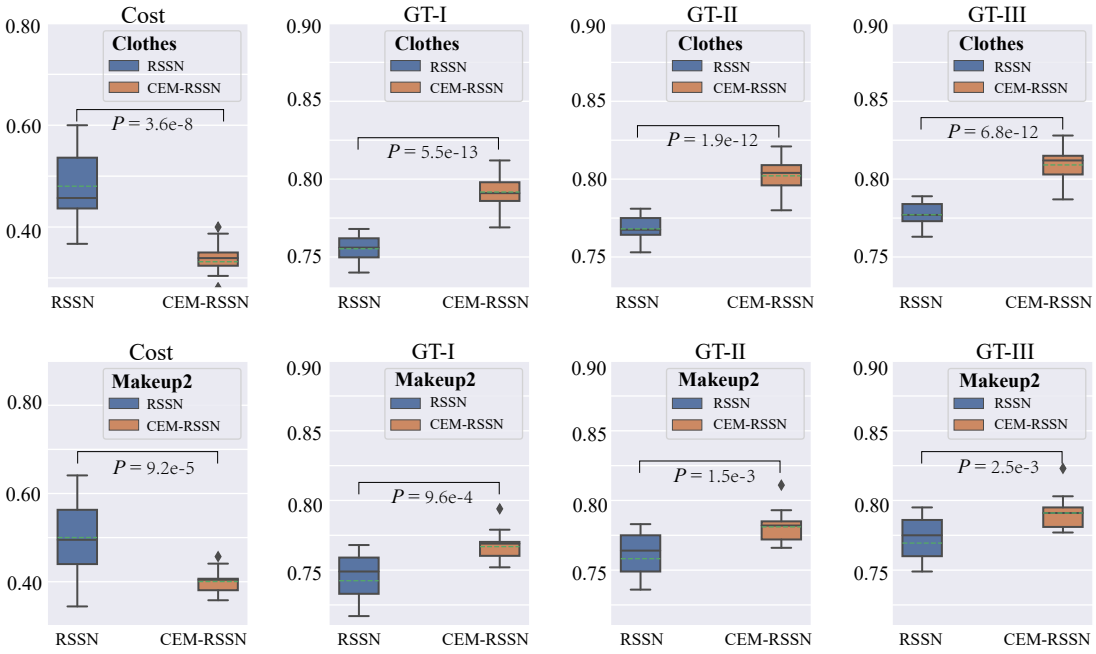

Figure 3: Comparison of labor cost (%) and GT-T in Clothes and Makeup2. Cost is IC defined in Eq.(10). We obtain multiple sets of results of RSSN and CEM-RSSN through multiple experiments, draw the box plots of Cost and GT-T, and perform two sided t-test to verify the whether the significant difference between RSSN and CEM-RSSN over four metrics.

Running Efficiency. On one RTX2080, we check the training speed and inference effectiveness for RSSN and CEM-RSSN(full) under the same setting on 10 runs: For training speed, they need 38.86s (standard deviation=1.14s) and 39.2s (standard deviation=1.80s), respectively; For inference effectiveness, they need 2.38s (standard deviation=0.63s) and 2.50s (standard deviation=0.50s). Although CEM takes more time than RSSN, the additional computational consumption is not large.

Case Analysis. To help evaluate the results of the experiments, studies were conducted using representative cases from the datasets. For example, in some dialogues in which users always make talk in a neutral or positive tone, the model can pinpoint locations that require manual switching based on semantic understanding. While other baselines, such as RSSN, tend to perform the manual switching operation several steps ahead of the optimal switching point. Such as the 28th, 29th, and 30th utterances of dialogue 23 on Clothes dataset; the 23rd, 24th, and 25th utterances of dialogue 1 on Makeup2 dataset. The reasons may be that due to the complexity of language expressions, RSSN may confuse some semantic information reflected in user utterances and make an incorrect decision in some cases. The proposed CEM model can solve the problem by incorporating confounding factors and a soft adjustment mechanism. In addition, the counterfactual framework can help the model find appropriate reference cases in the historical dialogues, thus further reducing the error rate.

