To Know What User Concerns: Conceptual Knowledge Reasoning for Task-oriented Dialogue Quality Estimation

Anonymous ACL submission

Abstract

Dialogue Quality Estimation (DQE) is crucial in assessing the effects of a conversational consultation system, which has wide applications in E-commerce and Social Media. In task-oriented scenarios, users usually seek personalized consultation about the target subjects they are concerned with rather than general knowledge commonly known by populations. It is essential to identify whether a dialogue solves the user's questions by task-oriented DQE. Existing studies mainly focus on analyzing dialogue semantics and user sentiment, neglecting to understand what the user is concerned about when requesting a consultation. It may cause fatal errors when the response is emotionally friendly but non-informative. In this paper, we propose a knowledge-enhanced DQE model named **CoReT**, which introduces the **Co**nceptual Knowledge **Re**asoning for **T**ask-oriented DQE. We first design a simple yet efficient entity linking and relation selection module enabling conceptual reasoning from a knowledge graph. Then, we propose a multi-turn textual encoder to capture the contextual information in dialogues. Finally, we introduce a knowledge enhancement module to fuse conceptual reasoning features into contextual embeddings to produce DQE results. For evaluation, we conduct experiments on two real-world datasets in e-commerce consultation systems, the results demonstrate the effectiveness and robustness of **CoReT** compared with the state-of-the-art baselines.

1 Introduction

Dialogue System (DS) is a human-computer interaction technology that aims to simulate natural conversations between humans. It has been widely used in many areas to provide intelligent assistants (Yan et al., 2017), customer service management (Cui et al., 2017), and automated question answering (Zhang et al., 2020). Especially in the e-commerce domain, intelligent assistants play a crucial role in supporting various customer-oriented services (Li et al., 2017; Ping et al., 2019), such as product consultation, complaints addressing, and feedback collection. The quality of DS is essential for user satisfaction and customer conversion.

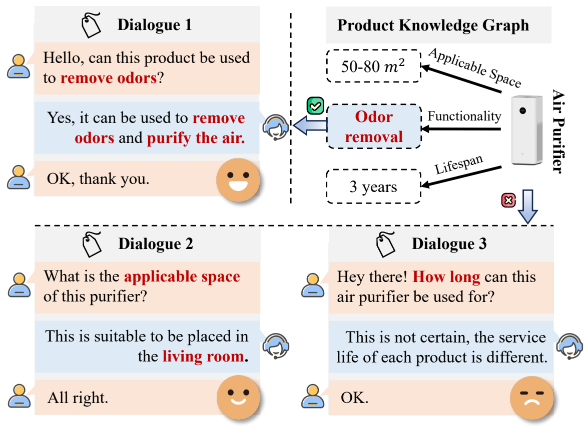


Figure 1: Three dialogues of air purifier. Customers are more satisfied when the server provides accurate answers to product inquiries.

However, task-oriented dialogue quality estimation (DQE) remains a challenging problem. It requires not only semantic understanding but also intent detection in textual dialogues, which makes it a complex and difficult task (Fan and Luo, 2020). Recently, task-oriented DQE (Song et al., 2019; Bodigutla et al., 2020; Cai and Chen, 2020) has become a crucial topic in dialogue system research. It has a natural label to measure the quality by user satisfaction. Existing works usually focus on intent detection by modeling user actions in each turn and ultimately fit them to user satisfaction (Sun et al., 2021; Deng et al., 2022b; Kim and Lipani, 2022), or address the semantic understanding by modeling the contextual information within the whole dialogue (Song et al., 2019; Feng et al., 2023). Such work seldom considers an important aspect in task-oriented dialogues, *the subject that the user is inquiring about*, which can provide valuable insights into the specific concerns of the user during the dialogue.

Figure 1 presents three dialogue cases in e-commerce consultation systems. The servers who offer accurate and informative answers to user's queries about the product are more likely to receive better evaluations. The satisfaction of users largely depends on whether the server effectively conveys the necessary information about the subject attributes. This involves the server's skills in critical information expression during the dialogue. Therefore, utilizing specific knowledge for extracting keywords related to subjects in dialogues is a necessary and pivotal work. For example, in the cases of these dialogues, we can capture the key terms in the dialogue by utilizing the triplet *<air purifier, applicable space, 50-60 square meters>* and *<air purifier, lifespan, 2-3 years>*. This information provides a conceptual understanding for servers to infer which aspects of the subjects the customer is concerned about and helps better evaluate the responses' quality.

Nevertheless, to infer the conceptual knowledge of subjects given a task-oriented dialogue still has to address the following challenges: (1) **Contextual entity matching**: How can we accurately link subject references in dialogues to specific entities in the conceptual knowledge graph? (2) **Differentiated knowledge reasoning**: How can we perform differentiated knowledge inference for various subjects to obtain concepts with different emphases?

