Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5787–5796
Marseille, 20-25 June 2022

E-ConvRec: A Large-Scale Conversational Recommendation Dataset for E-Commerce Customer Service

Meihuizi Jia[*,†], Ruixue Liu[†], Peiying Wang[†], Yang Song[†], Zexi Xi[†], Haobin Li[†], Xin Shen[†], Meng Chen(✉)[†], Jinhui Pang[*], Xiaodong He[†]
[*]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[†]JD AI, Beijing, China
{jmhuizi24, pangjinhui}@bit.edu.cn
{liuruixue, wangpeiying3, songyang23, xizexi, lihaobin8, shenxin10, chenmeng20, xiaodong.he}@jd.com

Abstract

There has been a growing interest in developing conversational recommendation system (CRS), which provides valuable recommendations to users through conversations. Compared to the traditional recommendation, it advocates wealthier interactions and provides possibilities to obtain users' exact preferences explicitly. Nevertheless, the corresponding research on this topic is limited due to the lack of broad-coverage dialogue corpus, especially real-world dialogue corpus. To handle this issue and facilitate our exploration, we construct E-ConvRec, an authentic Chinese dialogue dataset consisting of over 25k dialogues and 770k utterances, which contains user profile, product knowledge base (KB), and multiple sequential real conversations between users and recommenders. Next, we explore conversational recommendation in a real scene from multiple facets based on the dataset. Therefore, we particularly design three tasks: user preference recognition, dialogue management, and personalized recommendation. In the light of the three tasks, we establish baseline results on E-ConvRec to facilitate future studies.

Keywords: Conversational Recommendation, Dialogue Corpus, User Preference Recognition

1. Introduction

Recently, Conversational Recommendation System (CRS), has attracted attention in the dialog community as it collects dynamic and interactive information from users' requirements and provides useful recommendations (Christakopoulou et al., 2016; Sun and Zhang, 2018; Chen et al., 2019a; Radlinski et al., 2019; Lei et al., 2020a; Lei et al., 2020b; Jannach et al., 2021; Gao et al., 2021; Zhou et al., 2021). Intuitively, a high-qualified dataset is essential to facilitate the development of CRS. To drive the progress of CRS development, there are some corpora proposed recently. In general, existing corpora are constructed in roughly two ways. Some studies generate the dialogues in a Wizard-of-Oz setting (Shah et al., 2018) by connecting two crowd-workers to engage in a chat session (Moon et al., 2019; Liu et al., 2020; Xu et al., 2020; Hayati et al., 2020; Liu et al., 2021b; Liao et al., 2021). Some other studies construct datasets from user review corpus (Fu et al., 2020) or item rating website (Zhou et al., 2020).

The task of conversational recommendation in E-commerce domain is far more complex compared with the above mentioned scenarios. Figure 1 shows a real-world E-commerce conversation. Several characteristics can be observed from the conversation. Firstly, users describe the sought products in broader terms and a casual way, with either explicit expressions (e.g. *14 inches*, *512 SSD*) or implicit words (e.g. *cheaper one*), resulting in difficulties in eliciting users' preferences. Therefore, accurately recognizing user preference words in casual utterances lays the foundation for providing high-quality recommendations. Secondly, customers usually proceed in a coarse-to-fine manner to gradually make their decisions during a conversation (Fu et al., 2020). Thus, customer service staffs need to conduct effective interaction with them to collect required information or make a recommendation. To attract users' interests, more attention should be paid to effective dialogue management during the conversation. Thirdly, there are massive personalized user profiles and product knowledge in the E-commerce domain. The auxiliary information makes it easy to trace connections from users to specific items and provides a high-quality recommendation for customers. For example, in Figure 1, customer service staffs tend to recommend a computer suitable for a student-user based on the information obtained from the user profile (eg. *16-25 years old*, *undergraduate*, etc). Hence, it is also a key problem to make a personalized recommendation by fully combining user profile, product knowledge, and dialogue context into consideration.

