A Role-Selected Sharing Network for Joint Machine-Human Chatting Handoff and Service Satisfaction Analysis

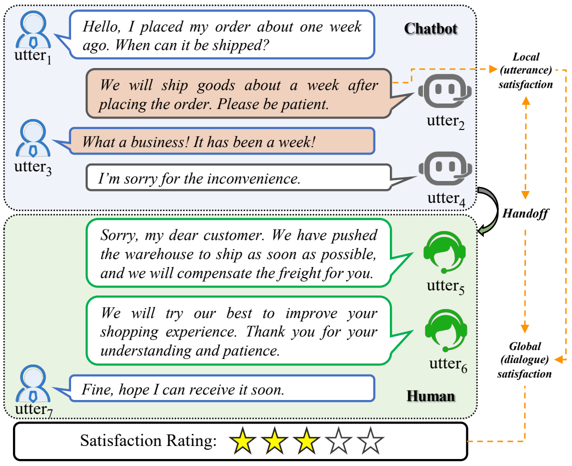
Jiawei Liu[1], Kaisong Song[2,3], Yangyang Kang[3], Guoxiu He[4], Zhuoren Jiang[5], Changlong Sun[3,5], Wei Lu[1*], Xiaozhong Liu[6*]

[1]Wuhan University, Wuhan, China [2]Northeastern University, Shenyang, China
[3]Alibaba Group, Hangzhou, China [4]East China Normal University, Shanghai, China
[5]Zhejiang University, Hangzhou, China [6]Worcester Polytechnic Institute, Worcester, USA

{laujames2017,weilu}@whu.edu.cn, {kaisong.sks,yangyang.kangyy}@alibaba-inc.com,
gxhe@fem.ecnu.edu.cn, jiangzhuoren@zju.edu.cn,
changlong.scl@taobao.com, xliu14@wpi.edu

Abstract

Chatbot is increasingly thriving in different domains, however, because of unexpected discourse complexity and training data sparseness, its potential distrust hatches vital apprehension. Recently, Machine-Human Chatting Handoff (MHCH), predicting chatbot failure and enabling human-algorithm collaboration to enhance chatbot quality, has attracted increasing attention from industry and academia. In this study, we propose a novel model, Role-Selected Sharing Network (RSSN), which integrates both dialogue satisfaction estimation and handoff prediction in one multi-task learning framework. Unlike prior efforts in dialog mining, by utilizing local user satisfaction as a bridge, global satisfaction detector and handoff predictor can effectively exchange critical information. Specifically, we decouple the relation and interaction between the two tasks by the role information after the shared encoder. Extensive experiments on two public datasets demonstrate the effectiveness of our model.



utter1: Hello, I placed my order about one week ago. When can it be shipped?

utter2: We will ship goods about a week after placing the order. Please be patient.

utter3: What a business! It has been a week!

utter4: I'm sorry for the inconvenience.

utter5: Sorry, my dear customer. We have pushed the warehouse to ship as soon as possible, and we will compensate the freight for you.

utter6: We will try our best to improve your shopping experience. Thank you for your understanding and patience.

utter7: Fine, hope I can receive it soon.

Satisfaction Rating: ★★★☆☆

Figure 1: A snippet of a moderately satisfied customer service dialogue. There is a satisfaction rating at the end of the conversation. The utterance with an orange background color denotes a transferable utterance.

1 Introduction

Chatbot, as one of the recent palpable AI excitements, has been widely adopted to reduce the cost of customer service (Qiu et al., 2017; Ram et al., 2018; Zhou et al., 2020). However, due to the complexity of human conversation, auto-chatbot can hardly meet all users' needs, while its potential failure perceives skepticism. AI-enabled customer service, for instance, may trigger unexpected business losses because of chatbot failures (Radziwill and Benton, 2017; Rajendran et al., 2019). Moreover, for chatbot adoption in sensitive areas, such as healthcare (Chung and Park, 2019) and criminal justice (Wang et al., 2020a), any subtle statistical miscalculation may trigger serious health and legal consequences. To address this problem, recently, scholars proposed new dialog mining tasks to auto-assess dialogue satisfaction, a.k.a. Service Satisfaction Analysis (SSA) at dialogue-level (Song et al., 2019), and to predict potential chatbot failure via machine-human chatting handoff (MHCH) at utterance-level (Huang et al., 2018; Liu et al., 2021). In a MHCH context, algorithm can transfer an ongoing auto-dialogue to the human agent when the current utterance is confusing.

Figure 1 depicts an exemplar dialogue of online customer service. In this dialogue, the chatbot gives an unsatisfied answer about shipping, thus causing the customer's complaint (local dissatisfaction $utter_2$ and $utter_3$). Ideally, chatbot should be able to detect the negative (local) emotion ($utter_3$) and tries to appease complaints, but this problem remains unresolved. If chatbot continues, the customer may cancel the deal and give a negative rating (dialogue global dissatisfaction). With MHCH

[*] Corresponding authors.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9731–9741
November 7–11, 2021. ©2021 Association for Computational Linguistics

(detects the risks of $utter_2$ and $utter_3$), the dialogue can be transferred to the human agent, who is better at handling, compensating, and comforting the customer and enhance customer satisfaction. This example illustrates the cross-impact between handoff and dialogue (local+global) satisfaction. Intuitively, MHCH and SSA tasks can be compatible and complementary given a dialogue discourse, i.e., the local satisfaction is related to the quality of the conversation (Bodigutla et al., 2019a, 2020), which can support the handoff judgment and ultimately affect the overall satisfaction. On the one hand, handoff labels of utterances are highly pertinent to local satisfaction, e.g., one can utilize single handoff information to enhance local satisfaction prediction, which ultimately contributes to the overall satisfaction estimation. On the other hand, the overall satisfaction is obtained by combining local satisfactions, which reflects the quality in terms of answer generation, language understanding, and emotion perception, and subsequently helps to facilitate handoff judgment.

