Tracking Satisfaction States for Customer Satisfaction Prediction in E-commerce Service Chatbots

Yang Sun[1], Liangqing Wu[2], Shuangyong Song[2], Xiaoguang Yu[2], Xiaodong He[2], Guohong Fu[1,3]*

[1]School of Computer Science and Technology, Soochow University, China
[2]JD AI Research, Beijing, China
[3]Institute of Artificial Intelligence, Soochow University, China
ysun23@stu.suda.edu.cn,ghfu@suda.edu.cn
{wuliangqing,songshuangyong,cdyuxiaoguang,xiaodong.he}@jd.com

Abstract

Due to the increasing use of service chatbots in E-commerce platforms in recent years, customer satisfaction prediction (CSP) is gaining more and more attention. CSP is dedicated to evaluating subjective customer satisfaction in conversational service and thus helps improve customer service experience. However, previous methods focus on modeling customer-chatbot interaction across different turns, which are hard to represent the important dynamic satisfaction states throughout the customer journey. In this work, we investigate the problem of satisfaction states tracking and its effects on CSP in E-commerce service chatbots. To this end, we propose a dialogue-level classification model named DialogueCSP to track satisfaction states for CSP. In particular, we explore a novel two-step interaction module to represent the dynamic satisfaction states at each turn. In order to capture dialogue-level satisfaction states for CSP, we further introduce dialogue-aware attentions to integrate historical informative cues into the interaction module. To evaluate the proposed approach, we also build a Chinese E-commerce dataset for CSP. Experiment results demonstrate that our model significantly outperforms multiple baselines, illustrating the benefits of satisfaction states tracking on CSP.

1 Introduction

Customer satisfaction prediction (CSP) in E-commerce service chatbots is dedicated to determining the customer satisfaction level such as *strongly satisfied*, *satisfied*, *neutral*, *dissatisfied*, or *strongly dissatisfied* with a specific conversational service she/he has just received, as shown in Figure 1. Due to the increasing use of service chatbots in E-commerce platforms in recent years (Song et al., 2019; Bodigutla et al., 2020), CSP is gaining more and more attention in the field of natural language processing. On the one hand, to deliver an effective conversational service and further enhance the ability of service chatbots, it is crucial to understand whether customers are satisfied with chatbot responses. On the other hand, CSP provides a straightforward way to dynamically monitor the performance of customer-chatbot interactions in terms of customer satisfaction and thus helps to intervene in problematic conversational services immediately (Liang et al., 2021). Once it is recognized that the customer is dissatisfied, we can immediately switch to manual service, so as to improve customer service experience and reduce customer churn (Yao et al., 2020).

Figure 1: An example of the CSP task. Customer satisfaction states (smiling or crying face) keep changing throughout the customer journey, contributing to the dialogue-level satisfaction.

Existing research on CSP focuses on two different tasks, namely the turn-level CSP (Pragst et al., 2017) and the dialogue-level CSP (Ultes, 2019). The former aims to determine the customer satis-

*Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 616–625
October 12–17, 2022.

faction at each turn of customer-chatbot interaction while the latter is a task to predict the overall customer satisfaction with the whole dialogue. As shown in Figure 1, in a real scenario of conversational service, a few customers are willing to give their feedback after service. Obviously, asking customers for turn-level feedback will undeniably lead to poor customer experience (Park et al., 2020). Therefore, in this study, we concentrate on the dialogue-level CSP.

Many approaches have been proposed for CSP with a focus on conversational context representation and customer-chatbot interaction modeling. While earlier works exploit manual features or recurrent neural networks (RNNs) to represent conversational context (Walker et al., 1997; Yang et al., 2010; Jiang et al., 2015; Choi et al., 2019), recent studies exert more efforts on modeling customer-chatbot interaction with attention mechanisms (Song et al., 2019) or similarity-based methods (Yao et al., 2020). Although these studies have greatly promoted the progress of the CSP technique, most of them concentrate on the interaction between customer questions and chatbot answers across different turns. However, chatbot answers from future turns are invisible to customers in a real scenario, so these methods are hard to represent important satisfaction states during the customer journey.

Actually, customer satisfaction states arise from customer-chatbot interaction and are dynamically changing throughout the customer journey (Lemon and Verhoef, 2016; Lee et al., 2020; Kvale et al., 2020). As shown in Figure 1, the customer is first dissatisfied at the turn (2) and then becomes satisfied at the turn (4) when the problem is solved smoothly, resulting in an overall satisfaction level *satisfied*. Furthermore, integrating historical context is helpful for representing the satisfaction states at each turn. For example, in the dialogue in Figure 1, the customer asks a more detailed question at the turn (3) based on the preceding response *"describe it again"* from the chatbot.