To bridge the gap between the complex problems in real scenario and the existing artificial CRS public datasets, in this paper, we present a real-world, large-scale, and informative E-commerce conversational recommendation dataset, namely **E-ConvRec**. It consists of 25,440 dialogues and 775,338 utterances derived from a leading Chinese E-commerce platform [1]. The dataset contains a wealth of information, including conversations, user profiles and product knowledge base. We hope that the natural dialogues and the en-

(✉) Corresponding Author

[1]https://www.jd.com/

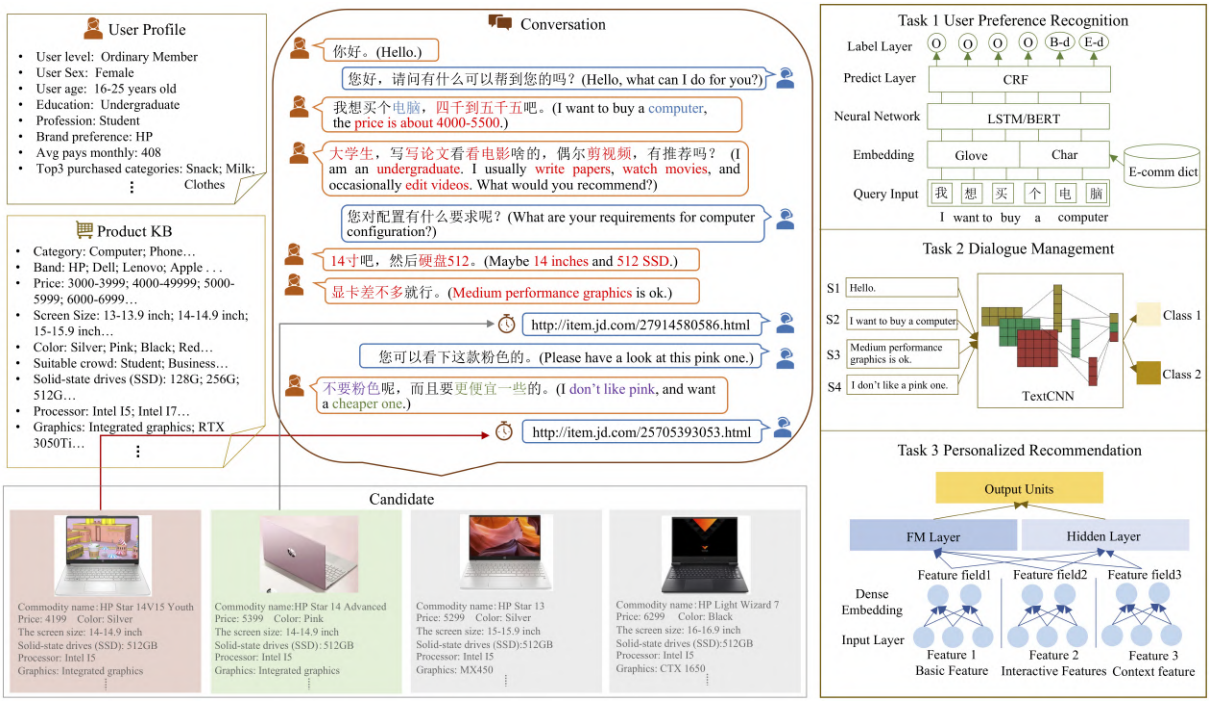

Figure 1: A dialogue example for E-ConvRec. This dataset provides conversation flows from real scenario of E-commerce, user profile and product KB to enrich recommendation. Moreover, three sub-tasks of user preference recognition, dialogue management and personalized recommendation are devised to facilitate research on CRS.

riched information in E-ConvRec open plenty of room for future studies on the conversational recommendation system.

Furthermore, we summarize three questions to address the aforementioned CRS paradigm: 1) What kind of products does the user prefer? 2) How should the CRS proceed the conversation by information collection or product recommendation? 3) Which product should CRS recommend to better attract users' interests? Specifically, we devise three meaningful tasks which are user preference recognition, dialogue management, and personalized recommendation with high-qualified annotated datasets. We conduct extensive experiments and provide baselines for the three tasks. Despite promising early results we get, E-ConvRec leaves ample scope to improve the CRS's performance.

In summary, the main contributions of this paper are listed as follows:

- We contribute E-ConvRec, a real-world, natural, and informative dataset from the E-commerce domain, which consists of 25k dialogues, user profiles, and product knowledge base.
- We design three worth investigating tasks to explore conversational recommendation in the real scenario from multiple facets, supported by the dataset.
- We conduct extensive experiments and provide several baseline results for three tasks to facilitate future research. The corpus and the annotated subtasks will be released soon.

2. Related Work

CRS is a recommendation system that elicits the dynamic preferences of users and takes actions based on their current needs through real-time multi-turn interactions (Gao et al., 2021). The growth of this field has been consistently supported by the development of novel datasets. We present a detailed comparison of E-ConvRec with existing datasets in Table 1.