In recent years, researchers (Bodigutla et al., 2019a,b; Ultes, 2019; Bodigutla et al., 2020) explore joint evaluation of turn and dialogue level qualities in spoken dialogue systems. In terms of general dialogue system, to improve the efficiency of dialogue management, Qin et al. (2020) propose a co-interactive relation layer to explicitly examine the cross-impact and model the interaction between sentiment classification and dialog act recognition, which are relevant tasks at the same level (utterance-level). However, MHCH (utterance-level) and SSA (dialogue-level) target satisfaction at different levels. More importantly, handoff labels of utterances are more comprehensive and pertinent to local satisfaction than sentiment polarities. Meanwhile, customer utterances have significant impacts on the overall satisfaction (Song et al., 2019), which motivates us that the role information can be critical for knowledge transfer of these two tasks.

To address the aforementioned issues, we propose an innovative Role-Selected Sharing Network (RSSN) for handoff prediction and dialogue satisfaction estimation, which utilizes role information to selectively characterize complex relations and interactions between two tasks. To the best of our knowledge, it is the pioneer investigation to leverage the multi-task learning approach for integrating MHCH and SSA. In practice, we first adopt a shared encoder to obtain the shared representations of utterances. Inspired by the co-attention mechanism (Xiong et al., 2016; Qin et al., 2020), the shared representations are then fed into the role-selected sharing module, which consists of two directional interactions: *MHCH to SSA* and *SSA to MHCH*. This module is used to get the fusion of MHCH and SSA representations. We propose the role-selected sharing module based on the hypothesis that the role information can benefit the tasks' performances. The satisfaction distributions of utterances from different roles (agent and customer) are different, and the effects for the tasks are also different. Specifically, the satisfaction of agent is non-negative. The utterances from agent can enrich the context of customer's utterances and indirectly affect satisfaction polarity. Thus, directly employing local satisfaction of agent into the interaction with handoff may introduce noise. In the proposed role-selected sharing module, we adopt local satisfaction based on the role information: only the local satisfaction from customer can be adopted to interact with handoff information. By this means, we can control knowledge transfer for both tasks and make our framework more explainable. The final integrated outputs are then fed to separate decoders for handoff and satisfaction predictions.

To summarize, our contributions are mainly as follows: (1) We introduce a novel multi-task learning framework for combining machine-human chatting handoff and service satisfaction analysis. (2) We propose a Role-Selected Sharing Network for handoff prediction and satisfaction rating estimation, which can utilize different role information to control knowledge transfer for both tasks and enhance model performance and explainability. (3) The experimental results demonstrate that our model outperforms a series of baselines that consists of the state-of-the-art (SOTA) models on each task and multi-task learning models for both tasks. To assist other scholars in reproducing the experiment outcomes, we release the codes and the annotated dataset[1].

2 Related Work

Due to the complexity of human conversation, current automatic chatbots are not mature enough and still fail to meet users' expectations (Brandtzaeg and Følstad, 2018; Jain et al., 2018; Chaves and Gerosa, 2020). Besides exploring novel dialogue models, dialogue quality estimation, service satis-

[1] https://github.com/WeijiaLau/RSSN

faction analysis, and human intervention are vital strategies to enhance chatbot performance.

Dialogue Quality and Service Satisfaction Analysis. Interaction Quality (IQ) (Schmitt et al., 2012) and Response Quality (RQ) (Bodigutla et al., 2019b) are dialogue quality evaluation metrics for spoken dialogue systems. Automated models to estimate IQ (Ultes et al., 2014; El Asri et al., 2014) and RQ (Bodigutla et al., 2019a,b, 2020) utilize various features derived from the dialogue content and output from spoken language understanding components. For chat-oriented dialogue system, Higashinaka et al. (2015a,b) introduce Dialogue Breakdown Detection task to detect a system's inappropriate utterances that lead to dialogue breakdowns. To efficiently analyze dialogue satisfaction, Song et al. (2019) introduce the task of service satisfaction analysis (SSA) based on multi-turn customer service dialogues. The proposed CAMIL model can predict the sentiment of all the customer utterances and aggregate those sentiments into overall service satisfaction polarity. Nevertheless, the sentiment of customer utterance is only one of the factors that influence service satisfaction.

Machine Human Chatting Handoff. Another perspective of further enhancing the chatbot's performance is to combine chatbots with human agent. Recently, there are several works about human-machine cooperation for chatbots. Huang et al. (2018) propose the crowd-powered conversational assist architecture, namely Evorus, which integrates crowds with multiple chatbots and a voting system. Rajendran et al. (2019) utilize reinforce learning framework to transfer conversations to human agents once encountered new user behaviors. Different from them, Liu et al. (2021) mainly focus on detecting transferable utterances which are one of the keys to improve user satisfaction. They propose a DAMI network that utilizes difficulty-assisted encoding and matching inference mechanisms to predict the transferable utterance.

Multi-task learning in dialogue system. For satisfaction estimation, Bodigutla et al. (2020) propose to jointly predict turn-level RQ labels and dialogue-level ratings. They utilize features from spoken dialogue system and BiLSTM (Hochreiter and Schmidhuber, 1997) based model to automatically weight each turn's contribution towards the rating. Ma et al. (2018) propose a joint framework that unifies two highly pertinent tasks. Both tasks are trained jointly using weight sharing to extract the common and task-invariant features while each task can still learn its task-specific features. To learn the correlation between two tasks, Qin et al. (2020) propose a DCR-Net. It adopts a stacked co-interactive relation layer to incorporate mutual knowledge explicitly. This model ignores the contextual information and isolated two types of information when performing interaction.

3 Methodology

Figure 2 shows the overall architecture of RSSN, which consists of three parts: Shared Utterance and Matching Encoder, Role-Selected Interaction Layer, and Decoder for MHCH and SSA. In this section, we will describe them in detail.