To address the aforementioned issues, we propose a dialogue-level classification model for CSP in E-commerce service chatbots, namely DialogueCSP. It consists of three main modules: Firstly, a dialogue encoding module exploits convolutional neural networks (CNNs) (Kim, 2014) and Long Short-Term Memory (LSTM) networks to capture conversational context. Secondly, an interaction module is used to represent the customer satisfaction states at each turn. In particular, the interaction module utilizes two Gated Recurrent Units (GRUs) (Chung et al., 2014) to perform a two-step customer-chatbot interaction, namely local question-answer interaction and satisfaction state interaction. Furthermore, we introduce dialogue-aware attentions, including question attention, answer attention, and state attention. While the former two attentions integrate historical cues into the interaction module, the latter captures dialogue-level satisfaction representations. Finally, a decoding module is applied to predict the customer satisfaction for each dialogue. We also construct a Chinese E-commerce customer satisfaction prediction dataset (CECSP) that contains approximately 30k conversational services. Experimental results demonstrate that the proposed model outperforms the current state of the art on CECSP and other two benchmark datasets.

In summary, we make the following contributions:

- We propose a dialogue-level classification model for customer satisfaction prediction.
- We explore a novel two-step interaction module to handle both local question-answer and customer satisfaction state interactions at each turn and further integrate it with historical cues using dialogue-aware attentions to handle dialogue-level satisfaction representations.
- We construct a large Chinese E-commerce CSP dataset (CECSP). Experimental results show that the proposed model outperforms multiple baselines.[1]

2 Related Work

Recently, CSP has attracted much attention due to the increasing use of service chatbots in many different aspects of our lives (Hashemi et al., 2018; Choi et al., 2019; Kachuee et al., 2021). Some studies focus on addressing turn-level satisfaction prediction with human annotations (Pragst et al., 2017; Rach et al., 2017). However, they are not scalable in terms of annotation costs due to the large volume of conversational services in E-commerce. Therefore, recent studies explore contrastive learning (Kachuee et al., 2021) and reinforcement learning

[1]Our code is available at https://github.com/McSumail/DialogueCSP, and the dataset will be released after encryption.

(Liang et al., 2021) to make them more suitable for E-commerce customer service.

Most of the existing works exert more effort on dialogue-level satisfaction prediction since few customers are willing to give their feedback after service. While earlier methods rely on manual features (Walker et al., 1997; Yang et al., 2010), recent studies use deep neural networks to model conversational context and customer-chatbot interaction. Hashemi et al. (2018) exploit LSTMs to capture the sequential context features within a dialogue and use the hidden states of the last turn for satisfaction prediction. To enhance dialogue-level representations, Ultes (2019) apply an attention mechanism over LSTM layers to capture information from each turn. To model customer-chatbot interaction, Song et al. (2019) use each customer question to capture relevant information from all chatbot answers, while Yao et al. (2020) compute the semantic similarity scores between customer questions and chatbot answers across different turns. However, these methods both exploit the information from future turns that are invisible to customers in a real scenario to capture turn-level features. Therefore, they are hard to model the customer journey and track the dynamic satisfaction states within a conversational service. This work differs in that we consider both question-answer and customer satisfaction state interactions at each turn, and thus design a novel two-step interaction module to track the satisfaction states throughout the customer journey.

3 Dataset

For our experiments, we collect conversational services from one of the largest E-commerce platforms and construct a Chinese E-commerce CSP dataset. In the following, we will introduce the annotation strategy and compare this dataset with other benchmark datasets (Song et al., 2019).

3.1 Dataset Annotation

We use **real customer feedback** as the dialogue-level satisfaction labels which include *strongly satisfied*, *satisfied*, *neutral*, *dissatisfied*, and *strongly dissatisfied*. For the quality of the annotation, we then assign several experienced customer service coordinator to check whether the feedback is consistent with the conversational service, and about 20% dialogues were excluded from the final dataset.

