CMCC: A Comprehensive and Large-Scale Human-Human Dataset for Dialogue Systems

Yi Huang[1,4], Xiaoting Wu[1,4], Si Chen[1,4], Wei Hu[1,4], Qing Zhu[1,4], Junlan Feng[1,4], Chao Deng[1,4], Zhijian Ou[3,4] and Jiangjiang Zhao[2]

[1]JIUTIAN Team, China Mobile Research,[2]China Mobile Online Marketing and Services Center
[3]Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University
[4]Tsinghua University-China Mobile Communications Group Co., Ltd. Joint Institute

[1] {huangyi,wuxiaoting,chensiyjy,huweiyjy,zhuqingai,fengjunlan,dengchao}@chinamobile.com
[3]ozj@tsinghua.edu.cn, [2]zhaojiangjiang@cmos.chinamobile.com

Abstract

Dialogue modeling problems severely limit the real-world deployment of neural conversational models and building a human-like dialogue agent is an extremely challenging task. Recently, data-driven models become more and more prevalent which need a huge amount of conversation data. In this paper, we release around 100,000 dialogue, which come from real-world dialogue transcripts between real users and customer-service staffs. We call this dataset as CMCC (China Mobile Customer Care) dataset, which differs from existing dialogue datasets in both size and nature significantly. The dataset reflects several characteristics of human-human conversations, e.g., task-driven, care-oriented, and long-term dependency among the context. It also covers various dialogue types including task-oriented, chitchat and conversational recommendation in real-world scenarios. To our knowledge, CMCC is the largest real human-human spoken dialogue dataset and has dozens of times the data scale of others, which shall significantly promote the training and evaluation of dialogue modeling methods. The results of extensive experiments indicate that CMCC is challenging and needs further effort. We hope that this resource will allow for more effective models across various dialogue sub-problems to be built in the future.

1 Introduction

Task-oriented dialogue systems (Young et al., 2013; Williams et al., 2017; Su et al., 2021; He et al., 2021; Jayanthi et al., 2021) are designed to assist user in completing daily tasks, which involve reasoning over multiple dialogue turns. Tremendous progress has been made recently, but building a human-like dialogue system is a challenging task remaining. To drive the progress of building dialogue systems using data-driven approaches, a number of conversational corpora have been released in the past. Task-oriented dialogue corpus, such as Frames (Asri et al., 2017), MultiWOZ (Budzianowski et al., 2018), CrossWOZ (Zhu et al., 2020), RiSAWOZ (Quan et al., 2020), are collected by two crowd workers playing the roles of the user and the system, which often leads to be small-scale, and can not sufficiently capture a number of challenges that arise with production scaling. More recently, some researchers construct dialogue datasets from real human-to-human scenario conversations, especially human-to-human customer service scenario, such as JDDC (Chen et al., 2020) and MobileCS (Ou et al., 2022). JDDC is collected from E-commerce scenario and annotates intent information. MobileCS is conducted from mobile customer service scenario and model the process as task-oriented conversations. Therefore, the entity information related to tasks is annotated. However, the complexity of the dialogue process is far more than TOD, in addition to task completion, it is also accompanied by emotional support that appease an angry customer and providing solutions.

Several emotional support conversation corpora (Welivita and Pu, 2020; Sharma et al., 2020; Rashkin et al., 2019; Sun et al., 2021) are designed to emotional chat or provide empathetic responding. Since the emotional supporters are not well-trained, existing datasets do not naturally exhibit examples or elements of supportive conversations. As a result, data-driven models which leverage such corpora are limited in their ability to explicitly learn how to provide effective support. ESConv (Liu et al., 2021) is collected by communication of trained individuals who play the roles of the seeker and the supporter, and guided by predefined emotional support conversation framework, however, it is more focused on alleviating the negative emotions that users encounter in their daily lives.

Despite the efforts in modeling emotional support, work that focuses specifically on modeling emotional care and support in task-oriented dia-

Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD), pages 48 - 61
December 7, 2022 ©2022 Association for Computational Linguistics

logue system is relatively limited. To this end, we design a customer service care-oriented taxonomy, and annotate care-oriented information for MobileCS dataset, covering 9 types of emotion labels and 17 types of customer service act labels finally. This new dataset consists of two parts, 8975 dialogues which are labeled with annotations of care-oriented information and other more than 90,000 unlabeled dialogues. We call this new dataset as **CMCC** (**C**hina **M**obile **C**ustomer **C**are) dataset. To be able to explain the patterns and trends of the conversation flow, we employ visualization methods to illustrate the most frequent exchanges and reveal how they temporally vary as dialogues proceed. Finally, we explore and demonstrate the effectiveness of care-oriented information for dialogue sub-tasks.

We highlight our contributions as follows:

- We provide a customer service care-oriented taxonomy, and conduct CMCC dataset on top of MoibleCS to facilitate the dialogue research.
- We employ visualization methods to illustrate the most frequent exchanges and reveal how patterns and trends temporally vary as dialogues proceed.
- We report the benchmark models and results of two evaluation tasks on CMCC, indicating that the dataset is a challenging testbed for future work.