Given a dialogue $D = [u_1, ..., u_L]$, it consists of a sequence of L utterances with corresponding handoff labels $[y_1^h, ..., y_L^h]$, where $t = \{1 \leq t \leq L | t \in \mathbb{N}\}$, $y_t^h \in \Psi$ and $\Psi = \{$*normal, transferable*$\}$. *Transferable* indicates the dialogue should be transferred to the human agent, whereas *normal* indicates there is no need to transfer. The satisfaction polarity of dialogue D is noted as y^s, where $y^s \in \Omega$ and $\Omega = \{$*well satisfied, met, unsatisfied*$\}$. Note that we perform the multi-task learning with the supervision of handoff labels and dialogue's satisfaction only. The local satisfaction distributions of utterances are only the latent estimation, which helps to predict the dialogue's satisfaction.

3.1 Shared Utterance and Matching Encoder

The shared encoder consists of a bidirectional LSTM (BiLSTM) to learn the utterance representation and a masked matching layer to capture the contextual matching information.

Suppose $u_t = [w_1, ..., w_{|u_t|}]$ represents a sequence of words in the t-th utterance. These words are mapped into corresponding word embeddings $\boldsymbol{E}_{u_t} \in \mathbb{R}^{n \times |u_t|}$, where n is the word embedding dimension. By adopting semantic composition models with word embeddings, we can learn the utterance representation. In this work, we adopt a BiLSTM model and concatenate hidden states of forward and backward LSTM to learn the context-sensitive utterance representation $\boldsymbol{v}_t \in \mathbb{R}^{2k}$, where k is the number of hidden units of LSTM cell. Formally, we have $\boldsymbol{v}_t = \text{BiLSTM}(\boldsymbol{E}_{u_t})$.

In a dialogue, preceding utterances for each utterance provide helpful context information to estimate local satisfaction. Thus, within a dialogue, there is a high probability of inter-dependency with

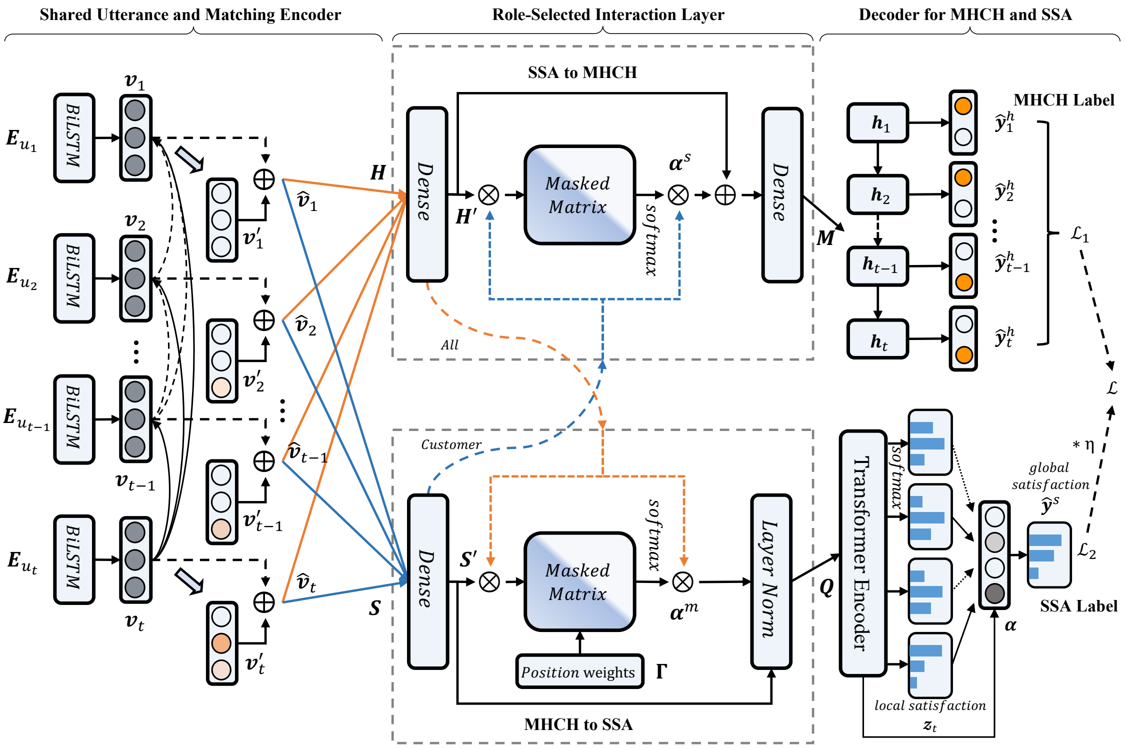


Figure 2: The architecture of our Role-Selected Sharing Network (RSSN) Network.

respect to their context clues. To encapsulate the contextual matching and information flow in the dialogue, we feed the utterance representation into a unidirectional matching mechanism:

$$v'_t = v_t^\top [v_1, v_2, ..., v_{t-1}] \qquad (1)$$

After masking out the future information of the present utterance, the matching features of dialogue D is a lower triangular matrix with the diagonal values removed. Then we concatenate the matching features with utterance representation to get $\hat{v}_t = [v'_t; v_t]$. Finally, we obtain the initial shared utterances representations of MHCH $\boldsymbol{H} = [\hat{v}_1, ..., \hat{v}_L]$ and SSA $\boldsymbol{S} = [\hat{v}_1, ..., \hat{v}_L]$.

3.2 Role-Selected Interaction Layer

In customer service dialogue, the roles of different participants would exhibit different characteristics (Song et al., 2019). Besides, we conjecture that MHCH and SSA have different impacts on each other. These two tasks indirectly establish a connection through various factors such as dialogue quality, satisfaction, and sentiment. At the same time, role information also plays an important role in both tasks. On the one hand, the utterances from agent can enrich the context of customer utterances and indirectly affect satisfaction polarity. In contrast, customer utterances tend to have a more direct impact on the dominating satisfaction polarity. On the other hand, the utterances of any participants can trigger machine-human chatting handoff. Thus, we propose the Role-Selected Interaction Layer, which contains two interaction directions: *SSA to MHCH* and *MHCH to SSA*, to model the relations and interactions between the two tasks separately.