Statistics items	CECSP	Clothes	Makeup
# of Train	22576	8000	2832
# of Val	2822	1000	354
# of Test	2801	1000	354
# of strongly dissatisfied	3158	-	-
# of dissatisfied	1417	2302	1180
# of neutral	2633	6399	1180
# of satisfied	10840	1299	1180
# of strongly satisfied	10151	-	-
Avg. # of turns per dialog	3.67	8.14	8.01
Max # of turns per dialog	10	18	16
Min # of turns per dialog	1	2	2
Multiple domains	Yes	No	No
Turn-level annotation	No	Yes	Yes

Table 1: The comparison of the three datasets in some key statistics. While **CECSP** is our constructed Chinese E-commerce CSP dataset, **Clothes** and **Makeup** are two benchmark datasets.

3.2 Comparison with Other Datasets

Table 1 shows some key statistics of the three datasets. As we can see, CECSP consists of more but shorter conversational service compared to **Clothes** (Song et al., 2019) and **Makeup** (Song et al., 2019). While **Clothes** and **Makeup** only collect conversational services in post-sale, CECSP consists of dialogues from multiple domains such as logistic, post-sale and VIP service. Due to ethical concerns, we follow Song et al. (2019) and transform segmented Chinese word[2] into word index in the final dataset.

4 Methodology

4.1 Problem Definition

Suppose there is a conversational service consisting of n turns of interaction $\{(q_1 : a_1), (q_2 : a_2), ..., (q_n : a_n)\}$, where q_i is the i-th question asked by the customer and a_i is its corresponding answer from the chatbot, the goal of CSP is to predict the satisfaction label for this dialogue, which is one of the five classes: *strongly satisfied*, *satisfied*, *neutral*, *dissatisfied*, and *strongly dissatisfied*.

4.2 Model Overview

As illustrated in Figure 2, the proposed framework for CSP consists of three main components, namely dialogue encoding, satisfaction states tracking, and satisfaction prediction. Firstly, we encode the utterances of input dialogues into context-dependent vectors. Next, an interaction module

[2]The segmentation toolkit is open source and available at https://github.com/fxsjy/jieba

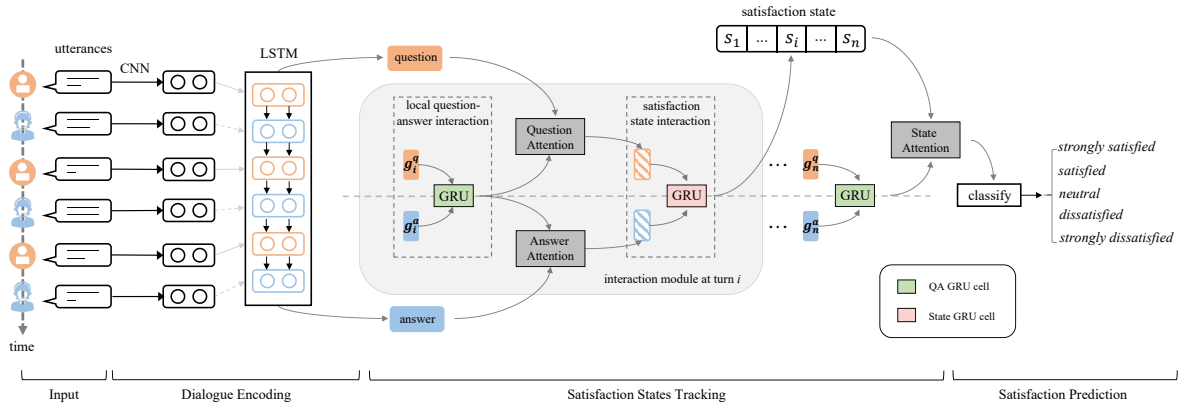

Figure 2: Overview of our proposed model for dialogue-level CSP, congruent to the illustration in Methodology.

with two GRU cells is applied to perform a two-step customer-chatbot interaction to represent the customer satisfaction states at each turn. Meanwhile, dialogue-aware attentions integrate the historical information into the interaction module and capture dialogue-level satisfaction representations. Finally, the dialogue-level satisfaction representations are used to predict satisfaction labels for dialogues. In the following sections, we will explain each component in detail.

4.3 Dialogue Encoding

The input of our model is a sequence of utterances consisting of word index. The goal of dialogue encoding is to encode the utterance sequence into context-dependent vectors using CNNs and LSTM for subsequent customer-chatbot interaction.[3]

4.3.1 Utterance Encoding

CNNs are capable of capturing n-gram information from an utterance (Kim, 2014). We leverage a CNN layer with max-pooling to extract context-independent features of each utterance. Concretely, the input is the 300 dimensional pre-trained 840B GloVe vectors (Pennington et al., 2014). We employ three filters of size 3, 4, and 5 with 50 feature maps each. These feature maps are further processed by max-pooling and ReLU activation (Nair and Hinton, 2010). Then, these features are concatenated and fed to a 100 dimensional fully connected layer, whose activations form the representations of the utterances.