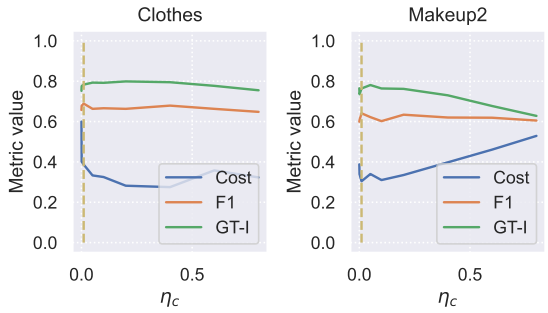

Figure 4: The impact of the cost simulator over different $\eta_c \in [0, 1]$. This experiment is about CEM-RSSN on Clothes and Makeup2 dataset. Cost is IC which is defined in Eq.(10). F1 and GT-I is the metrics about the MHCH accuracy.

5.7 Parameter Sensitivity

We compare different η_c in Eq.9 to explore the impact of the cost simulator on MHCH models. We take the value of η_c from 0 to 1, and the experimental results about CEM-RSSN on Clothes and Makeup2 are shown in Fig.4. It can be seen that

the different η_c has little effect on the green (GT-I) and orange lines (F1), which indicates the stability of CEM in cost control.

Comparing the metric values under different η_c on both two datasets, it can be found that when the weight exceeds a certain threshold, the larger the weight, the higher the labor cost and the lower the accuracy, indicating the labor cost is associated with the accuracy of models. Besides, it also shows that the trade-off between the accuracy and labor cost can be achieved by adjusting the loss weight, so as to obtain a model with low cost and high accuracy. Meanwhile, when the η_c is lower than a certain threshold, the effect of the cost simulator can be ignored, which will affect the performance of CEM. Finally, we chose 0.01 as the value of η_c to make a better trade-off in our experiments.

6 Conclusions

In this paper, we propose a novel CEM module for the MHCH task where we use causal inference to enhance the models of the MHCH task and take into account the labor cost. And the empirical results on four datasets and two types of models indicate that CEM improves model accuracy consistently and effectively saves invalid labor cost. Since there are minor modifications to the model architecture and loss function on the existing MHCH method and achieve significant improvement, CEM can be easily plugged into the different MHCH methods.

7 Future works

Based on CEM, we can consider the following future work:
(1) When CEM is used to enhance the existing MHCH method, there is no additional parameters. From the perspective of inference speed and model deployment, this is the advantage. However, the consideration of user state and labor cost in CEM is mainly based on intuition to construct explicit transformations, which can not ensure good enough performance. Therefore, we can consider adding neural networks to CEM in the future.
(2) For MHCH task, conventional neural networks are generally used, such as fully connected neural networks, LSTM, BiLSTM, etc. This makes the structure of the model lack careful consideration, which has the potential to greatly improve the performance of MHCH. Specifically, we should consider a lot of fine-tuning methods for the neural network to ensure their performance, including the use of structure search techniques (He et al., 2021; Liu et al., 2018; Huang et al., 2020a), elaborate modules (Hu et al., 2018; Huang et al., 2020b), specific parameters (Liang et al., 2020), etc.
(3) Compared with other artificial intelligence fields, such as image segmentation and image classification, the data volume shown in Table.2 is not very large. However, the quality and quantity of data have a huge impact on their training, so the data-driven cost simulator may have a bias in its estimation of labor cost. Therefore, we can consider data augment methods(Cubuk et al., 2018; Lin et al., 2021) to effectively improve model training.