We first apply two Dense layers over the handoff information and satisfaction information respectively to make them more task-specific, which can be noted as $\boldsymbol{H}' = \text{Dense}(\boldsymbol{H})$ and $\boldsymbol{S}' = \text{Dense}(\boldsymbol{S})$, where $\boldsymbol{H}' \in \mathbb{R}^{L\times d}$ and $\boldsymbol{S}' \in \mathbb{R}^{L\times d}$. Note that d is the number of hidden units of the Dense layer.

SSA to MHCH. Co-attention is an effective and widely used method to capture the mutual knowledge among the correlated tasks (Xiong et al., 2016; Qin et al., 2020). Inspired by the basic co-attention mechanism, we design the interaction mechanism separately according to the characteristics of tasks. In this way, task-relevant knowledge can be transferred mutually between two tasks. Specifically, the *SSA to MHCH* module produces comprehensive handoff representations incorporating the local satisfaction information. Since the agent utterances indirectly affect satisfaction polarity, directly employing local satisfaction of agent into the interaction with handoff may introduce noise. As a

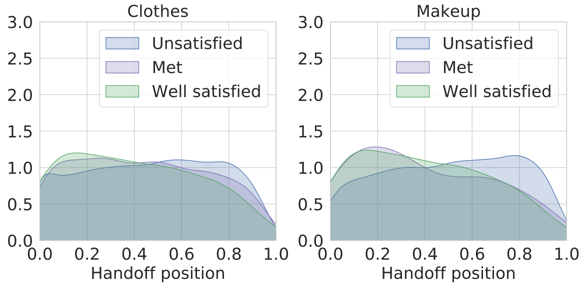


Figure 3: The relative handoff position distributions in three different service satisfaction ratings.

consequence, we only adopt the local satisfaction information of customer to interact with handoff information. The process can be defined as follows:

$$\boldsymbol{\alpha}^s = \text{softmax}(\text{Mask}_c(\boldsymbol{H}'(\boldsymbol{S}')^\top)) \quad (2)$$

$$\boldsymbol{M} = \text{Dense}([\boldsymbol{\alpha}^s\boldsymbol{S}'; \boldsymbol{H}']) \quad (3)$$

where $\boldsymbol{M} \in \mathbb{R}^{L\times d}$ and Mask_c denotes that we mask out (setting to $-\infty$) all values of the future information and agent utterances.

MHCH to SSA. As shown in Figure 3, we observe that the dialogue satisfaction rating is related to the handoff position. Intuitively, a handoff can be triggered by the local unsatisfied attitude of the customer, and the later handoff means users are unsatisfied before the end of the conversation. Prior study, Song et al. (2019), also found that user satisfaction at the dialogue level is usually determined by the attitudes of the last few utterances. We can derive that handoff at the later period of the conversation may result in a lower satisfaction rating. Thus, we adjust the interactive attention by positional weights, which can be computed as below:

$$\boldsymbol{\beta}_t = \text{softmax}([\frac{1}{L}, ..., \frac{t}{L}, ..., 1] \odot \boldsymbol{I}_p(u_t)) \quad (4)$$

where $\odot$ is element-wise product and $I_p(\cdot)$ denotes a zero masking identity matrix to mask out future information. Finally, the positional weights $\boldsymbol{\Gamma} = [\boldsymbol{\beta}_1; ...; \boldsymbol{\beta}_L]$, where $\boldsymbol{\Gamma} \in \mathbb{R}^{L\times L}$. The mechanism gives more weight to the later handoff information. We apply the positional weights to the interaction:

$$\boldsymbol{\alpha}^m = \text{softmax}(\text{Mask}(\boldsymbol{S}' \cdot (\boldsymbol{H}')^\top \cdot \boldsymbol{\Gamma})) \quad (5)$$

$$\boldsymbol{Q} = \text{LayerNorm}(\boldsymbol{\alpha}^m \cdot \boldsymbol{H}' + \boldsymbol{S}') \quad (6)$$

where Mask denotes that we mask out the future information (setting to $-\infty$), and LayerNorm denotes the layer normalization (Ba et al., 2016).

3.3 Decoder for MHCH and SSA

After the role-selected interaction layer, we can get the outputs $\boldsymbol{M} = [\boldsymbol{m}_1, ..., \boldsymbol{m}_L]$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, ..., \boldsymbol{q}_L]$. Then we adopt separate decoders to predict handoff and satisfaction rating.

In terms of machine-human chatting handoff, the tendency of handoff also depends on the dialogue context. Thus, we feed the outputs of the interaction layer into an LSTM to connect the sequential information flow in the dialogue:

$$\boldsymbol{h}_t = \text{LSTM}(\boldsymbol{m}_t, \boldsymbol{h}_{t-1}) \quad (7)$$

where $\boldsymbol{h}_t \in \mathbb{R}^k$ is the hidden state for u_t. Since there are no dependencies among labels, we simply use a softmax classifier for handoff prediction:

$$\hat{\boldsymbol{y}}_t^h = \text{softmax}(W_\hbar \boldsymbol{h}_t + \boldsymbol{b}_\hbar) \quad (8)$$

where $W_\hbar \in \mathbb{R}^{|\Psi|\times k}$ and $\boldsymbol{b}_\hbar \in \mathbb{R}^{|\Psi|}$. $\hat{\boldsymbol{y}}_t^h \in \mathbb{R}^{|\Psi|}$ is the predicted handoff probability distribution of u_t.