To address these challenges, we propose **CoReT** to incorporate **Co**nceptual knowledge **Re**asoning of subjects in dialogues for **T**ask-oriented DQE. Specifically, it consists of three main components, (1) **Conceptual reasoning module:** We trained the TuckER (Balažević et al., 2019) on a large-scale e-commerce knowledge graph OpenBG (Deng et al., 2022a) for conceptual knowledge reasoning. To align inquired subjects with the knowledge graph, we utilize text and semantic similarity matching to link them to corresponding entities. A single subject may be linked to multiple entities, thereby obtaining broader knowledge related to the subject. Additionally, we leverage category-based relations for knowledge inference, enabling differentiated keyword extraction for different types of subjects. By employing knowledge reasoning, we obtain the subjects' attributes and concepts, which serve as the basis for extracting dialogue keywords. (2) **Hierarchical Text Mining Module:** To make full use of the information in multi-turn dialogues, we employ a Transformer encoder (Vaswani et al., 2017) to model the semantic actions of both turn-level and dialogue-level. This module enables the model to comprehensively capture the crucial information within each turn and the historical context across turns. (3) **Knowledge enhancement Module:** Based on the inquired subjects and extracted keywords, we employ parallel multi-head attention mechanisms to enhance the dialogue embeddings. This module aims to integrate conceptual reasoning features into contextual embeddings, enabling the model to achieve a deeper understanding of the informative and crucial aspects of the dialogue. In summary, our main contributions are as follows:

- Conceptually, we are the first to model task-oriented DQE by incorporating conceptual knowledge reasoning. Our model captures the attributes of specific subjects and the generalized concepts to make up for the lack of informativeness measurement in DQE.
- Technically, we propose CoReT, a knowledge-enhanced model that enhances DQE by combining conceptual knowledge reasoning with contextual mining.
- Experimentally, our model achieves the new SOTA on two real-world datasets. Additionally, we conduct experiments in multi-task and low-resource scenarios. The results indicate that CoReT shows robust and impressive performances in various settings.

2 Problem Definition

A dialogue can be represented as a sequence of turns $\mathcal{D} = \{(u_{u_1}, u_{s_1}), ..., (u_{u_T}, u_{s_T})\}$, where (u_{u_t}, u_{s_t}) represents the user's question and the server's answer in the t-th turn. A complete dialogue usually starts with a question or consultation about a specific subject. Task-oriented DQE requires not only understanding the informativeness of the response whether it addresses the question but also the sentimental reactions to customers' concerns. Therefore, the evaluation of the dialogue quality about how it meets customer satisfaction is summarized as follows: given a dialogue $\mathcal{D}$, the subject information $\mathcal{P}$, and the conceptual knowledge graph $\mathcal{G}$, the objective is to model an estimator $\mathcal{E}$ that accurately predicts the user satisfaction $\mathcal{Y}^s$ by the end of the dialogue session. This process

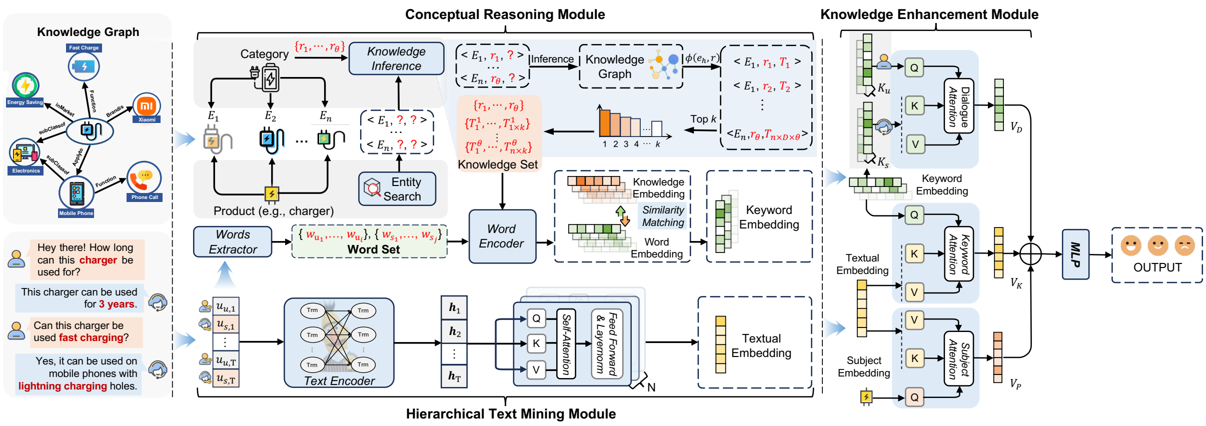


Figure 2: Overview of CoReT: (1) Conceptual Reasoning Module perform subject knowledge inference and extract the keywords from the dialogue; (2) Hierarchical Text Mining Module encodes the dialogue; (3) Knowledge Enhancement Module fuse the keywords, contextual representation and subject information.