Limitations

Although we have fully demonstrated the effectiveness of CEM experimentally, we ignore the analysis of CEM from the mathematical point of view of causal inference. This makes it impossible for us to guarantee that CEM can be used in more complex and sophisticated MHCH methods in the future or other applications in more extensive fields. Moreover, since the cost simulator is trained by neural networks, we can not ensure whether the cost given by the simulator can not have a well enough performance to reflect the true labor cost.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This research was supported by National Natural Science Foundation of China (NSFC) under Grant No.72074231 and Grant No.62206314, GuangDong Basic and Applied Basic Research Foundation under Grant No.2022A1515011835, Key Project of Guangdong "Pandeng" program in China under grant pdjh2021a0001.

References

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

Praveen Kumar Bodigutla, Aditya Tiwari, Josep Valls Vargas, Lazaros Polymenakos, and Spyros Matsoukas. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. *arXiv preprint arXiv:2010.02495*.

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*, pages 225–234.

T. Claassen, J. M. Mooij, and T. Heskes. 2014. Proof supplement - learning sparse causal models is not np-hard (uai2013). *Statistics*.

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. 2018. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501.

Zhigang Dai, Jinhua Fu, Qile Zhu, Hengbin Cui, Yuan Qi, et al. 2020. Local contextual attention with hierarchical structure for dialogue act recognition. *arXiv preprint arXiv:2003.06044*.

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

Yi Ding and Xue Li. 2005. Time weight collaborative filtering. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 485–492.

Halpern, Joseph, Y., Pearl, and Judea. 2005. Causes and explanations: A structural-model approach. part i: Causes. *British Journal for the Philosophy of Science*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei He, Zhongzhan Huang, Mingfu Liang, Senwei Liang, and Haizhao Yang. 2021. Blending pruning criteria for convolutional neural networks. In *Artificial Neural Networks and Machine Learning - ICANN 2021 - 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14-17, 2021, Proceedings, Part IV*, volume 12894 of *Lecture Notes in Computer Science*, pages 3–15. Springer.

T. Heskes. 2013. Bayesian probabilities for constraint-based causal discovery.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Ting Hao 'Kenneth' Huang, Joseph Chee Chang, and Jeffrey P. Bigham. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. *ACM*.

Zhongzhan Huang, Senwei Liang, Mingfu Liang, Wei He, and Haizhao Yang. 2020a. Efficient attention network: Accelerate attention by searching where to plug. *arXiv preprint arXiv:2011.14058*.

Zhongzhan Huang, Senwei Liang, Mingfu Liang, and Haizhao Yang. 2020b. Dianet: Dense-and-implicit attention network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4206–4214.

T. Ide and D. Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation.

Nathan Kallus. 2020. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 265–274.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Senwei Liang, Zhongzhan Huang, Mingfu Liang, and Haizhao Yang. 2020. Instance enhancement batch normalization: An adaptive regulator of batch noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4819–4827.

SENWEI Liang, Zhongzhan Huang, and Hong Zhang. 2022. Stiffness-aware neural network for learning hamiltonian systems. In *International Conference on Learning Representations*.

Junfan Lin, Zhongzhan Huang, Keze Wang, Xiaodan Liang, Weiwei Chen, and Liang Lin. 2021. Continuous transition: Improving sample efficiency for continuous control problems via mixup. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9497. IEEE.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 899–907.

Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. 2018. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34.

Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A

general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 831–840.

Jiawei Liu, Zhe Gao, Yangyang Kang, Zhuoren Jiang, Guoxiu He, Changlong Sun, Xiaozhong Liu, and Wei Lu. 2021a. Time to transfer: Predicting and evaluating machine-human chatting handoff. In *Proc. of AAAI*, pages 5841–5849.

Jiawei Liu, Kaisong Song, Yangyang Kang, Guoxiu He, Zhuoren Jiang, Changlong Sun, Wei Lu, and Xiaozhong Liu. 2021b. A role-selected sharing network for joint machine-human chatting handoff and service satisfaction analysis. *arXiv preprint arXiv:2109.08412*.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, and Lawrence Carin. 2020. Reconsidering generative objectives for counterfactual reasoning. *Advances in Neural Information Processing Systems*, 33:21539–21553.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*, pages 585–593.