Based on various resources of data collection, existing corpora can be roughly divided into two categories: 1) corpus generated by crowd-sourced workers and 2) dataset constructed from reviews or item ratings. The former one collects human-to-human and human-to-machine conversation data by recruiting crowd-sourced workers to interact in real-time under pre-defined search or recommendation settings. There are crowdsourcing sites, such as Amazon Mechanical Turk (AMT) [2], where researchers can find participants to accomplish their data collection task (Li et al., 2018; Kang et al., 2019; Moon et al., 2019; Liu et al., 2020; Xu et al., 2020; Hayati et al., 2020; Liu et al., 2021b; Liao et al., 2021). The latter is constructed in an automatic or semi-automatic manner. Researchers utilize the real data records or reviews from popular review websites (e.g., Douban Movie, Amazon, Facebook) and simulate the recommendation scenario to

[2]https://www.mturk.com/

Dataset	#Dialogue	#Utterance	#Domain	Language	Source	Extra Info
Facebook_Rec	1M	6M	Movie	EN	item rating	-
ReDial	10,006	182,150	Movie	EN	crowd-sourced	KB
GoRecDial	9,125	170,904	Movie	EN	crowd-sourced	User data
OpenDialKG	15,673	91,209	Movie, book, etc.	EN	crowd-sourced	KB
DuRecDial	10,190	155,447	Movie, food, etc.	ZH	crowd-sourced	User data, KB
TG-ReDial	10,000	129,392	Movie	ZH	item rating	User data
MGConvRex	7.6K+	73K	Restaurant	EN	crowd-sourced	User data
INSPIRED	1.001	35,811	Movie	EN	crowd-sourced	-
DuRecDial 2.0	16,482	255,346	Movie, music, etc.	EN-ZH	crowd-sourced	-
MMConv	5106	39,759	Travel	EN	crowd-sourced	User data,KB
COOKIE	-	11.6M	E-commerce	EN	item review	KB
E-ConvRec	25,440	775,338	E-commerce	ZH	natural dialogue	User data, KB

Table 1: Comparison of E-ConvRec with other Conversational Recommendation Datasets.

construct the dataset. (Dodge et al., 2015; Zhou et al., 2020; Fu et al., 2020).

During data construction, apart from dialogue information, most datasets provide additional information to improve CRS's performance. OpenDialKG (Moon et al., 2019) imports knowledge graph sources from Freebase, aiming to model of dialogue logic by walking over the knowledge graph. GoDialKG (Moon et al., 2019) and TG-ReDial (Zhou et al., 2020) import user data to capture human-level reasoning for the personalized recommendation. DuRecDial (Liu et al., 2020) leverages multi-type of dialogues in conversation recommendation and DuRecDial 2.0 (Liu et al., 2021b) prepares bi-lingual corpus for the task. From the perspective of the application domain, most of the corpora such as ReDial (Li et al., 2018) and INSPIRED (Hayati et al., 2020) focus on movie recommendation. MGConvRex (Xu et al., 2020) concentrates on restaurant booking and MMConv (Liao et al., 2021) presents multi-domain conversation during traveling. However, in most of the mentioned domains, the recommendation is quite straightforward. This is largely due to the fact that, during movie or restaurant recommendations, the user's intention is usually under-control and their requirements can be easily defined in limited aspects. Compared with them, the conversational recommendation in the E-commerce domain is quite different. Faced with millions of customers and thousands of different categories of products, the CRS not only needs to interpret various expressed requirements from users but are also required to bridge the gap between user's description and tens of attributes information corresponding to each individual product. Even though COOKIE (Fu et al., 2020) takes the first step to present an E-commerce recommendation dataset, it only covers on four categories of products, which limits the complexity in this domain.

3. Dataset Collection and SubTasks Annotation

E-ConvRec is constructed from the real scenario application in the E-commerce domain. In this section, we will introduce the detailed information for data collection and annotation procedures for sub-tasks.

3.1. Data Collection

Our data sources mainly include the following three parts: 1) The dialogue flows from online E-commerce conversation; 2) The personalized user profile; 3) Knowledge base for products. We will introduce them respectively in the following part.

Dialogue Flows. We first collect the dialogue dataset which contains conversations on pre-sales topics between users and customer service staff in an E-commerce scenario. We pre-select the conversations with a high intention of placing an order from a broader set of dialogues. After crawling, we de-duplicated the raw data, desensitized and anonymized private information. As illustrated in Figure.1, conversation collected from the real-world application contains more linguistic variety with natural expressions, and users involved tend to present more complicated requirements compared with the synthesized corpus.