can be formalized as:

$$\mathcal{Y}^s = \mathcal{E}(\mathcal{D}, \mathcal{P}, \mathcal{G}) \tag{1}$$

3 Methods

Figure 2 shows the architecture of our proposed model. It is composed of three parts: (1) **Conceptual Reasoning module**: We begin by pretraining a knowledge-inferring model on the knowledge graph. We link each subject to its corresponding entities as head entities to perform knowledge inference. Using the head entities and relations, we predict the tail entities on the inferring model, thereby obtaining key attributes and concepts for each subject. This enables us to extract the keywords from the dialogue. (2) **Hierarchical Text Mining Module**: Utilizing the utterance encoder and Transformer for turn-level and dialogue-level encoding to obtain dialogue contextual representation. (3) **Knowledge Enhancement module**: Applying attention mechanisms to fuse different representations with distinct focuses.

3.1 Conceptual Reasoning Module

To conduct unique knowledge inference for each subject, thus obtaining its attributes and concepts, we introduce the Conceptual Reasoning Module. It enables the model to effectively capture informative keywords in a dialogue.

3.1.1 Pretraining Knowledge Inferring Model

We select the TuckER (Balažević et al., 2019) model for knowledge graph inference. For a triple (e_h, r, e_t) in the knowledge graph, we define the embedding vectors for the head entity and relation as $\mathbf{v}_h \in \mathbb{R}^{d_e}$ and $\mathbf{v}_r \in \mathbb{R}^{d_r}$, respectively. We also introduce the parameter matrix $\mathbf{W} \in \mathbb{R}^{d_e \times d_e \times d_r}$ and the unified representation matrix $\mathbf{U} \in \mathbb{R}^{d_e \times M}$ that contains embeddings for all entities, where M denotes the number of entities. Given e_h and r, the inferring of tail entity is as follows:

$$\phi(e_h, r) = \mathbf{W} \times \mathbf{v}_r \times \mathbf{v}_h \times \mathbf{U} \tag{2}$$

Here, $\phi(e_h, r)$ represents the probability distribution of the predicted entities according to the head entity and relation: $(e_h, r, ?)$. We consider the entity with the highest probability as the predicted result and use the cross-entropy loss function to train the knowledge inferring model. In practical applications, we choose the top-K entities with the highest probabilities as potential tail entities. For each subject, we consider both the relations $\{r\}$ utilized in the inference and the resulting tail entities $\{T\}$ as the key attributes $\{attr_1, ..., attr_h\}$.

3.1.2 Matching Subject Entities & Selecting Relations

Before inferring the conceptual knowledge corresponding to the inquiry subjects, we have to map the subjects to the entity nodes in knowledge graphs. Considering the efficiency of subject entity matching between dialogues and knowledge bases, we propose a simple yet efficient method to match the subjects $\mathcal{P}$ mentioned in the dialogue to the corresponding entity E in the knowledge graph by judging whether their text is the same ($\mathcal{P} = E$) or their semantic cosine similarity is greater than a threshold ($\cos\_smi(\mathrm{BERT}(\mathcal{P}), \mathrm{BERT}(E)) > \tau$). To fully explore the relevant entities of a subject, we link each subject to multiple entities $\{E_1, ..., E_n\}$, using them as candidate head entities in the inference process. This approach allows us to leverage a broader range of knowledge and capture a comprehensive understanding of the subject.

To acquire the relations utilized during the inference process for different subjects, we design a statistical-based heuristic algorithm. For different categories of subjects, it retrieves and counts the relations that appear in the training set dialogues. Based on the frequency of occurrence, we ultimately select the top-θ most frequently occurring relations within each category.

3.1.3 Extracting Keywords in Dialogues

Based on the candidate entities matched to the subjects and corresponding relation sets describing the inferring aspects of differentiated subject categories, we utilize the key attributes $\{attr_1, ..., attr_h\}$ which contains both the relations and inferred tail entities to extract keywords in the dialogue. Then, we segment the dialogue utterances to obtain the word sequences for the user and the server, denoted as $\{w_{u_1}, ..., w_{u_i}\}$ and $\{w_{s_1}, ..., w_{s_j}\}$ respectively. A semantic matching mechanism is designed to calculate the similarity between the segmented words $\{w\}$ and the entity attributes $\{attr\}$ using the following formula:

$$\mathrm{Sim}(w, attr) = \mathrm{cos\_sim}(\mathrm{BERT}(w), \mathrm{BERT}(attr)) \tag{3}$$

The BERT model is employed to encode the words. If the similarity between a $(w, attr)$ pair exceeds a threshold τ, we consider the word w as a keyword. As a result, for each dialogue, we have the keywords $\{w_{uk_1}, ..., w_{uk_n}\}$ and $\{w_{sk_1}, ..., w_{sk_m}\}$ from user and server respectively.