[3]We also used pre-trained BERT-Base to encode the original conversational service from CECSP, but the results were not satisfactory.

4.3.2 Context Encoding

The LSTM introduces gating mechanism into recurrent neural networks to capture long-term dependencies from input sequences. In this part, we use a LSTM network to capture sequential context information,

$$g_i = \text{LSTM}\left(g_{i-1}, u_i\right) \tag{1}$$

where $i = 1, 2, \ldots, n$, u_i and g_i are context-independent and sequential utterance representations, respectively. Then, we denote question and answer representations as $M^q = [g_1^q, g_2^q, \ldots, g_n^q]$ and $M^a = [g_1^a, g_2^a, \ldots, g_n^a]$.

4.4 Satisfaction States Tracking

Since customer satisfaction states keep changing throughout the customer journey, we design an interaction module to perform a two-step customer-chatbot interaction to represent the customer satisfaction states at each turn. Figure 2 shows the details of the interaction module at turn i.

4.4.1 Dialogue-aware Attention

Attention mechanisms aim to capture the most relevant information and are widely applied on different natural language processing tasks (Bahdanau et al., 2015; Luo et al., 2018; Sinha et al., 2018). Given the query q, the key k, and the value v, the attention output o is computed as follows:

$$w = f\left(q, k\right) \tag{2}$$

$$\tilde{w} = w - m \tag{3}$$

$$o = \text{softmax}(\tilde{w})v \tag{4}$$

where f is a function that computes a single scalar from q and k. The attention mask m is a matrix with the same shape as the attention weights w.

The value of m_j is set to be $+\infty$ only when the attention for the j-th vector in k is masked, and set to be 0 otherwise.

In conversational service, the customer satisfaction state at turn i are most related to the questions and answers at turn (1)~(i) (Lemon and Verhoef, 2016). Therefore, the attention mechanism used by Song et al. (2019) that model the customer-chatbot interaction across different turns is hard to capture satisfaction states throughout the customer journey. To address this issue, we design dialogue-aware attentions by using different inputs and masking strategies to integrate historical cues into the interaction module and capture dialogue-level satisfaction representations.

4.4.2 Local Question-Answer Interaction

Since customer satisfaction states arise from the customer-chatbot interaction (Lee et al., 2020; Kvale et al., 2020), we adopt a QA GRU cell to model the local question-answer interaction and capture satisfaction features,

$$s_i^{qa} = \mathrm{GRU}^{qa}\left(g_i^a, g_i^q\right) \tag{5}$$

where $i = 1, 2, \ldots, n$.

4.4.3 Question Attention

Due to the nature of dialogues, contextual information plays an vital role in customer satisfaction states (Lemon and Verhoef, 2016; Kvale et al., 2020). Therefore, we design an attention mechanism to match relevant historical cues from the question representations:

$$q, k, v = s_i^{qa}, M^q, M^q \tag{6}$$

$$m_j^{que} = \begin{cases} +\infty, & j \notin \{\tilde{g}_1^q, \tilde{g}_2^q, \ldots, \tilde{g}_i^q\} \\ 0, & \text{Otherwise} \end{cases} \tag{7}$$

$$\tilde{q}_i = \mathrm{QueAttn}\left(q, k, v, m_j^{que}\right) \tag{8}$$

The masking strategy m_j^{que} separates future turns from the interaction at the current turn, which is more consistent with the customer journey.

4.4.4 Answer Attention

We also devise another attention mechanism to capture historical cues from the answer representations:

$$q, k, v = s_i^{qa}, M^a, M^a \tag{9}$$

$$m_j^{ans} = \begin{cases} +\infty, & j \notin \{\tilde{g}_1^a, \tilde{g}_2^a, \ldots, \tilde{g}_i^a\} \\ 0, & \text{Otherwise} \end{cases} \tag{10}$$

$$\tilde{a}_i = \mathrm{AnsAttn}(q, k, v, m_j^{ans}) \tag{11}$$

4.4.5 Satisfaction State Interaction

With the attention mechanisms described above, we successfully collect informative cues from the historical questions and answers. Then, we use a State GRU cell to lever these cues to represent the customer satisfaction state s_i at turn i,

$$s_i = \mathrm{GRU}^s\left(\tilde{a}_i, \tilde{q}_i\right) \tag{12}$$

where $i = 1, 2, \ldots, n$.