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

Judea Pearl. 2009. *Causality*. Cambridge university press.

L. Qin, W. Che, Y. Li, M. Ni, and T. Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):8665–8672.

Nicole Radziwill and Morgan Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional*, 19(3):25–35.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Janarthanan Rajendran, Jatin Ganhotra, and Lazaros C Polymenakos. 2019. Learning end-to-end goal-oriented dialog with maximal user task success and minimal human agent use. *Transactions of the Association for Computational Linguistics*, 7:375–386.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

D. B. Rubin. 2006. *Matched Sampling for Causal Effects*. Matched sampling for causal effects /.

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Kaisong Song, Lidong Bing, Wei Gao, Jun Lin, Lujun Zhao, Jiancheng Wang, Changlong Sun, Xiaozhong Liu, and Qiong Zhang. 2019. Using customer service dialogues for satisfaction analysis with context-assisted multiple instance learning.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. 2021. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34.

R. Xu, C. Tao, D Jiang, X. Zhao, D Zhao, and R. Yan. 2020. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving ad click prediction by considering non-displayed events. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 329–338.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong

Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–20.

A The detail of baselines

We compare our proposed approach with the following state-of-the-art dialogue classification models and multi-task models, which mainly come from MHCH, SSA and other similar tasks. We briefly categorize these baselines and introduce them below.

Baselines for the MHCH task. HRN (Lin et al., 2015): It uses a bidirectional LSTM to encode utterances and then fed these utterance features into a standard LSTM for context representation. **HAN** (Yang et al., 2016): HAN is a hierarchical network with two levels of attention mechanisms on word-level and utterance-level. **BERT** (Devlin et al., 2018): It uses a pre-trained BERT model to construct the single utterance representations for classification. **HEC** (Kumar et al., 2018): It builds a hierarchical recurrent neural network using bidirectional LSTM as a base unit and the conditional random field (CRF) as the top layer to classify each utterance into its corresponding dialogue act. **CRF-ASN** (Chen et al., 2018): It extends the structured attention network to the linear-chain conditional random field layer, which takes both contextual utterances and corresponding dialogue acts into account. **HBLSTM-CRF** (Kumar et al., 2018): It is a hierarchical recurrent neural network using bidirectional LSTM as a base unit and two projection layers to combine utterances and contextual information. **DialogueRNN** (Majumder et al., 2019): It is a method based on RNNs that keeps track of the individual party states throughout the conversation and uses the information for emotion classification. **CASA** (Raheja and Tetreault, 2019): It leverages the effectiveness of a context-awar self-attention mechanism to capture utterance level semantic text representations on prior hierarchical recurrent neural network. **LSTMLCA** (Dai et al., 2020): It is a hierarchical model based on the revised self-attention to capture intra-sentence and inter-sentence information. **CESTa** (Wang et al., 2020): It employs LSTM and Transformer to encode context and leverages a CRF layer to learn the emotional consistency in the conversation. **DAMI** (Liu et al., 2021a): It utilizes difficulty-assisted encoding to enhance the representations of utterances, and a matching inference mechanism is introduced to capture the contextual matching features.

Multi-task baselines. MT-ES (Ma et al., 2018): It proposes a joint framework that unifies the two highly pertinent tasks. **JointBiLSTM** (Bodigutla et al., 2020): It minimizes an adaptive multi-task loss function in order to jointly predict turn-level Response Quality labels provided by experts and explicit dialogue-level ratings provided by end users. **DCR-Net** (Qin et al., 2020): It considers the cross-impact and model the interaction between the two tasks by introducing a co-interactive relation layer. **RSSN** (Liu et al., 2021b): It integrates both dialogue satisfaction estimation and handoff prediction in one multi-task learning framework.