3.2 Hierarchical Text Mining Module

To address multi-turn customer service dialogues, we obtain contextual representations that capture both the turn-level and the dialogue-level information in this module. Firstly, for turn-level context encoding, we utilize the BERT model as the text encoder. The user and the server utterances (u_{ut}, u_{st}) are concatenated as a complete sequence with a special token $[SEP]$ to feed the text encoder. Then the turn-level representation is calculated as follows:

$$\begin{aligned} \mathbf{h}_t &= \mathrm{TextEncoder}(u_{ut}, u_{st}) \\ &= \mathrm{BERT}([CLS]u_{ut}[SEP]u_{st}[SEP]) \end{aligned} \tag{4}$$

By calculating all $\mathbf{h}_t$ for each turn, we can obtain a set of turn-level representations $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_T\}$, where T is the number of turns in the dialogue.

To capture the inter-turn relationships and obtain an overall representation of the dialogue, we employ a dialogue-level encoder with L transformer layers. Similar to the turn-level encoder, we also add position embedding vectors to ensure the temporal information between different turns. The inputs are constructed as follows:

$$\mathbf{H}^{(0)} = \{\mathbf{h}_t + \mathrm{position\_encoding}(t) | 1 \leq t \leq T\} \tag{5}$$

where the position encoding is computed using sine and cosine functions following the approach in (Vaswani et al., 2017). Then we utilize the Multi-Head Attention mechanism (MHA) to fuse information across different turns.

The computation of the dialogue-level encoder that we employ is as follows:

$$\begin{aligned} \mathbf{H}' &= \mathrm{MHA}(\mathbf{H}^{(l)}, \mathbf{H}^{(l)}, \mathbf{H}^{(l)}) \\ \mathbf{H}^{(l+1)} &= \mathrm{FFN}(\mathbf{H}' + \mathbf{H}^{(l)}) + \mathbf{H}' + \mathbf{H}^{(l)} \end{aligned} \tag{6}$$

where l represents the l-th encoder layer. The Feed Forward Network (FFN) is also utilized to facilitate information flow between the encoder layers. Within the FFN, we use the ReLU function as the activation function between the two linear layers.

Finally, we retrieve the last layer's output of the encoder, denoted as $\mathbf{H}^{(L)} = \{\mathbf{h}_0^{(L)}, ..., \mathbf{h}_T^{(L)}\}$. From $\mathbf{H}^{(L)}$, we select the encoded information of the last turn $\mathbf{h}_T^{(L)}$ as the contextual representation of the whole dialogue, denoted as $\mathbf{D}_e$.

3.3 Knowledge Enhancement Module

In this module, we employ multiple multi-head attention blocks to effectively integrate keyword information, dialogue text information, and subject information. This integration allows for comprehensive exploration of key information within the dialogue from various perspectives.

Firstly, we introduce the keyword-level feature fusion. Given the obtained keywords $\{w_{uk_1}, ..., w_{uk_n}\}$ and $\{w_{sk_1}, ..., w_{sk_m}\}$ from the conceptual reasoning module, we utilize the BERT model to capture the conceptual features about subject keywords. The keyword representations are calculated as follows:

$$\begin{aligned} \mathbf{K}_u &= \mathrm{BERT}([CLS]w_{uk_1}[SEP]...w_{uk_n}[SEP]) \\ \mathbf{K}s &= \mathrm{BERT}([CLS]w_{sk_1}[SEP]...w_{sk_m}[SEP]) \end{aligned} \tag{7}$$

Then we apply a multi-head attention block to fuse the two vectors $\mathbf{K}_u$ and $\mathbf{K}_s$:

$$\begin{aligned}\mathbf{V}_D &= \text{DialogueAttention}(\mathbf{K}_u, \mathbf{K}_s, \mathbf{K}_s) \\ &= \text{MHA}(\mathbf{K}_u, \mathbf{K}_s, \mathbf{K}_s)\end{aligned} \tag{8}$$

where we obtain $\mathbf{V}_D$ to provide essential interactive information between the user and the server in a dialogue.

Furthermore, we specifically conduct information mining to target contextual representation with integrated keyword representation:

$$\begin{aligned}\mathbf{V}_K &= \text{KeywordAttention}(\mathbf{K}_{us}, \mathbf{D}_e, \mathbf{D}_e) \\ &= \text{MHA}(\mathbf{K}_{us}, \mathbf{D}_e, \mathbf{D}_e)\end{aligned} \tag{9}$$

where $\mathbf{K}_{us}$ is the integrated keyword representation by averaging $\mathbf{K}_u$ and $\mathbf{K}_s$.