For service satisfaction analysis, we first apply a transformer block (Vaswani et al., 2017) to model the long-range context of the dialogue further. Formally, we have $\hat{\boldsymbol{Q}} = \text{Transformer}(\boldsymbol{Q})$, where $\hat{\boldsymbol{Q}} = \{\hat{\boldsymbol{q}}_1, ..., \hat{\boldsymbol{q}}_L | \hat{\boldsymbol{q}}_t \in \mathbb{R}^k\}$.

Then we utilize a softmax function for estimating local satisfaction distribution $\boldsymbol{z}_t \in \mathbb{R}^{|\Omega|}$ of u_t:

$$\boldsymbol{z}_t = \text{softmax}(W_\xi \hat{\boldsymbol{q}}_t + \boldsymbol{b}_\xi) \quad (9)$$

where $W_\xi \in \mathbb{R}^{|\Omega|\times k}$ and $\boldsymbol{b}_\xi \in \mathbb{R}^{|\Omega|}$.

Since only a fraction of customer utterances can contribute to the final satisfaction rating, we introduce an attention strategy that enables our model to attend to customer utterances of different importance when merging the local satisfaction distribution. Formally, we measure the importance of each customer utterances as below:

$$\boldsymbol{\alpha} = \text{softmax}(\text{Mask}'_c(\boldsymbol{g}^\top \tanh(W_\mu \hat{\boldsymbol{Q}}^\top + \boldsymbol{b}_\mu))) \quad (10)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^L$. $W_\mu \in \mathbb{R}^{z\times k}$, $\boldsymbol{b}_\mu \in \mathbb{R}^z$, and $\boldsymbol{g} \in \mathbb{R}^z$ are trainable parameters. z is the number of attention units. Mask'_c denotes the masking function used to reserve customer utterances. $\boldsymbol{g}$ can be perceived as a high-level representation of a fixed query "Which is the critical utterance?".

Finally, we obtain the overall satisfaction distribution $\hat{\boldsymbol{y}}^s \in \mathbb{R}^{|\Omega|}$ as the weighted sum of local customer satisfaction distribution:

$$\hat{\boldsymbol{y}}^s = \sum_{t=1}^{L} \alpha_t \boldsymbol{z}_t \quad (11)$$

where α_t is the t-th weight of utterance u_t in $\boldsymbol{\alpha}$.

Statistics items	Clothes	Makeup
# Dialogues	10,000	3,540
# US (unsatisfied)	2,302	1,180
# MT (met)	6,399	1,180
# WS (well satisfied)	1,299	1,180
# Transferable Utterances	16,921	7,668
# Normal Utterances	237,891	86,778
Avg # Utterances	25.48	26.67
Avg # Tokens	7.64	7.87
Kappa	0.85	0.88

Table 1: Statistics of the datasets.

3.4 Joint Training

The objective function of MHCH is formulated as:

$$\mathcal{L}_1 = -\frac{1}{L}\sum_{t=1}^{L}\sum_{i=1}^{|\Psi|} y_{i,t}^{h}\log(\hat{y}_{i,t}^{h}) \tag{12}$$

The objective function of SSA is formulated as:

$$\mathcal{L}_2 = -\sum_{i=1}^{|\Omega|} y_i^s \log(\hat{y}_i^s) \tag{13}$$

Finally, we minimize the joint cross-entropy loss $\mathcal{L}$, which is obtained as follow:

$$\mathcal{L}(\Theta) = \mathcal{L}_1 + \eta * \mathcal{L}_2 + \delta \left\|\Theta\right\|_2^2 \tag{14}$$

where $\eta \in \mathbb{R}^+$ denotes the trade-off parameter, δ denotes the L_2 regularization weight, and Θ denotes all the trainable parameters of model. We use backpropagation to compute the gradients of the parameters, and update them with Adam (Kingma and Ba, 2015) optimizer.

4 Experiments and Results

4.1 Dataset and Experimental Settings

Our experiments are conducted based on two publicly available Chinese customer service dialogue datasets, namely **Clothes** and **Makeup**[2], collected by Song et al. (2019) from Taobao[3]. Both datasets have service satisfaction ratings from customer feedbacks and annotated sentiment labels of utterances. Note that the sentiment labels do not participate in our training process and are only used for test. Meanwhile, we also annotate the *transferable/normal* labels for both datasets according to the existing specifications (Liu et al., 2021). Two annotators with professional linguistics knowledge participated in the annotation task.

A summary of statistics, including Kappa value (Snow et al., 2008) for both datasets are given in Table 1. Clothes is a corpus with 10K dialogues in the *Clothes* domain, which has an imbalanced satisfaction distribution at dialogue level. Makeup is a corpus with 3,540 dialogues in the *Makeup* domain, which has a balanced satisfaction distribution dialogue level. Note that we do not adopt the original word segmentation. Figure 3 shows the relative handoff position distributions in different satisfaction ratings, where we take explicit request, negative emotion, and unsatisfactory answer handoffs into consideration. It indicates that handoff at the later phase of the conversation is more likely to get a lower service satisfaction rating.

Except BERT-based model, all texts are tokenized by a popular Chinese word segmentation utility called jieba[4]. The datasets are partitioned for training, validation, and test with an 80/10/10 split. For the BERT-based methods, we fine-tune the pre-trained model. For the other methods, we apply the pre-trained word vectors initially trained on Clothes and Makeup corpora by using CBOW (Mikolov et al., 2013). The dimension of word embedding is set as 200. Other trainable model parameters are initialized by sampling values from the Glorot uniform initializer (Glorot and Bengio, 2010). The sizes of hidden state k, Dense units d, attention units z, and batch size are selected from $\{32, 64, 128, 256, 512\}$. The dropout (Srivastava et al., 2014) rate and the loss weight η are selected from $(0, 1)$ by grid search. Finally, we train the models with an initial learning rate of 1.5×10^{-3} and 2×10^{-5} for regular baselines and BERT-based models. All the methods run on a server configured with a Tesla V100, 32 CPU, and 32G memory.

4.2 Baselines

We compare our model with 14 strong dialogue classification baseline models, which come from MHCH, SSA, and other similar tasks.