4.4.6 State Attention

After applying the two-step interaction module at each turn, we denote the customer satisfaction states as $S = [s_1, s_2, \ldots, s_n]$. Then, we use state attention to capture the dialogue-level satisfaction representations $\tilde{s}$:

$$q, k, v = s_n^{qa}, S, S \tag{13}$$

$$m_j^{sta} = 0 \tag{14}$$

$$\tilde{s} = \mathrm{StaAttn}(q, k, v, m_j^{sta}) \tag{15}$$

4.5 Satisfaction Prediction

Finally, we classify each conversational service using a fully connected network:

$$h = \mathrm{ReLU}(W_r \tilde{s} + b_r) \tag{16}$$

$$\mathcal{P} = \mathrm{softmax}(W_{smax} h + b_{smax}) \tag{17}$$

$$\hat{y} = \underset{k}{\mathrm{argmax}}\left(\mathcal{P}[k]\right) \tag{18}$$

To train the model, we choose the cross-entropy loss function:

$$\mathcal{L}(\theta) = -\sum_{v \in y_V} \sum_{z=1}^{Z} Y_{vz} \ln P_{vz} \tag{19}$$

where $y_{\mathcal{V}}$ is the set of dialogue indices that have labels and Y is the label indicator matrix, and θ is the collection of trainable parameters in DialogueCSP.

5 Experimental Settings

In this section, we present the experimental settings including implementation details and baselines.

5.1 Implementation Details

We use the validation set to tune hyperparameters. The batch size is set to be {128,64,64} for CECSP, Clothes, and Makeup. We adopt Adam (Kingma and Ba, 2015) as the optimizer with an initial learning rate of {1e-3,1e-4,1e-4} and L2 weight decay of {1e-4, 1e-5, 1e-5} for CECSP, Clothes, and

Makeup, respectively. The dropout (Srivastava et al., 2014) is set to be 0.5. We train all models for a maximum of 100 epochs and stop training if the validation loss does not decrease for 20 consecutive epochs.

5.2 Baseline Methods

For a comprehensive evaluation of our proposed DialogueCSP, we compare it with the following baseline methods:

LSTMCSP (Hashemi et al., 2018): This model adopts a Bi-directional LSTM network to capture the contextual information of conversational services and uses the hidden states of the last turn for satisfaction prediction.

LSTM+Attn (Ultes, 2019): This model applies an attention mechanism over Bi-directional LSTM layers to capture information from all turns within a service.

DialogueGCN (Ghosal et al., 2019): It is a graph-based model which encodes the relative positions between customers and chatbots within a window context.

CAMIL (Song et al., 2019): This model uses each question to capture information from all answers to model customer-chatbot interaction. Additionally, it exploits turn-level sentiment information by multiple instance learning.

LSTM+MTL (Bodigutla et al., 2020): It is a multi-task learning network that uses the hidden states of LSTM layers to predict dialogue-level and turn-level satisfaction jointly.

LSTM-Cross (Yao et al., 2020): It is the latest work for dialogue-level CSP which uses LSTM networks to capture contextual features and computes the semantic similarity scores between customer questions and chatbot answers across different turns. Then, these similarity scores are concatenated with the contextual features for satisfaction prediction.

6 Results and Analysis

6.1 Overall Results

Table 2 shows the comparison results for CSP in conversational services. Our proposed DialogueCSP consistently achieves better performance than the baseline methods on all datasets, while being statistically significant under the paired t-test ($p<0.05$). Besides, we can make another three observations as follows, which help to understand the CSP task and the advantages of DialogueCSP.

Model	CECSP		Clothes		Makeup	
	Acc.	F1	Acc.	F1	Acc.	F1
LSTMCSP	51.85	49.57	75.59	75.78	76.31	76.56
LSTM+Attn	53.09	51.02	77.12	77.28	77.56	77.52
DialogueGCN	53.69	51.35	76.89	76.82	77.72	77.78
CAMIL	55.43	52.92	78.30$^{\#}$	78.40	78.50$^{\#}$	78.64
LSTM+MTL	–	–	78.21	78.12	78.18	78.08
LSTM-Cross	55.51	53.11	78.91	79.33	79.88	79.58
DialogueCSP	**57.48**	**54.98**	**81.18**	**80.93**	**81.30**	**81.62**

Table 2: Overall performance on the three datasets. We use the accuracy and the weighted F1 score to evaluate each model. Scores marked by "#" are reported results, while others are based on our re-implementation.