Finally, an embedding layer is utilized to represent each unique subject $\mathbf{P}_e$. We then fuse the subject embedding with the contextual representation as follows:

$$\begin{aligned}\mathbf{V}_P &= \text{SubjectAttention}(\mathbf{P}_e, \mathbf{D}_e, \mathbf{D}_e) \\ &= \text{MHA}(\mathbf{P}_e, \mathbf{D}_e, \mathbf{D}_e)\end{aligned} \tag{10}$$

In summary, we obtain three representations in this module: $\mathbf{V}_D$, $\mathbf{V}_K$, and $\mathbf{V}_P$. Among them, $\mathbf{V}_D$ represents the key interaction between the user and the server. $\mathbf{V}_K$ represents the contextual representation enhanced by the subject-related keywords. And for $\mathbf{V}_P$, it represents the contextual representation autonomously enhanced by the subject information. These three types of representations enable us to explore the dialogue from multiple perspectives.

3.4 Learning Objectives

In the output module, we concatenate the representations $\mathbf{V}_D$, $\mathbf{V}_K$, and $\mathbf{V}_P$ and feed them into a multi-layer perceptron (MLP) for the final satisfaction estimation:

$$y_p^s = \text{softmax}(\text{MLP}(\text{Concat}(\mathbf{V}_D, \mathbf{V}_K, \mathbf{V}_P))) \tag{11}$$

where y_p^s represents the prediction results produced by our model. We utilize the cross-entropy loss as the objective function to train the proposed model:

$$\mathcal{L} = -y^s \log(y_p^s) \tag{12}$$

where y^s represents the ground truth. During the training process, the parameters of the pre-trained BERT model are also fine-tuned.

Table 1: The statistics of datasets.

Dataset	# Dialogues	# Aver. utterance	# Labels
JDDC	3,300	10.9	3
EDP	8,115	10.2	5

4 Experiments

4.1 Dataset & Evaluation metrics

To evaluate the effectiveness of our model, we conduct the experiments on an open-source dataset (**JDDC** (Chen et al., 2019)) and a real-world customer service dialogue dataset (**EDP**) collected from 300 users as volunteers. The dialogues are conversational consultations in e-commerce platforms. We use the user satisfaction score as the dialogue quality label for evaluations. The statistics of the datasets are shown in Table 1.

We consider the inquired products in the dialogue as the subjects, which cover a total of 602 distinct products in the EDP dataset. When there is no specific subject in a dialogue, we general tokens of the most frequent products within each category to serve as the possible subjects. For the conceptual reasoning, we utilize OpenBG (Deng et al., 2022a) as the product knowledge graph.

We adopted accuracy, macro-average precision, recall, and F1-score as evaluation metrics in our experiments, which are consistent with previous works (Ye et al., 2023; Feng et al., 2023; Deng et al., 2022b). The implementation code of our model is available here[1].

4.2 Baselines

We selected the following models as baselines:

HAN (Yang et al., 2016) It utilizes a two-level attention mechanism to capture information at different granularities in dialogues.

Transformer (Vaswani et al., 2017) It passes the dialogue embedding as input to the Transformer for representation learning.

BERT (Devlin et al., 2018) It embeds multi-turn dialogue by concatenating each turn.

Speaker (He et al., 2021) It leverages the modeling of interactions between the participants to recognize the dialogue acts.

CDCN (Li et al., 2020b) It employs a dynamic convolutional network as an utterance encoder to capture local information in dialogues.

[1] https://anonymous.4open.science/r/CoRe-USE-582C/

SG-USM (Feng et al., 2023) It adopts an attention mechanism to model task fulfillment in task-oriented dialogues.

USDA (Deng et al., 2022b) It employs a hierarchical Transformer structure for dialogue encoding.

ASAP (Ye et al., 2023) It applies the Hawkes process to model the intent changes in DQE tasks.

To ensure fairness, we employ the BERT model as the backbone encoder for all baseline models.

4.3 Parameter Settings

The base BERT consists of 12 Transformer layers and outputs a final dimensionality of 768. In the conceptual reasoning module, we set the cosine similarity threshold τ to 0.80. Each subject is mapped to 10 different entities and we select 50 relations for each subject category. For each inference, we select the top 5 tail entities as the reliable results. To verify the reliability of the reasoning model, we conducted tail entity prediction tasks on the benchmark of OpenBG. The model achieved impressive results with hits@10=71.0%, hits@3=59.0%, and hits@1=41.4%. For the utterances-level representation module, we set the number of Transformer layers to 3 and the number of attention heads to 8. The dropout probability of MLP is set to 0.1. We utilized the Adam optimizer and trained the model for 10 epochs.

4.4 Overall Performance

The results of DQE on the EDP and JDDC datasets are shown in Table 2. It can be observed that our model significantly and consistently outperforms all baselines on both datasets. Since JDDC lacks explicit subject information for each dialogue, we utilize the most frequent products within each category as the head entities used in the inference process for experiments conducted on this dataset. Even without explicit subject information, our approach still achieves a 1.5% improvement in F1-score compared to ASAP, which is the state-of-the-art model. Moreover, for EDP where subject information is available, our approach consistently outperforms other baseline models with an average 10% higher F1-score. This indicates that our modeling of subject information enables the model to focus on the crucial aspects that impact user satisfaction in dialogues, thereby effectively enhancing the accuracy of dialogue quality evaluation.