Generic Baselines: **HAN** (Yang et al., 2016) and **BERT**(Devlin et al., 2019)**+LSTM**. We adopt outputs and the last hidden of RNN to predict handoff labels and the satisfaction rating, respectively.

Baselines for the MHCH task: **HEC** (Kumar et al., 2018), **DialogueRNN** (Majumder et al., 2019), **CASA** (Raheja and Tetreault, 2019),

[2] https://github.com/songkaisong/ssa

[3] https://www.taobao.com

[4] https://pypi.org/project/jieba

Models	Clothes										Makeup									
	MHCH					SSA					MHCH					SSA				
	F1	Mac. F1	GT-I	GT-II	GT-III	WS F1	MT F1	US F1	Mac. F1	Acc.	F1	Mac. F1	GT-I	GT-II	GT-III	WS F1	MT F1	US F1	Mac. F1	Acc.
HAN	59.8	78.7	71.7	73.1	74.0	51.5	81.7	70.4	67.9	75.5	54.3	75.4	68.5	70.1	71.3	68.4	71.3	84.8	74.8	74.8
BERT+LSTM	60.4	78.9	73.4	74.9	75.9	42.2	84.2	72.9	66.4	77.6	59.1	78.0	72.0	73.0	73.7	66.7	72.9	87.2	75.6	76.0
HEC	59.8	78.7	71.2	72.3	73.0	-	-	-	-	-	57.1	76.8	68.0	69.5	70.5	-	-	-	-	-
DialogueRNN	60.8	79.2	73.1	74.6	75.6	-	-	-	-	-	58.3	77.4	68.8	70.5	71.6	-	-	-	-	-
CASA	62.0	79.8	73.6	75.0	75.9	-	-	-	-	-	58.4	77.5	70.6	72.7	73.9	-	-	-	-	-
LSTMLCA	62.6	80.1	72.4	73.9	74.8	-	-	-	-	-	57.4	77.0	70.2	71.7	72.6	-	-	-	-	-
CESTa	60.6	79.1	73.4	74.8	75.6	-	-	-	-	-	59.3	78.0	69.6	71.2	72.2	-	-	-	-	-
DAMI	66.7	82.2	74.2	75.9	77.1	-	-	-	-	-	61.1	79.0	73.3	74.4	75.2	-	-	-	-	-
MILNET	-	-	-	-	-	38.2	82.3	70.8	63.8	75.3	-	-	-	-	-	72.0	68.9	84.9	75.3	75.1
HMN	-	-	-	-	-	44.1	83.3	69.6	65.7	76.3	-	-	-	-	-	73.5	73.1	83.4	76.6	76.8
CAMIL	-	-	-	-	-	55.4	84.4	71.5	70.4	78.3	-	-	-	-	-	73.8	74.5	87.4	78.6	78.5
MT-ES	61.7	79.7	74.6	75.9	76.8	47.7	82.4	74.1	68.1	76.4	57.1	76.9	69.9	71.7	72.8	72.0	68.7	84.3	75.0	75.1
JointBiLSTM	62.0	79.9	75.0	76.1	76.9	26.7	82.1	69.4	59.4	74.5	59.3	78.0	70.1	72.0	73.1	74.5	72.2	83.7	76.8	76.8
DCR-Net	62.1	79.9	71.4	72.8	73.7	49.8	82.7	**76.6**	69.7	77.3	58.8	77.7	70.0	72.1	73.4	74.8	69.1	88.6	77.5	77.7
RSSN(ours)	**69.2***	**83.6***	**78.4***	**79.5***	**80.3***	**56.0**	**85.1**	74.0	**71.7***	**79.5***	**65.9***	**81.5***	**75.1***	**76.6***	**77.6***	**77.4***	**76.1***	**88.9**	**80.8***	**80.8***

Table 2: Experimental results of performance (%) comparison with base models on Makeup and Clothes test datasets. *Underline* shows the best performance for baselines. - means not applicable. ***Bold*** shows the best performance. * indicates statistical significance at $p < 0.05$ level compared to the best performance of baselines.

LSTMLCA (Dai et al., 2020), **CESTa** (Wang et al., 2020b), and **DAMI** (Liu et al., 2021).

Baselines for the SSA task: **MILNET** (Angelidis and Lapata, 2018), **HMN** (Shen et al., 2018), and **CAMIL** (Song et al., 2019).

Multi-task baselines: **MT-ES** (Ma et al., 2018), **JointBiLSTM** (Bodigutla et al., 2020), and **DCR-Net** (Qin et al., 2020). Specifically, We modify **DCR-Net** for our tasks by keeping the core self-attention and co-interactive relation layer.

For *DAMI*, we adopt the open-sourced code[5] to get the results. For *DialogueRNN*, we adapt the open-sourced code[6] to MHCH by keeping the core component unchanged. For *HAN, MILNET, HMN*, and *CAMIL* of SSA, we adopt the reported results from Song et al. (2019). We re-implement the other models. For *BERT+LSTM*, we adopt Chinese BERT-base model[7].

4.3 Comparative Study

Following Song et al. (2019), we adopt *Macro F1* (*Mac. F1*) and *Accuracy* (*Acc.*) for evaluating the SSA task. For evaluating the MHCH task, we adopt *F1*, *Macro F1* (*Mac. F1*), and *Golden Transfer within Tolerance (GT-T)* (Liu et al., 2021). *GT-T* considers the tolerance property of the MHCH task by the tolerance range T, which allows a "biased" prediction within it. The adjustment coefficient λ of *GT-T* penalizes early or delayed handoff. Likewise, we set λ as 0, and set T to range from 1 to 3 corresponding to *GT-I*, *GT-II*, and *GT-III*. The results of comparisons are shown in Table 2.