2 Data Annotation

2.1 Motivation

We collect the CMCC dataset from the user-customer service conversations in real-life scenarios. These dialogues are inherently rich in user and customer service acts and emotional information. Therefore, our data annotation process integrates such features in the data and concentrates on how the customer service provides caring and empathetic acts according to a dynamic in the user's emotions. We present a novel data annotation approach by adding "User Emotion", "Expanded Customer Service Caring Act", and "Satisfaction" labels to emphasize the importance of emotions and "care-oriented" in the conversations. To our best knowledge, limited datasets have demonstrated such features in previous studies.

2.2 Guideline for Annotations

Our dataset is developed in multiple ways, which are provided in detail throughout the following sections. Compared to the MobileCS dataset, three new dimensions are added in our data annotation: user emotions, expanded customer service caring acts, and satisfaction. We also redefine the user intents to clarify the differences between intents and emotions.

2.2.1 User Emotion

We notice that users express various emotions throughout the conversations with customer service representatives, which can have a large impact on data division and annotation. Limited studies were conducted to consider this factor. As a result, we capture subtle user emotions throughout the conversations to derive and divide them into 8 labels for annotations. The refined annotation is necessary because customer service can act accordingly with "care-oriented" methods. We develop the "User Intent" labels from the MobileCS dataset, and add "Propose suggestion" and "Propose criticism" labels to separate intents from emotions. We pre-define an annotation schema and an intent set consisting of the 8 user emotion labels. At each turn, if emotions are explicitly expressed, the user's utterances are allowed to be annotated with one or more labels, which is common since multiple emotions could be expressed in one sentence in real-life conversations. The annotators are instructed to determine if the user's utterances contain emotions according to the schema and common sense. For example, "上次打电话说好了好了好了谁给我开的我要投诉他" (That's fine on the last phone call. Who opened the business for me last time, I want to complain to him), the label for this sentence is "Emotionally More Agitated". "这样哦要像每个人这样扣的话，还得了" (Would it be worth it if everyone's package was deducted like this?) is labeled with "Complain About A Problem".

2.2.2 Expanded Customer Service Caring Act

It's essential that good customer service provides "care-oriented" responses for emotional support. Adopting the original customer service acts from the MobileCS dataset, we derive and pre-define an "Expanded Customer Service Caring Act" set from the conversations. At each turn, the annotators are instructed to determine if the customer service utterances contain caring and empathetic acts to respond to user emotions and intents, allowing the use of multiple labels in one sentence. In addition, we extract keywords in each customer service utterance, such as "放心" (relax), "理解" (understand)， and "别着急" (don't worry), etc., indicating different customer service caring acts. For example, "还有剩下的是基本费用请您放心好吧" (The rest is the basic fee, please rest assured.) is labeled as "comfort". "确实是您的心情我非常理解" (I really understand how you feel) is labeled as "empathy".

2.2.3 Satisfaction

The satisfaction labels are pre-defined based on the context of conversations. Each conversation is required to be annotated with one of the three labels. "3" indicates the user is satisfied; "2" indicates the user accepts the suggestion provided by the customer service representative while the problem is unsolved; "1" indicates the user is unsatisfied. The annotators are instructed to comprehend the context of the conversation and annotate each conversation with one of the three satisfaction labels. For example, customer service: "请问还有其他可以帮到您吗？ " (Is there anything else I can help you with ?) user: "没有啦谢谢" (No thanks) is labeled as "3", suggesting that the user is very satisfied with the solution and result that the customer service provided.

2.3 Annotation Results

We improve the MobileCS dataset and further develop it by incorporating user emotions, expanded customer service caring acts, and satisfaction in the dialogues. Our novel dataset not only is motivated by the inherent nature of customer service-user dialogues but also aims to emphasize a "care-oriented" focus. Also, the experiment results support that the CMCC dataset is advancing and valuable in user-customer service conversations. The label set consists of 4 expanded customer service caring acts, 13 original customer service acts, 9 user emotions, 14 user intents, and 3 satisfaction labels in total.

2.4 Quality Control

Since the annotations are conducted on several dimensions simultaneously and differently on multiple criteria, missing and incorrect labels are inevitable problems we might face. To ensure a high-quality annotation result, we review and revise the missing or incorrect annotations based on several effective strategies. First, we conduct keyword extractions to check for the missing and incorrect labels, which are manually filtered out and re-labeled by the qualified annotators. For example, "您稍等一下好吗，我这边的话肯定会站在你的角度去想" (Can you wait a moment, I will definitely think from your point of view) misses the "empathy" label during the first round of annotation, and it's added during the manual check. Based on this strategy, we review and re-label the dataset two more times, which guarantees the efficiency and completeness of our annotation. Additionally, for the satisfaction annotation, we randomly sample 10% of conversations to check for the annotation quality. For example, "唉算了算了反正还有几天就" (Oh, forget it, there are still a few days left) is labeled as "3" in the first round of annotation, but it should be "2" instead.

Upon review, the missing labels and incorrect labels from the dataset are all revised and corrected for the quality control process. As a result, this ensures the high quality of our data annotation process.

3 Data Characteristics

This section mainly introduces the characteristics of the data. In addition to showing the number of conversations and labels in the dataset, we also demonstrate the characteristics of customer service dialogue data by visualizing the transition between customer service acts and user emotion in dialogues.