Firstly, although LSTM+Attn only applies a vanilla attention mechanism compared to LSTM-CSP, the improvements on the three datasets are significant. This indicates that dialogue-level CSP must capture information from all turns in conversational services. Since chatbots respond to each customer question immediately, the relative positions between customer questions and chatbot answers are fixed. Therefore, the position model in DialogueGCN does not work here.

Secondly, CAMIL takes turn-level sentiment information into account and achieve better performance than previous strategies. However, the improvement of the method on CECSP is more obvious than that on Clothes and Makeup. After examining the datasets, we find that the average conversational service length is 3.67 turns in CECSP which is much shorter than that in Clothes and Makeup. When the lengths are short, especially only 1 or 2 turns, overall satisfaction is more related to turn-level sentiment information (Bodigutla et al., 2020).

Thirdly, CAMIL and LSTM-Cross achieve better performance than other baselines due to their customer-chatbot interaction modeling methods. While these methods focus on questions and answers across different turns, our proposed DialogueCSP exploits a two-step interaction module to better model the customer journey and thus capture important customer satisfaction states.

6.2 Different Interaction Modeling Methods

In this section, we make a comparison between different interaction modeling methods. To this end, we modify our two-step interaction module with the following two methods. The first one is the same as LSTM-Cross (Yao et al., 2020). We compute the semantic similarity scores between the question and answer at the same turn. Then we con-

Method	Weighted F1 score		
	CECSP	Clothes	Makeup
DialogueCSP	**54.98**	**80.93**	**81.62**
DialogueCSP-similarity	54.01	79.71	80.38
DialogueCSP-global attn	54.34	80.07	80.90

Table 3: Results of comparison between different interaction modeling methods. We modify our interaction module with another two methods and evaluate them on the three datasets.

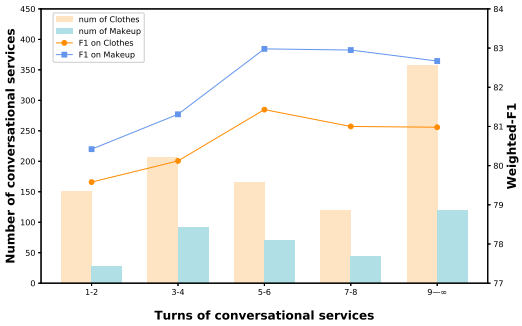

Figure 3: The influence of conversational service length on CSP. We divide the test set of Clothes and Makeup into five subsets in terms of conversational turns and further evaluate DialogueCSP over these subsets.

catenate them with obtained contextual features for satisfaction prediction. The second one is replacing the dialogue-aware attentions with the global attention (Song et al., 2019) to capture contextual information.

The results of different interaction modeling methods are shown in Table 3. We observe that our interaction modeling method is around 1% better than other methods in weighted F1 scores. Since customers can directly choose the options provided by chatbots, high semantic similarity scores don't always mean the high customer satisfaction. For instance, if customers choose "None of the above" from provided options, they may be dissatisfied. Besides, chatbot answers from future turns are invisible to customers within a conversational service. Therefore, global attention used in Song et al. (2019) is hard to capture the customer satisfaction states during the customer journey, leading to its inferior performance.

6.3 Influence of Conversational Service Length

In this section, we experiment on Clothes and Makeup to examine the influence of conversational service length.

Method	Weighted F1 score		
	CECSP	Clothes	Makeup
DialogueCSP	**54.98**	**80.93**	**81.62**
- 1st-step inter	54.60(↓ 0.38)	80.61(↓ 0.32)	80.94(↓ 0.68)
- 2nd-step inter	54.02(↓ **0.96**)	80.08(↓ **0.85**)	80.46(↓ **1.16**)
- Question Attn	54.51(↓ 0.47)	80.49(↓ 0.44)	80.80(↓ 0.82)
- Answer Attn	54.43(↓ 0.55)	80.37(↓ 0.56)	80.90(↓ 0.72)
- State Attn	54.64(↓ 0.34)	80.21(↓ 0.72)	80.64(↓ 0.98)

Table 4: Results of ablation study on the three datasets. 1st-step inter and 2nd-step inter stand for first-step interaction and second-step interaction, respectively.