Models	Clothes					Makeup				
	MHCH			SSA		MHCH			SSA	
	F1	Mac. F1	GT-I	Mac. F1	Acc.	F1	Mac. F1	GT-I	Mac. F1	Acc.
Average	65.3	81.6	74.6	63.8	73.7	64.1	80.4	73.2	73.6	73.7
Voting	61.7	79.7	71.7	27.5	61.2	62.0	79.4	68.3	34.6	42.1
Last	66.4	82.1	74.6	67.0	75.6	62.8	79.8	71.5	76.6	76.8
w/o Interact	65.6	81.6	70.6	65.5	72.3	62.0	79.5	71.4	74.8	74.6
w/o Select	64.4	81.0	73.6	67.0	74.3	61.7	79.2	71.7	72.9	72.9
w/o Position	66.1	81.9	73.2	68.4	76.0	64.7	80.8	73.5	76.8	76.8
Full Model	69.2	83.6	78.4	71.7	79.5	65.9	81.5	75.1	80.8	80.8

Table 3: Ablation study performance (%) on Clothes and Makeup test datasets. w/o denotes "without".

We can observe that: (1) The proposed method outperforms all state-of-the-art models specific to one task in terms of all metrics on two datasets. This indicates that our proposed model can effectively capture useful information in both tasks by utilizing role and positional information to explicitly control the interaction between the two tasks. Hence, the performance of the two tasks can be boosted mutually. (2) By integrating MHCH with SSA, the multi-task learning model can obtain further improvements. Specifically, we find that the MHCH task has a positive influence on detecting the unsatisfied dialogue. Overall, DCR-Net and our model perform better than standalone models on *US F1* of satisfaction prediction. Intuitively, it is mainly because the interaction with handoff can more comprehensively reflect the local dissatisfaction dialogues than solely sentiment polarity analysis, which helps the joint model better identify dissatisfied dialogues for the SSA task.

[5] https://github.com/WeijiaLau/MHCH-DAMI

[6] https://github.com/senticnet/conv-emotion

[7] https://github.com/google-research/bert

Dialogue Content	RSSN	CAMIL	DAMI
C₁: When will I receive it? *Normal | Neutral*	*N | NE*	*NE*	*N*
A₂: Normally, you can receive it in two or three days. *Normal | -*	*N*	-	*N*
C₃: My skin is dull, I don’t know which one to choose, please recommend suitable one to me. *Normal | Neutral*	*N | NE*	*NE*	*N*
A₄: Do not know which one to choose? Suitable for your own is the best. We recommend you directly order and experience it. It can be returned for no reason within 7 days. If the problem has not been solved, you can ask for “human”. *Transferable | -*	*T*	-	*N*
C₅: For nothing… *Transferable | Negative*	*T | NG*	*NG*	*T*
C₆: Told you, my skin is dull. I don't know which one to buy. I won't buy it if there is no recommendation. *Transferable | Negative*	*T | NG*	*NG*	*T*
Satisfaction Rating: *Unsatisfied*	*US*	*MT*	-

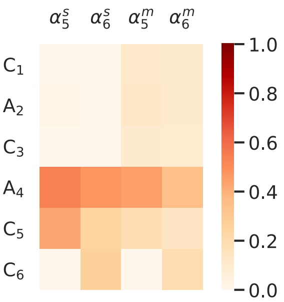


Figure 4: An example dialogue with predictions and attention distribution. $\mathbf{C}_i/\mathbf{A}_i$ denotes Customer/Chatbot utterance, followed by true *labels*. The sentiment labels of Customer utterances are also given along with the handoff labels. The other columns are the predictions of our model, CAMIL and DAMI, respectively. The satisfaction ratings of ground truth and predictions are in the last row of the table. N/T denotes *Normal/Transferable*.

4.4 Ablation

We perform several ablation tests in our model on two datasets and the results are recorded in Table 3. The results demonstrate the effectiveness of different components of our model.

w/o Interact: We modify the full version of our model by only sharing parameters of the Utterance and Matching Encoder. The performance degradation demonstrates the effectiveness of modeling the relations between two tasks with interaction. **w/o Select:** We remove the Role-Select mechanism to ignore the role information during the interaction process. The performance degradation indicates that straightforward interaction may bring noisy information for both tasks. **w/o Position**: We remove the positional weights in the MHCH to SSA sub-module. It performs well but worse than **Full Model** since the position information provide prior knowledge for controlling context interaction. **Average, Voting**, and **Last: Average** takes the average of the local satisfaction distributions of customer utterances for classification. **Voting** directly maps the majority local satisfaction distributions of customer into satisfaction prediction. **Last** takes the last customer’s satisfaction distribution as classification result. Average, Voting and Last are sub-optimal choices and perform worse than the Full Model. This is because the local satisfaction distributions contribute unequally to the overall satisfaction polarity. Also, the majority satisfaction polarity does not directly correlate with the overall satisfaction.

4.5 Case Study

Figure 4 illustrates our prediction results with an example dialogue, which is translated from Chinese text. In this case, three utterances (A_4, C_5 and C_6) are labeled as *transferable*, and two of them (C_5 and C_6) are labeled as “negative emotion”. Among them, A_4 is an unsatisfactory response, which arouses negative emotions of the customer. DAMI only predicts C_5 and C_6 as *transferable* utterances. However our model successfully detects all the *transferable* utterances. By mapping local satisfaction distribution of utterances to sentiment of utterances, our model is able to predict reasonable sentiment polarities for customer utterances (detailed analysis is in Subsection 4.6). Considering the context, the customer describes his/her skin problem at C_3 and asks for a recommendation. However, the chatbot does not give any recommendations and returns an irrelevant answer at A_4. We provide the attention distributions of the utterances on the right side of the example dialog. α_5^s and α_6^s are the *SSA to MHCH* attention distributions of C_5 and C_6; α_5^m and α_6^m are the *MHCH to SSA* attention distributions of C_5 and C_6. We can observe that attention distributions are concentrated on A_4 rather than other utterances. It is because A_4 is the main cause of negative emotion and dissatisfaction. This again demonstrates that our model can capture the mutual influence between local satisfaction and handoff, which is useful for prediction. In terms of final satisfaction rating, although CAMIL correctly predicts the sentiments of customer utterances, it gives a wrong prediction of satisfaction rating. Our model correctly predicts the satisfaction rating as *Unsatisfied* by considering the negative emotions and its cause of the unsatisfied response.