3.1 Data Statistics

The basic information of the labeled part in this dataset is shown in Table 1. The labeled data contains a total of 8,975 dialogues. The maximum

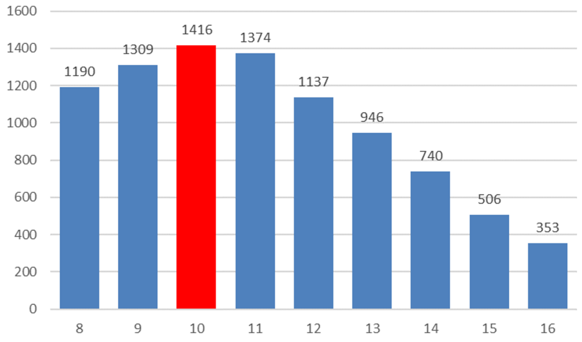

Figure 1: The histogram of dialogue turns. The horizontal axis is the number of dialogue turns, and the vertical axis is the number of dialogues, filtering the dialogues with less than 10 dialogues.

Criteria	Statistics
Total no. of dialogues	8,975
Total no. of dialogue turns	100,139
Average no. of turns per dialogue	22.31
Maximum no. of turns per dialogue	16 (353 dialogues)
Minimum no. of turns per dialogue	5 (1 dialogue)
Total no. of customer service turns	100,139
Total no. of user turns	100,138
Average no. of customer service tokens per dialogue turn	25.27
Average no. of user tokens per dialogue turn	14.58

Table 1: Dialogue statistics in the dataset.

number of dialogue turns included in the dataset is 16. Figure 1 is a histogram of dialogue turns. It can be observed that most of the dialogue turns in the dataset are concentrated between 8 and 13. This means that the dialogue between the user and the customer service typically ends in around 10 turns. If there are situations such as user's problems that are difficult to solve, the number of turns in this dialogue will increase significantly.

The histogram of user negative emotion labels is shown in Figure 2. The statistical scope is all negative emotions of users in the dialogue, excluding neutral emotions. The largest proportion of the entire user emotion label is "Complain About A Problem". This label is about the user emotion that often appears on the user side in the field of customer service dialogue. It generally occurs when users complain about networks, fees, business use, business handling, and e-commerce after-sales. The second-largest user emotion label is "Emotionally More Agitated". This label indicates that various businesses or services have seriously affected the user experience, or that customer service has not effectively helped users to solve problems.

Figure 3 is a statistical histogram of customer service intent labels. It can be seen that the labels with the largest proportion of intent are "Inform" and "Passive Confirmation". "Inform" means that the customer service informs the user of certain information, usually definite information, such as the customer service will perform a certain operation, the problem will be solved within a certain period of time, etc. "Passive Confirmation" means the act of confirming based on the user's inquiry or information provided above. Since the common content of dialogues in the field of customer service is to solve the user's problem, the labels of "Inform" and "Passive Confirmation" will be ubiquitous in each turn of dialogue.

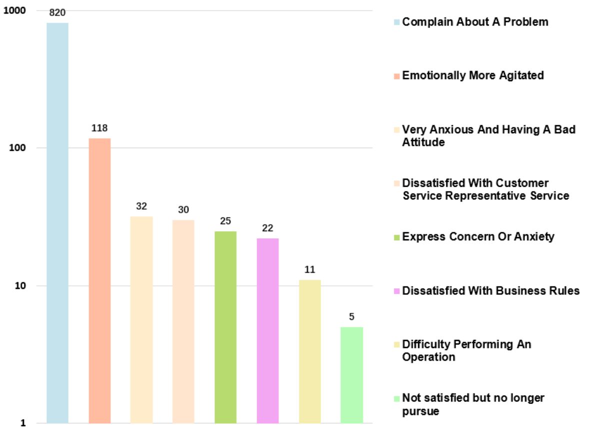

Figure 2: The histogram of user negative emotion. The horizontal axis is user emotion labels, and the vertical axis is the number of emotions.

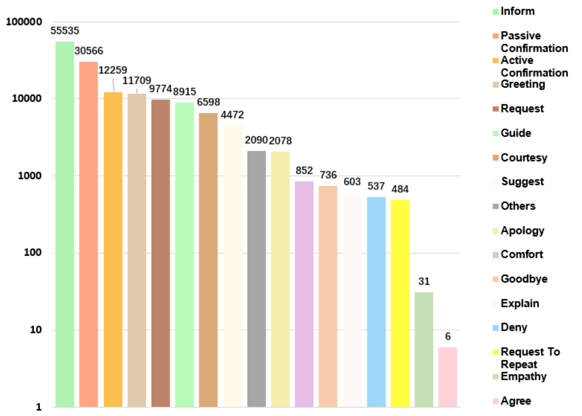

Figure 3: The histogram of customer service act. The horizontal axis is the customer service act label, and the vertical axis is the number of acts.