As shown in Figure 3, whether on Clothes or Makeup, as the turns of conversational services increase, the performance of our proposed approach first rises significantly and then decreases. When conversational services length is short, there are few changes of customer satisfaction states (Lemon and Verhoef, 2016; Lee et al., 2020). Therefore, in these cases, the interaction module in DialogueCSP that captures satisfaction states does not work. Moreover, DialogueCSP uses dialogue-aware attentions to integrate historical information into customer-chatbot interaction. When the turns of services increase, there are more informative cues from preceding questions and answers which contribute to customer satisfaction states. As a result, DialogueCSP achieves weighted F1 scores of 81.43% and 82.98% on the subsets where the turns are 5 or 6. Further, it is still a challenge to handle the intricate context information when the turns are over 6, leading to the decline of DialogueCSP.

6.4 Ablation Study

In this ablation study, we analyze the impact of five components by removing one of them at a time from DialogueCSP. The results are presented in Table 4.

We can observe that the performance of DialogueCSP drops on the three datasets when any of the components is removed, suggesting that all these components contribute to the improvement of DialogueCSP. However, their contributions can be distinguished. By eliminating second-step interaction, our model drops the most by 0.96% on CECSP, 0.85% on Clothes, and 1.16% on Makeup in weighted F1 scores, which implies the importance of modeling the satisfaction state interaction.

Moreover, we found that Question Attention and Answer Attention also play important roles in our model. This phenomenon supports our argument that customer satisfaction states have close bonds with not only the questions and answers at

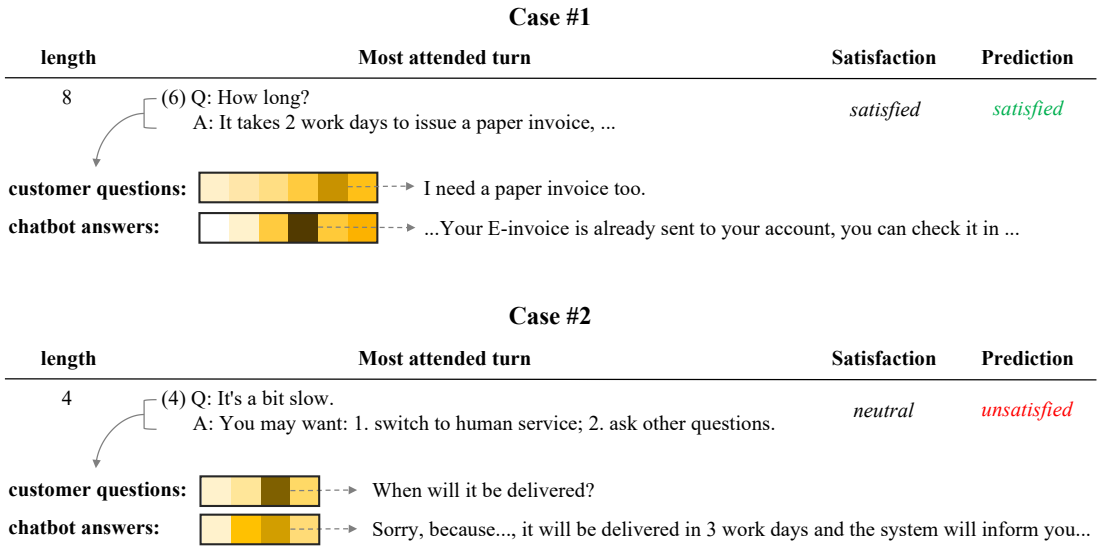

Figure 4: Results of case analysis, where some turns of two conversational services are provided, along with the visualization of attention weights between different context memories and the most attended turn (selected according to the highest attention weight computed by State Attention). The darker colors mean larger attention weights.

the current single turn but also historical information. Further, while State Attention is more important than Question Attention and Answer Attention on Clothes and Makeup, it is the opposite on CECSP. After delving into the datasets, we found that the average conversational service length is around 8 turns in Clothes and Makeup, which is much longer than that in CECSP. Therefore, it is important to weigh multiple satisfaction states to generate dialogue-level representations on Clothes and Makeup.