4.5 Ablation Study

In the ablation experiments, we conducted a total of five groups of experiments. Among them, “w/o text mining" refers to the removal of dialogue-level encoding from the Hierarchical Text Mining Module, and directly using the **[CLS]** token output by the BERT model for dialogue text representation. The results of ablation experiments on two datasets are shown in Figure 4. It can be observed that utilizing the transformer model for dialogue-level encoding is beneficial for the overall DQE. Furthermore, we also conducted ablation experiments on each of the three attention modules within the Knowledge Enhancement Module. The experimental results also confirmed that all three attention modules we designed positively contribute to the experimental outcomes. The final group of ablation experiments targeted the Conceptual Reasoning Module. In this experiment, we performed inference on the head entities using all relations in the knowledge graph, rather than solely relying on the most relevant and important relations for a particular subject category. From the experimental results, it is evident that if the relationships in the knowledge graph are not filtered, the knowledge reasoning module may introduce additional noise, thus negatively impacting the experimental results.

4.6 Robustness Assessment

4.6.1 Low-resource Scenarios

Due to the scarcity of high-quality training data for customer service dialogues in the real world, we also conducted evaluations of model performance in low-resource scenarios. In this experiment, we select USDA and ASAP, which performed well in the DQE task, as the baselines. We limited the number of dialogues used for training to 500, 1,000, 2,000, and 4,000, respectively, and conducted multiple experiments by randomly sampling from the two datasets for 5 times to test the stability of model performance. The results in the low-resource scenario are shown in Figure 3. The lines in the graph represent the average performance of the models under different sampling quantities, while the shaded areas represent the upper and lower ranges of model performance under different samplings. From the graph, it can be observed that our model performs better in terms of stability and performance by incorporating subject knowledge.

Table 2: Comparison of overall experimental results between CoReT and baselines. The **bold** demonstrates the best performance, while the second best performance is indicated with an <u>underline</u>.

Model	EDP				JDDC			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
HAN	56.3	39.3	39.9	38.2	58.4	54.2	50.1	52.0
Transformer	61.7	57.2	50.1	51.5	58.1	58.8	64.9	58.9
BERT	63.9	55.2	53.7	53.8	60.4	59.8	58.8	59.5
Speaker	59.7	55.7	47.9	50.0	63.4	61.8	62.2	61.9
CDCN	60.4	55.4	49.1	48.3	62.4	59.1	56.1	57.2
SG-USM	61.8	49.4	51.2	50.1	63.3	63.1	64.1	63.5
USDA	62.9	56.1	52.9	54.0	61.8	<u>62.8</u>	63.7	61.7
ASAP	<u>66.8</u>	<u>61.4</u>	<u>54.4</u>	<u>56.0</u>	<u>64.9</u>	62.3	<u>65.4</u>	<u>63.5</u>
CoReT	**71.3**	**62.5**	**61.4**	**60.5**	**65.4**	**63.7**	**67.3**	**65.0**

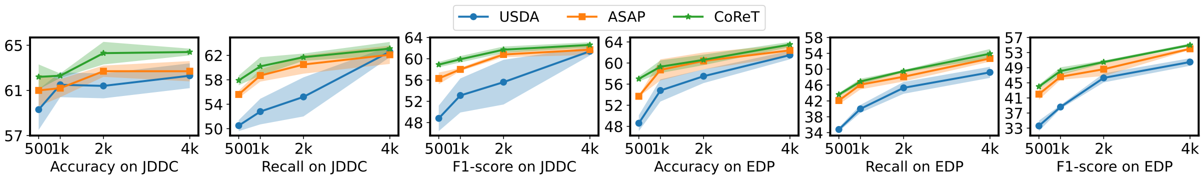
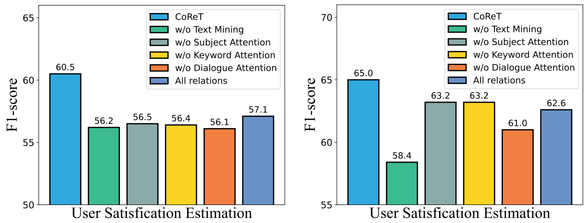


Figure 3: Experimental results in low-resource scenarios.



(a) EDP (b) JDDC

Figure 4: Ablation study on two datasets.

Table 3: Comparison of multi-task experimental results.