3.2 Data Structure

For a better understanding of the data structure, we investigate which customer service acts are frequently associated with users when responding to different emotional situations. We list the labeled instances of customer service act, user emotion, examples and the proportion of all labels, respectively (detailed in the appendix). Most conversations have multiple intent labels or emotion labels. For example, "Hello, nice to serve you, sorry to keep you waiting" includes "Apology" and "Greeting". Based on the statistics of user emotions and customer service acts, we observe the overall distribution of labels on the dataset.

In the following part, we will explore more about the conversion relationship between user emotions and customer service acts in the process of a dialogue. Figure 4 is a chord diagram of emotion-act labels. It represents the dialogue relationship between the user's emotion and the customer service act in the dialogue.The nodes and edges of the same color in the graph represent the user emotion and the customer service act corresponding to the next round of dialogue. It can be seen from the figure that the largest act dialogue is from "Complain About A Problem" to "Inform". This shows that when the user encounters a business problem, the customer service is more inclined to explain the cause or solution to the problem to the user. This phenomenon is in line with the most common scenario in the field of customer service, that is, customer service helps users solve related problems.

In order to intuitively observe the conversion relationship between user emotion and customer service act in multiple turns of dialogue, we draw a Sankey diagram of the dialogue between user emotion and customer service act in multi-turns. Figure 5 is the dialogue flow diagram of user emotions and customer service acts in four turns of dialogues. The first and third turns are user emotions, and the second and fourth turns are customer service acts. After the second turn of customer service replies to the user's dialogue with negative emotions, it can be observed that the user's emotion in the next turn, which is also the third turn, has become more "Neutral". This shows that as the customer service responds to the user's questions, the user's negative emotions will gradually disappear.

4 Experiments

In this section, we conduct experiments on the CMCC dataset. We focus on two tasks: dialogue response generation and user emotion recognition.

4.1 Dialogue Response Generation

Our experiments in this part mainly focus on the question: Can extra care-oriented information improve the generative dialogue model?

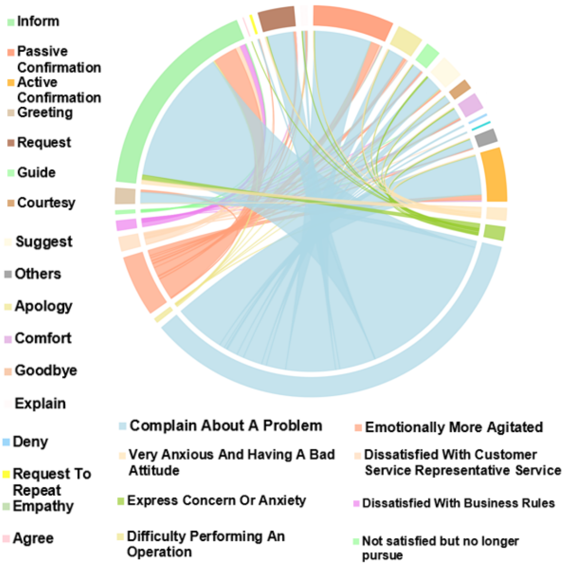

Figure 4: The chord diagram for user emotion and customer service act relationship. More details on labels can be found in the appendix. Best viewed in color.

4.1.1 Comparable Models

Similar to (Ou et al., 2022), we employ a Markovian generative architecture (MGA) (Liu et al., 2022) based on Chinese GPT-2 as baseline and build the following variant model:

Baseline The baseline model is a MGA generative model, which is designed to be $p_\theta(e_t, ui_t, a_t, r_t|e_{t-1}, u_t)$. u_t denotes the user utterance, e_t is entity names of dialogue history, ui_t is the user intent, and r_t is the customer service response, respectively, at turn t = 1, ..., T, for a diaogue of T turns.

Variants with care-oriented information To incorporate the care-oriented annotations into the baseline model, we add user emotion generation and expand original customer service acts to with caring acts in it. As is shown in Figure 6, for each customer service response, we append user emotion before corresponding customer service act. Then MGA generative process can ber represented as $p_\theta(e_t, ui_t, uemo_t, a_t, r_t|e_{t-1}, u_t)$, where $uemo_t$ is the user emotion at turn t. The model generates the response conditioned on the predicted user emotion and customer service act.

We study two variants that use care-oriented annotations in the experiments. (1) End2End: customer service response is generated conditioned on predicted customer service act and predicted user emotion, user emotion and customer service act are