6.5 Case Analysis

For a comprehensive understanding of our proposed method, we visualize its performance by a case analysis on the test set of CECSP. In short, we found that integrating historical information into customer-chatbot interaction can be a double-edged sword. As illustrated in Figure 4, the dialogue-aware attentions can capture useful historical information and help make a good prediction (Case #1). However, focusing too much on historical information may hinder the understanding of neutral utterances of customers (Case #2). Therefore, it is necessary to explore other mechanisms rather than merely relying on popular attention to handle historical information for CSP.

Besides, we also observe from these two cases that the most attended turns of customer satisfaction states are among the end of the dialogues. After examining the whole test sets of the three datasets, we found that 40% of the most attended turns are the last turn of conversational services, which is in tune with the conclusion from the previous studies (Hashemi et al., 2018; Yao et al., 2020).

7 Conclusion

In this paper, we investigate the importance of satisfaction states tracking in dialogue-level CSP in E-commerce service chatbots. We propose a dialogue-level classification model and design a two-step interaction module to handle both local question-answer and customer satisfaction state interactions throughout the customer journey. To capture dialogue-level satisfaction representations, we further introduce dialogue-aware attentions to integrate historical information into the interaction module. Besides, we also build a Chinese E-commerce dataset for CSP to evaluate the proposed approach. Experimental results on this dataset and two released corpora show that our proposed model outperforms all the baselines. Our further analysis illustrates that tracking the satisfaction states is more helpful for modeling customer-chatbot interaction than previous strategies. In addition, our experiments also show that integrating historical information with customer-chatbot interaction is of great value to CSP.

In our future work, we would like to explore more effective methods to model customer-chatbot interaction. Moreover, we also plan to investigate the importance of customer intentions in handling informative cues for CSP.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (No.62076173, U1836222), the High-level Entrepreneurship and Innovation Plan of Jiangsu Province (No.JSSCRC2021524), and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

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

Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and online satisfaction prediction in open-domain conversational systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1281–1290.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

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

Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A Crook. 2018. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1183–1192.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*, pages 506–516.

Mohammad Kachuee, Hao Yuan, Young-Bum Kim, and Sungjin Lee. 2021. Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4053–4064.

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

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Knut Kvale, Eleonora Freddi, Stig Hodnebrog, Olav Alexander Sell, and Asbjørn Følstad. 2020. Understanding the user experience of customer service chatbots: What can we learn from customer satisfaction surveys? In *International Workshop on Chatbot Research and Design*, pages 205–218. Springer.

Ching-Hung Lee, Qiye Li, Yu-Chi Lee, and Chih-Wen Shih. 2020. Service design for intelligent exhibition guidance service based on dynamic customer experience. *Industrial Management & Data Systems*.

Katherine N Lemon and Peter C Verhoef. 2016. Understanding customer experience throughout the customer journey. *Journal of marketing*, 80(6):69–96.

Runze Liang, Ryuichi Takanobu, Fenglin Li, Ji Zhang, Haiqing Chen, and Minlie Huang. 2021. Turn-level user satisfaction estimation in e-commerce customer service.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

Dookun Park, Hao Yuan, Dongmin Kim, Yinglei Zhang Spyros Matsoukas, Young-Bum Kim, Ruhi Sarikaya Chenlei Guo Yuan Ling, Kevin Quinn, Tuan-Hung Pham, and Benjamin Yao Sungjin Lee. 2020. Large-scale hybrid approach for predicting user satisfaction with conversational agents.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Louisa Pragst, Stefan Ultes, and Wolfgang Minker. 2017. Recurrent neural network interaction quality estimation. In *Dialogues with Social Robots*, pages 381–393. Springer.

Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2017. Interaction quality estimation using long short-term memories. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 164–169.

Koustuv Sinha, Yue Dong, Jackie Chi Kit Cheung, and Derek Ruths. 2018. A hierarchical neural attention-based text classifier. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 817–823, Brussels, Belgium. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 198–207, Hong Kong, China. Association for Computational Linguistics.

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

Stefan Ultes. 2019. Improving interaction quality estimation with bilstms and the impact on dialogue policy learning. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 11–20.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Zhaojun Yang, Baichuan Li, Yi Zhu, Irwin King, Gina Levow, and Helen Meng. 2010. Collaborative filtering model for user satisfaction prediction in spoken dialog system evaluation. In *2010 IEEE Spoken Language Technology Workshop*, pages 472–477. IEEE.

Riheng Yao, Shuangyong Song, Qiudan Li, Chao Wang, Huan Chen, Haiqing Chen, and Daniel Dajun Zeng. 2020. Session-level user satisfaction prediction for customer service chatbot in e-commerce (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, pages 13973–13974.