Model	DQE			DAR		
	Acc	Recall	F1	Acc	Recall	F1
JointDAS	58.5	55.1	55.4	63.4	43.6	41.1
Co-GAT	60.6	63.7	61.0	66.7	48.9	47.5
JointUSE	63.8	58.6	59.2	66.8	48.7	47.3
Speaker	<u>64.8</u>	60.1	60.2	65.8	47.0	45.8
USDA	63.0	**65.7**	<u>62.6</u>	<u>69.7</u>	<u>53.0</u>	<u>51.3</u>
ASAP	64.0	61.8	61.7	68.2	50.3	48.5
CoReT	**65.2**	<u>64.7</u>	**63.8**	**70.1**	**53.2**	**51.5**

4.6.2 Multi-task Scenarios

To evaluate the practicality of our model in real-world scenarios and enable the platform to track the real-time changes in user demands, we test our model on the JDDC dataset for the Dialogue Act Recognition (**DAR**) task. The DAR task aims to classify the user's intent for each turn in a dialogue. In our experiment, we conducted joint training of our model on both DQE and DAR tasks to obtain experimental results. To ensure fair performance comparison among different models, we selected three baseline models, namely JointDAS (Cerisara et al., 2018), Co-GAT (Qin et al., 2021), and JointUSE (Bodigutla et al., 2020), which are designed for multi-task learning. The experimental results in the multitasking scenario are shown in Table 3. Our model outperformed the other baselines in both tasks, indicating its excellent adaptability in turn-level user intent recognition tasks.

Table 4: Results Comparison of LLM and KG.

Knowledge Source	Acc	Recall	F1	Time
GPT-4	66.3	55.9	57.4	1.74s
KG	**71.3**	**61.4**	**60.5**	**0.53ms**

4.7 Discussion on LLM vs. KG

To investigate how well the LLMs perform in conceptual reasoning rather than a knowledge graph, we use GPT-4 to replace the knowledge graph(KG) for reasoning on subjects. The head entity and the corresponding relations are used to construct prompts of LLMs. The experimental results on the EDP are shown in Table 4. It can be seen that the knowledge obtained from KG is more conducive to accurately estimate the dialogue quality. Moreover, with cuda acceleration, each reasoning on the KG only takes 0.53ms, which is much faster than using LLM inferring APIs.

4.8 Case Study

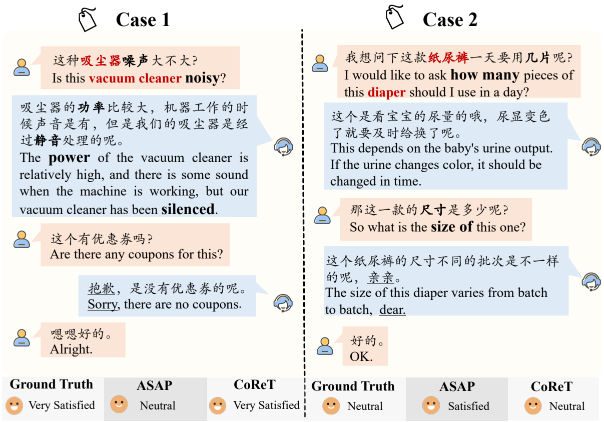


Figure 5: Two dialogue cases selected from the dataset. The mentioned products in the dialogue are highlighted in red, the product-related keywords are highlighted in **bold**, and the sentiment-related words are indicated with an underlined.

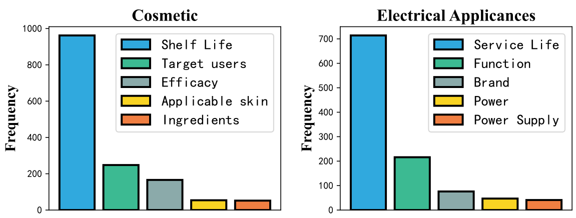


Figure 6: Relation selection of two product categories.

To further illustrate the working principle of our model in dialogue quality evaluation, We selected two dialogue examples from the dataset which are presented in Figure 5. It can be observed that models that rely solely on text analysis tend to focus more on emotion-related words in the dialogue. If the customer service attitude is friendly, it may receive a higher satisfaction rating, while neglecting whether the server answered the user's inquiry about products. However, our model can effectively address this issue and provide accurate evaluations.

We further select two examples to showcase the relations obtained by the algorithm for different product categories, as shown in Figure 6. Our algorithm can provide different relations based on different product categories and the selected relations are closely related to the corresponding subjects.

5 Related Work

5.1 Dialogue Quality Estimation

As intelligent dialogue systems are being widely applied in service platforms, Dialogue Quality Estimation is receiving more and more attention. Due to the high-cost and time-consuming characteristics of manual evaluation, recent works have increasingly relied on deep learning methods. Such methods follow two main lines. Some researchers focus on modeling user intents at the turn level to assess the dialogue quality (Kim and Lipani, 2022; Deng et al., 2022b). The other methods analyze at the dialogue level, focusing on contextual information the dialogue (Mendonca et al., 2022; Gupta et al., 2021; Ye et al., 2023). However, current methods seldom consider effectively modeling the subjects inquired about in task-oriented dialogues, which can imply the specific concerns of users. This limitation hinders the models' ability to determine whether the dialogue quality meets users' satisfaction.