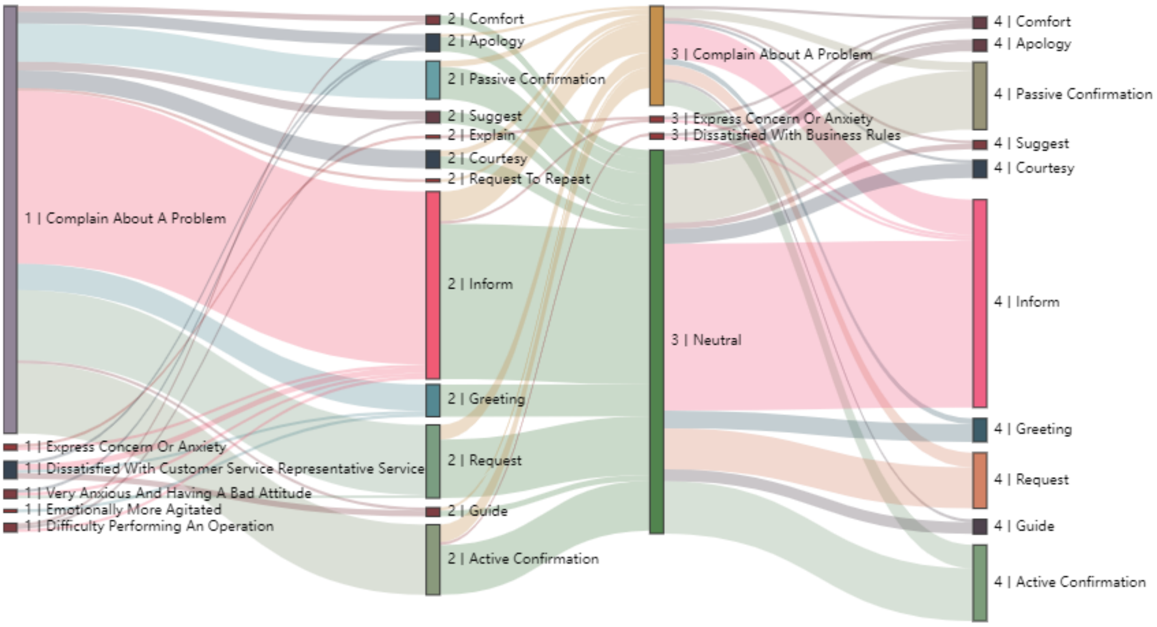

Figure 5: Dynamic transformation of user emotion vs. customer service act in the first four rounds of dialogue. Best viewed in color.

generated conditioned on KB result, KB result is queryed conditioned on predicted entity name and user intent. (2) Oracle: customer service response is generated conditioned on gold reference of customer service act, entity name, user intent and KB result.

4.1.2 Evaluation Measures

To investigate the impact of utilizing care-oriented information on the model performance with Chinese GPT-2 as backbone, we compare the performance of End2End and Oracle variants with the Baseline model. The automatic metrics include F1 score, Success rate and BLEU score. F1 is calculated for both predicted user intent and customer service act. Success rate (Budzianowski et al., 2018) is the percentage of generated dialogues that achieve user goals. BLEU-4 score (Papineni et al., 2002) evaluates the fluency of generated responses.

4.1.3 Experimental Results

The experimental results are shown in Table 2, which demonstrates the effectiveness of our model. There are 3 major findings from the experiments. (1) The Variant model has improved the Baseline model's performance of user intent F1, success rate and BLEU-4 of response, but the F1 of the customer service act has decreased slightly. It may be because the variant model expands the original cutomer service act labels, those with less data affects the overall performance. (2) Whether it is End2End or Oracle results, variant model is better than baseline model in BLEU-4 of response, we attribute it to the fact that care-oriented information matters and it enhances the dialogue generation positively. Care-oriented information includes user emotion and expanded customer service caring act, which part brings more gain will be analyzed in ablation experiments. (3) End2End results are lower than Oracle's results, because if predicted intermediate results is different from the ground truth, the generated response will be much different from the reference response.

Models	F1 for user intent	Success rate	F1 for customer service act	BLEU-4
Baseline Model (End2End)	0.642	0.315	0.575	4.137
Variant Model (End2End)	**0.656**	**0.357**	0.567	**4.669**
Baseline Model (Oracle)	–	–	–	6.230
Variant Model (Oracle)	–	–	–	**7.385**

Table 2: Results of automatic evaluation. The results in bold are better than the baseline.

4.1.4 Analysis

Our variant models consider care-oriented information, user emotion and customer service caring act. To investigate more, we conduct extra experiments and the analysis results.

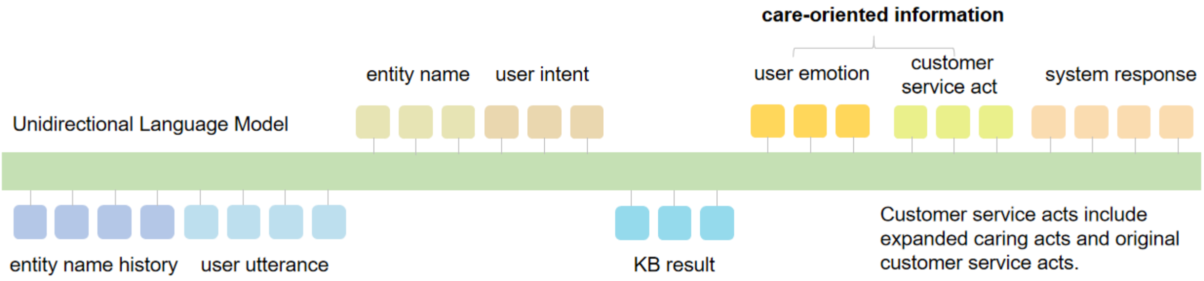
care-oriented information
entity name
user intent
user emotion
customer service act
system response
Unidirectional Language Model
entity name history
user utterance
KB result
Customer service acts include expanded caring acts and original customer service acts.

Figure 6: Variant model architecture with care-oriented information.