4.6 Results on Sentiment Classification

Song et al. (2019) utilize multiple instance learning to predict the satisfaction rating and the sentiment of customer utterances with the supervision

Models	**Clothes**					**Makeup**				
	PO F1	NE F1	NG F1	Mac. F1	Acc.	PO F1	NE F1	NG F1	Mac. F1	Acc.
MILNET	44.1	81.4	40.4	55.3	71.3	44.7	38.7	41.6	41.7	41.0
CAMIL	48.4	89.3	55.5	64.4	82.4	**54.4**	**72.5**	51.6	**59.5**	**64.7**
RSSN	**63.5**	**90.1**	**58.4**	**70.7**	**83.8**	51.3	67.8	**54.6**	57.9	61.6

Table 4: Results of sentiment classification by different models on Clothes and Makeup test datasets.

of the dialogue's satisfaction labels only during the training process. Similarly, our satisfaction prediction is based on the estimation of local satisfaction distributions while the utterance sentiment or satisfaction labels are unobserved. To compare and analyze the performance of utterance-level sentiment classification, we map these distributions into sentiments of utterances as the sentiment prediction results according to the distribution polarities, i.e., unsatisfied $\rightarrow$ negative (NG), met $\rightarrow$ neutral (NE), well-satisfied $\rightarrow$ positive (PO).

In Table 4, we compare the sentiment prediction results of MILNET, CAMIL, and our model. On Clothes dataset, our RSSN performs better than other baselines, while it performs worse than CAMIL on Makeup dataset. It is worth noting that our model achieves the best performance on both Clothes and Makeup datasets in terms of NG F1 metric. It indicates that MHCH task is sensitive to negative emotion and contributes more to negative emotion recognition than separate SSA models. From Table 2, we can also see that our model performs better than separate SSA models in terms of US F1, which is consistent with the findings of sentiment classification.

5 Conclusions and Future works

In this paper, we propose an innovative multi-task framework for service satisfaction analysis and machine-human chatting handoff, which deliberately establishes the mutual interrelation for each other. Specifically, we propose a Role-Selected Sharing Network for joint handoff prediction and satisfaction estimation, utilizing role and positional information to control knowledge transfer for both tasks. Extensive experiments and analyses reveal that explicitly modeling the interrelation between the two tasks can boost the performance mutually.

However, our model has not been calibrated to account for user preferences and biases, which we plan to address in future work. Moreover, we will further explore how to adjust the handoff priority with the assistance of personalized information.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by the National Natural Science Foundation of China (61876003, 62106039), the National Key R&D Program of China (2020YFC0832505), the Fundamental Research Funds for the Central Universities, and Alibaba Group through Alibaba Research Intern Program and Alibaba Research Fellowship Program.

References

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *TACL*, 6:17–31.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019a. Multi-domain conversation quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1911.08567*.

Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint turn and dialogue level user satisfaction estimation on mulit-domain conversations. In *Proc. of EMNLP: Findings*, pages 3897–3909.

Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. 2019b. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. *arXiv preprint arXiv:1908.07064*.

Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: Changing user needs and motivations. *Interactions*, 25(5):38–43.

Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *IJHCI*, pages 1–30.

Kyungyong Chung and Roy C Park. 2019. Chatbot-based heathcare service with a knowledge base for cloud computing. *Cluster Computing*, 22(1):1925–1937.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Layla El Asri, Hatim Khouzaimi, Romain Laroche, and Olivier Pietquin. 2014. Ordinal regression for interaction quality prediction. In *Proc. of ICASSP*, pages 3221–3225.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proc of AISTATS*, pages 249–256.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015a. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proc. of SIGDIAL*, pages 87–95.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015b. Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proc. of EMNLP*, pages 2243–2248.

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

Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P. Bigham. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proc. of CHI*, pages 1–13.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proc. of DIS*, pages 895–906.

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

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proc. of AAAI*, pages 3440–3447.

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

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

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proc. of AAAI*, pages 6818–6825.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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 *Proc. of AAAI*, pages 8665–8672.

Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. AliMe chat: A sequence to sequence and rerank based chatbot engine. In *Proc. of ACL*, pages 498–503.

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

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proc. of NAACL*, pages 3727–3733.

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. *TACL*, 7:375–386.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational AI: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In *Proc. of LREC*, pages 3369–3373.

Chenlin Shen, Changlong Sun, Jingjing Wang, Yangyang Kang, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2018. Sentiment classification towards question-answering with hierarchical matching network. In *Proc. of EMNLP*, pages 3654–3663.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, pages 254–263.

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. In *Proc. of EMNLP*, pages 198–207.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958.

Stefan Ultes. 2019. Improving interaction quality estimation with BiLSTMs and the impact on dialogue policy learning. In *Proc. of SIGDIAL*, pages 11–20.

Stefan Ultes, Robert ElChab, and Wolfgang Minker. 2014. Application and evaluation of a conditioned hidden markov model for estimating interaction quality of spoken dialogue systems. *Natural Interaction with Robots, Knowbots and Smartphones*, pages 303–312.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS*, pages 5998–6008.

Tianyi Wang, Yating Zhang, Xiaozhong Liu, Changlong Sun, and Qiong Zhang. 2020a. Masking orchestration: Multi-task pretraining for multi-role dialogue representation learning. In *Proc. of AAAI*, pages 9217–9224.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020b. Contextualized emotion recognition in conversation as sequence tagging. In *Proc. of SIGDIAL*, pages 186–195.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *Proc. of ICML*, pages 2397–2406.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proc. of NAACL*, pages 1480–1489.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.