5.2 Knowledge-enhanced Dialogue Understanding

In the field of natural language processing, knowledge is increasingly being used in various works to enhance the model's comprehension. Some works proposed to encode knowledge by pre-training and then apply the embedding to downstream tasks, such as Know-BERT (Peters et al., 2019b), KEPLER (Wang et al., 2021), etc. While some other studies directly use discrete triplets for knowledge enhancement. This type of work first retrieves relevant knowledge from the knowledge graph and then applies it to the model (Peters et al., 2019a; Li et al., 2020a; Gao et al., 2019). However, existing works on knowledge enhancement mostly focus on utilizing external knowledge to provide additional information, neglecting the exploration of the key content inherent in data.

6 Conclusion

In this study, we propose CoReT, which incorporates conceptual knowledge reasoning of subjects into dialogue quality estimation in multi-turn e-commerce dialogues. By capturing product-related concepts and attributes, our model can focus on the key content related to subject inquiries in the dialogue, which is also the part that concerns users the most. We conducted experiments on the EDP and JDDC datasets. The results demonstrated that our model achieved state-of-art performance on both datasets. To validate the impact of the knowledge on model robustness, we also conduct robustness tests in multi-task and low-resource scenarios. The results demonstrate that our model exhibits stable and reliable performance across various scenarios.

Limitations

While our model achieves the new state-of-the-art performance, it still has several limitations. Firstly, although we employ the hierarchical structure to model multi-turn dialogues, the abstractive extraction of key sentences in such dialogues still leaves an open issue when faced with long texts. Secondly, when there are no explicit subjects attached to given dialogues, we proposed to use a set of general subjects that frequently occurred in all dialogues and provide the model with commonsense reasoning comprehension corresponding to the e-commerce scenarios. Ideally, the framework should be able to understand what subjects the user is concerned about automatically from the whole contextual information rather than certain words at the beginning of the dialogues. Lastly, the evaluations are conducted in e-commerce consultation scenarios, the scalability of our model could be further analyzed in other domains.

Ethics Statement

In this work, we do not use any persona profiles or other user information for modeling. We do not collect or handle any personal data in our experiments. When applying our model to real-world applications, there is no need to capture any other information except the raw dialogues. The implementation code of our model is publicly available and compliant with anonymity standards.

References

Ivana Balažević, Carl Allen, and Timothy M Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590.*

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

Wanling Cai and Li Chen. 2020. Predicting user intents and satisfaction with dialogue-based conversational recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 33–42.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. *arXiv preprint arXiv:1807.05013.*

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2019. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. *arXiv preprint arXiv:1911.09969.*

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, system demonstrations*, pages 97–102.

Shumin Deng, Hui Chen, Zhoubo Li, Feiyu Xiong, Qiang Chen, Mosha Chen, Xiangwen Liu, Jiaoyan Chen, Jeff Z Pan, Huajun Chen, et al. 2022a. Construction and applications of open business knowledge graph. *arXiv preprint arXiv:2209.15214.*

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022b. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *Proceedings of the ACM Web Conference 2022*, pages 2998–3008.

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

Yifan Fan and Xudong Luo. 2020. A survey of dialogue system evaluation. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1202–1209. IEEE.

Yue Feng, Yunlong Jiao, Animesh Prasad, Nikolaos Aletras, Emine Yilmaz, and Gabriella Kazai. 2023. Schema-guided user satisfaction modeling for task-oriented dialogues. *arXiv preprint arXiv:2305.16798.*

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 429–437.

Saurabh Gupta, Xing Fan, Derek Liu, Benjamin Yao, Yuan Ling, Kun Zhou, Tuan-Hung Pham, and Chenlei Edward Guo. 2021. Robertaiq: An efficient framework for automatic interaction quality estimation of dialogue systems.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. *arXiv preprint arXiv:2109.05056.*

To Eun Kim and Aldo Lipani. 2022. A multi-task based neural model to simulate users in goal oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2115–2119.

Feng-Lin Li, Hehong Chen, Guohai Xu, Tian Qiu, Feng Ji, Ji Zhang, and Haiqing Chen. 2020a. Alimekg: Domain knowledge graph construction and application in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2581–2588.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017. Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2495–2498.

Jingye Li, Hao Fei, and Donghong Ji. 2020b. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 616–626.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. Qualityadapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019a. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019b. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Ng Lian Ping et al. 2019. Constructs for artificial intelligence customer service in e-commerce. In *2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS)*, pages 1–6. IEEE.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13709–13717.

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.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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

Fanghua Ye, Zhiyuan Hu, and Emine Yilmaz. 2023. Modeling user satisfaction dynamics in dialogue via hawkes process. *arXiv preprint arXiv:2305.12594*.

Weisheng Zhang, Kaisong Song, Yangyang Kang, Zhongqing Wang, Changlong Sun, Xiaozhong Liu, Shoushan Li, Min Zhang, and Luo Si. 2020. Multi-turn dialogue generation in e-commerce platform with the context of historical dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1981–1990.