In order to verify the improvement brought by each added part (user emotion, expanded customer service caring act), we drop these two parts from the original variant model and check the performance changes. Results are presented in Table 3. We have the following observations: (1) In most circumstances, when user emotion is removed, BLEU-4 dropped more and success rate dropped less. (2) When expanded customer service caring act is removed, situation differs. That is, BLEU-4 dropped less and success rate dropped more. It indicates that expanded customer service caring act provides more gain for the entity-related part of the response, while user emotion plays more for the non-entity-related part (e.g., caring or empathetic responding).

Models	F1 for user intent	Success rate	F1 for customer service act	BLEU-4
Variant Model (End2End)	0.656	0.357	0.567	4.669
w/o user emotion	0.611	0.356	0.567	4.462
w/o expanded customer service caring act	0.656	0.340	0.577	4.657

Table 3: Evaluation results of ablation study.

In Table 4, examples are presented to compare the response generated by variant model and the baseline model. The first column is user utterance, the second column is the response of manual customer service, the third and fourth columns are the responses generated by variant model and baseline model respectively. In the first example, user reports that the broadband network is not working well, and accompanied by complaints. The variant model can generate the response with the soothing keyword "马上" (right now). In the second example, user's emotion is neutral and the variant model is still able to generate a more friendly response with "请您放心" (please do not worry) keyword. Intuitively, the variant model which is introduced with care-oriented information achieves better performance than the baseline model.

4.2 User Emotion Recognition

In this part, we focus on the effect of different models used in the emotion recognition task: the classification-based model and generation-based one. We will conduct experiments on the dataset CMCC and answer the question: are both models suitable to solve the emotion recognition problem?

4.2.1 Classification-based Model

We first treat the emotion recognition task as a mutli-label classification problem because a user utterance may contain multiple emotions, e.g., *complain about a problem* and *dissatisfied with buisiness rules* at the same time. Taking the pre-trained models bert-base-chinese[1] as the backbone, the classification model takes dialogue utterances X as input and predicts a binary result for each possible dialogue emotion:

$$P = Sigmoid(W(G(X))) \in R^N, \qquad (1)$$

where $W \in R^{d_b * N}$ is a trainable weight matrix, d_b is the hidden size and G is used to encode the input utterance to our representation. The model is trained with binary cross-entropy loss. The task uses the threshold of 0.3 to predict whether the emotion is triggered. We evaluate micro-F1 and macro-F1 scores on our dataset for the emotion recognition task.

4.2.2 Generation-based Model

The Generation-based model is consistent with the variant model in the previous section 4.1.1. The user emotion is generated conditioned on KB result which is queryed conditioned on predicted entity name and user intent. Micro-F1 and macro-F1 are aggregation methods for the user emotion recognition task. Specifically, micro-F1 score gives equal

[1] https://huggingface.co/bert-base-chinese

User Utterance	Response of Manual Customer Service	Prediction of Variant Model	Prediction of Baseline Model
用户：每天晚上一到这个时候啊到十二点过钟连看电视都看不到 User: Every night at this time, I can't even watch TV until after twelve o'clock.	客服：嗯抱歉先生现在连我的远程系统都看不了任何状态了_这边的话请您稍后再来电正在恢复当中恢复好了请您这边再来电好吗 Customer Service: I'm sorry sir, I can't even see any status of my remote system now. Please call back later. It is recovering. Please call again when it is restored.	系统：嗯好的那我这边的话呢马上给您反映到后台去给您核实处理一下 System: Okay, I'll report it to you i mmediately to the backstage, and check it out.	系统：嗯那您这个问题的话我这边帮您记录反馈到后台去给您处理一下_然后呢我们这边的话呢会尽快的给您回 System: I will help you record the feedback and send it to the background to handle it for you, and then we will call you back as soon as possible.
用户：帮我登记一下啊安排你们师傅过来看 User: Help me register and arrange for your master to come and see.	客服：诶好的 Customer Service: Oh, okay.	系统：好的那我这边的话呢已 经为您登记好了请您放心 System: Okay, I've already registered it for you, please don't worry.	系统：好的那我这边的话就帮您登记一下然后的话呢您就可以使用了 System: Okay, I'll help you to register, then you can use it.

Table 4: Responses generated from variant model and baseline model.

importance to each observation. When the classes are imbalanced, those classes with more observations will have a larger impact on the micro-F1 score. Thus the final micro-F1 score tends to hide the performance of the minority classes and amplify the majority. On the other hand, macro-F1 score gives equal importance to each class. This means that a majority class will contribute equally along with the minority, allowing macro-F1 to still return objective results on imbalanced datasets. As shown in Table 5, our experiments show that the generation_based approach can help us improve emotion classification performance on the imbalanced classes, from a classification_based baseline performance of 30.1% macro-F1 to 39.3%, an increase of 9.2 points.

Models	micro-F1	macro-F1
Generation-based	0.832	0.393
Classification-based	0.859	0.301

Table 5: Emotion recognition performance using two different models (the generation-based model and the classification-based one).

5 Conclusion

In this paper, we present CMCC, to date the largest human-to-human real-life dataset annotated with rich care-oriented information on top of MobileCS. We not only manually label each dialogue with comprehensive user emotion, customer service act and satisfaction annotations for various sub-tasks of multi-domain dialogue systems, but also further investigate approach to facilitate the research of care-oriented way via empirical experiments. In addition, the process of data annotation and visualization is described in detail. We also report the benchmark models and results of two evaluation tasks on CMCC, indicating that the dataset is a challenging testbed for future work. We will enrich the dataset annotations (e.g., solutions, external knowledge and API calls) from various aspects in future work. We hope it can bring more imagination and benefit future research in dialogue systems.

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 459–466. European Language Resources Association.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2021. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *arXiv preprint arXiv:2111.14592*.

Sai Muralidhar Jayanthi, Varsha Embar, and Karthik Raghunathan. 2021. Evaluating pretrained transformer models for entity linking in task-oriented dialog. *arXiv preprint arXiv:2112.08327*.

Hong Liu, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. 2022. Revisiting markovian generative architectures for efficient task-oriented dialog systems. *arXiv preprint arXiv:2204.06452*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Zhijian Ou, Junlan Feng, Juanzi Li, Yakun Li, Hong Liu, Hao Peng, Yi Huang, and Jiangjiang Zhao. 2022. A challenge on semi-supervised and reinforced task-oriented dialog systems. *arXiv preprint arXiv:2207.02657*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv preprint arXiv:1909.02188*.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. Risawoz: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 930–940. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5263–5276. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *arXiv preprint arXiv:2109.14739*.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1489–1503. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 665–677. Association for Computational Linguistics.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE*, 101(5):1160–1179.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Trans. Assoc. Comput. Linguistics*, 8:281–295.

Category	Examples	Frequency
通知(Inform)	嗯，我帮您看看您的手机有没有开通业务了，我先帮你查一查 Well, let me help you to see if your mobile phone has been opened for business, let me check for you first	37.72%
被动确认(passive confirmation)	对咱们这面办不了 Yeah, we can’t do it here	20.76%
问候(Greeting)	您好很高兴为您服务 Hello, glad to serve you	8.33%
主动确认(Active Confirmation)	您好感谢您耐心等待，有一个十元一百兆的安心包确定要取消是吗 Hello, thank you for your patience, there is a peace of mind package of ten yuan and one hundred trillion, are you sure you want to cancel it?	7.95%
询问(request)	二十四小时之内先生，一般都很快的，那个您是主要在省内用吗 Within 24 hours, sir, i t’s usually very fast. A re you mainly using it in the province?	6.64%
引导(Guide)	嗯请问什么其他可帮你吗先生 Well, what else can I help you with, sir?	6.05%
客套(Courtesy)	不客气已经帮您改好了稍后查看一下 You’re welcome, I’ve fixed it for you, check it out later	4.48%
建议(Suggest)	那建议您测试一下好吗 I suggest you test it	3.04%
其他(Other)	嗯 Um	1.42%

Table 6: Types, instances, and proportions of customer service acts.

Category	Examples	Frequency
抱歉(Apology)	您好，很高兴为您服务， 抱歉让您久等了 Hello, nice to serve you, sorry to keep you waiting	1.41%
安抚(Comfort)	哦，这个的话是可以使用的 ，这您放心 Oh, this one can be used, don't worry	0.58%
再见(Goodbye)	好，麻烦了，感谢来电再见 Okay, sorry for your troubles, thanks for calling, bye	0.50%
解释(Explain)	它是每天早上八点到晚上六点 之间办公的_就说现在已经下班 了明天早上八点以后才可以拨打 It works between 8:00 am and 6:00 pm every day. It means that it is already off work now and can only be called after 8:00 am tomorrow.	0.41%
否认(Deny)	不好意思，不是，那个假日流量 只是三天时间_并且_时候才可 以的 Sorry, no, that holiday traffic is only available for three days _ and _	0.36%
请求重复(Request To Repeat)	呃我没听清 uh i didn't hear	0.33%
同理心(Empathy)	你这个心情，我非常理解， 给您带来不便，是向您致 一下歉 I understand your feelings very much. I apologize for the inconvenience caused to you.	0.02%
赞同(Agree)	嗯对是的，那您说的没错 Yes, then you are right	0.004%

Table 7: Types, instances, and proportions of customer service acts.

Category	Examples	Frequency
抱怨某问题(Complain About A Problem)	就是在那个上网的时间网络老是出现那个网络异常怎么回事儿 It’s the time when the Internet is online, the network always has that network abnormality,what’s the matter?	77.14%
情绪较为激动(Emotionally More Agitated)	因为我觉得我这个要求不是很高的他们确实有有点做的过分 Because I don’t think my requirements are very high, they are indeed overdoing it.	11.10%
非常着急且态度恶劣(Very Anxious And Having A Bad Attitude)	没有啊_打电话了他给我说我办下，还什么办了个三十G的咋了，我说你这些人[UNK]你说话怎么这么_嘴里跑火车着呢 No, _ called and he told me that I would do it, and even a 30G package . why are you full of crap?	3.01%
对客服代表服务不满(Dissatisfied With Customer Service Representative Service)	尽快呀，快到什么时候啊_对呀，我想问下快到什么，四五天了耶，然后重点是我报修也报了三天之后也没个人给我打个电话啊 As soon as possible, it’s been four or fiv e days, and the point is that I appl ied for repairs for three days and no one called me.	2.82%
表示担忧或焦虑(Express Concern Or Anxiety)	六十多岁了我能不着急吗我这个 I’m in my sixties, can I be in a hurry?	2.35%
对业务规定不满(Dissatisfied With Business Rules)	不可能吧哪有这种霸王条款我不想用我_我取掉的话它为啥不让取 Impossible, how can there be such an overlord clause, I don’t want to use it, I’ll jus t cancel it, why not let me cancel	2.07%
执行某项操作有困难(Difficulty Performing An Operation)	咋咋个下载法我也搞不清楚 i don’t know how to download	1.03%
不太满意但不再追究(Not Satisfied But No Longer Pursue)	可以我希望你们后台人员无论处理出怎样怎么样的结果_可以在最短的时间内告知我 Yes, I hope that no matter what the result is from your backstage staff, you can let me know in the shortest possible time.	0.47%

Table 8: Types, instances, and proportions of user negative emotions.

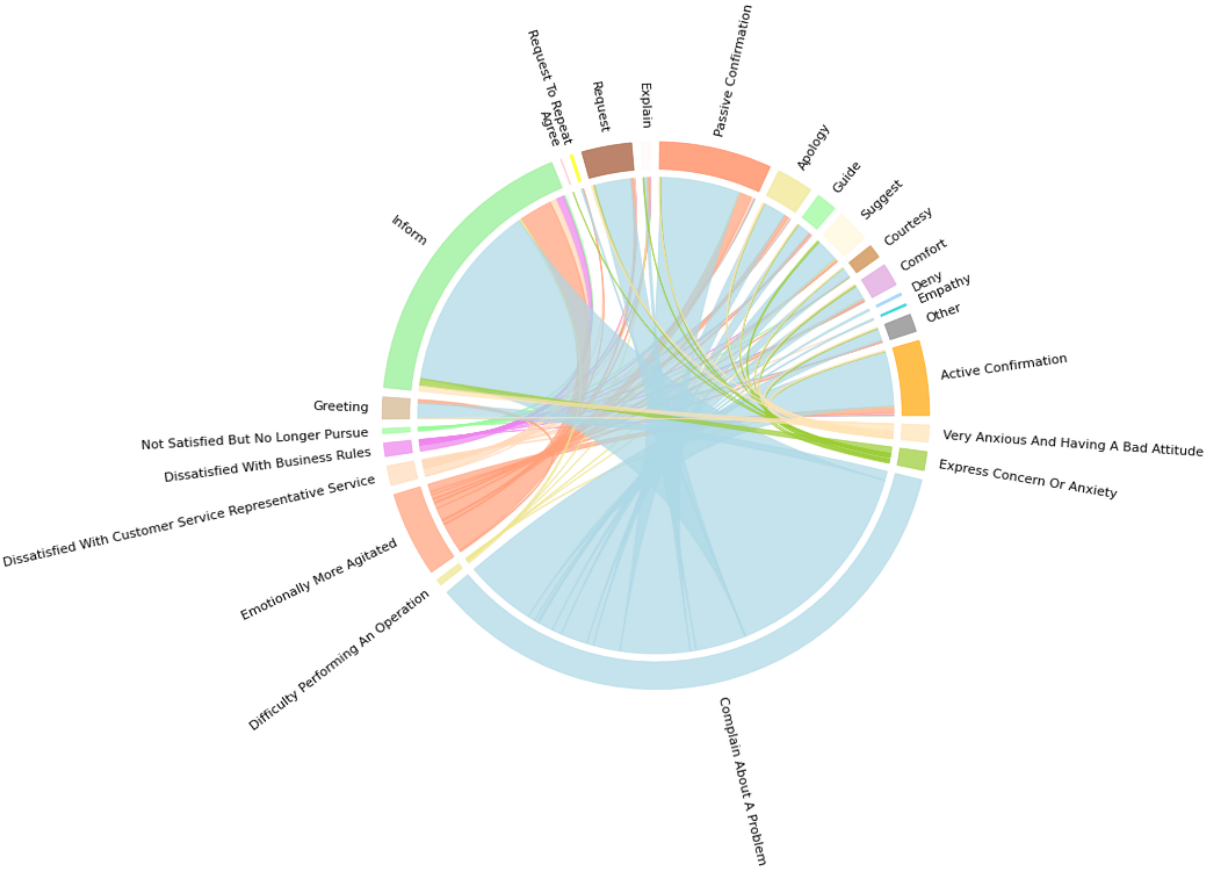

Figure 7: User emotion-customer service act conversion relationship chord diagram

Model	entity (F1)	emotion (Acc)
Stack-propagation Model	0.525	0.989
w/o user emotion	0.524	-
Baseline Model	0.382	-

Table 9: The joint performance on the stack-propagation model (Qin et al., 2019) using the CMCC dataset with or without emotion labeling.

Table 9 gives the result of the experiment comparison for entity extraction task. From results of the first two rows, we observe that without the emotion labels, simply incorporating the sequence labeling information, the entity extraction performance (micro-F1) drops slightly, which demonstrates that directly leveraging the emotion information can slightly improve the performance of the entity extraction task.