We also analyze the number of sessions, words, and average turn to give an overview of the conversation dataset. As illustrated in Table 2, we can see that, our conversation dataset contains more than 25k sessions, including 32k cases and 775k utterances in total. Besides, the number of turns ranges from 2 to 100 for each session, with an average of 12. Figure 2 describes the histogram of dialogue length in the dataset. We only present dialogue with less than 30 turns for space limitation. It illustrates that most conversations are between 3 to 12, and the session of 7 turns has the largest portion. It indicates that, in the real application, users may be impatient, and professional customer service staffs need to make product recommendation in an appropriate timing, which is also a challenge for CRS.

User Profile. User profile plays a critical role in the personalized recommendation system as it encourages the CRS to make decisions tailored to each individual user's interest without requiring the user to make an explicit query (Zhang and Koren, 2007; Massari, 2010; Ni et al., 2018). Thus, we collect the user profile from the pre-processed user profile library. During information processing, to protect user's privacy, we anonymize the username, delete the phone number and

Total cases	32,609
Total sessions	25,440
Total turns	305,441
Average turns per session	12
Max turns	100
Min turns	2
Total utterances	775,338
Max utterances	180
Min utterances	3
Total words	6,782,956
Average words per utterance	8.7

Table 2: Session statistics.

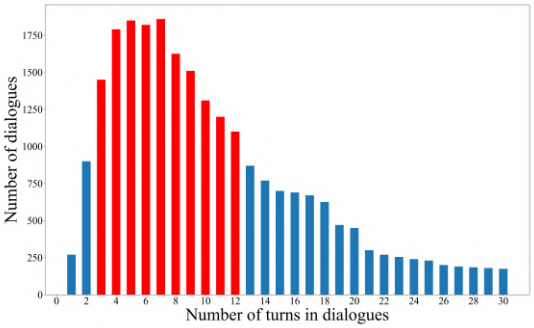

Figure 2: The distribution of dialogue with respect to the number of dialogue turns.

detailed address information, and remove all fields that related to the user identification. We further convert their identifications into string of 10 random characters in our dataset.

In general, 20 different types user profiles are provided. The user profiles are collected in two ways: some of the attributes are collected from their personal information during their registrations, such as *user level, sex, age, marital status, education and profession*, etc. The others are analyzed based on their historical shopping activities such as *brand preference, average payment per month, purchasing power, Top 3 purchased categories with the largest values in last three months*, etc. Detailed examples can be found in Figure.3. It's observed that, during popular item recommendation, the user profile such as *average payment monthly* may play a key role as it provides insights for user's preference on product price.

User Profile	User-level	User sex	User age	Avg pay monthly	Top 3 order recently
	Ordinary Member	Female	16-25 years old	103.4	phone bill, snack, milk
	Golden Member	Male	36-45 years old	1566.0	phone, earphone, waist support

Figure 3: Examples of user profile.

Product Knowledge Base. Product Knowledge Base (KB) provides abundant information about a product. Recently study (Fu et al., 2020) also proves its effectiveness to CRS with its explainability and transparency, which easily bridges the gap between user's description and product. Inspired by this, we crawl KB information of products mentioned in the conversation from the E-Commerce platform. To enrich information variety, we also collect knowledge from other products in the same category which are sold in the same online shop. Specifically, the extracted KB contains the product name, category, product title and various attributes and their values (e.g, *screen size:5.7 inch, Color:Silver*). Figure 4 shows two examples of product KB.

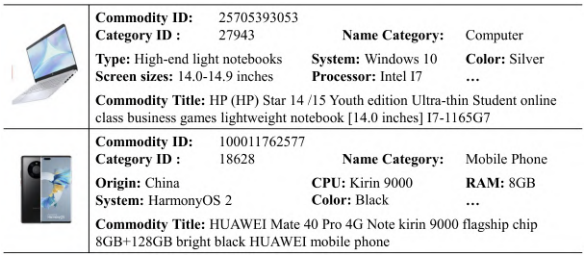

Figure 4: Examples of Product KB.

Table 3 shows the coverage of our provided user profile and product KB. The extracted KB covers 118k product items and the user profile provides anonymized information for 24k users.

Total items	118,086
Total users	24,358
Users without the profile	348
Items without the KB	21,907

Table 3: User and product statistics.

3.2. Task Formulation

As demonstrated in Figure 1, customers often proceed in a coarse-to-fine manner to gradually make their decisions during conversation flows. As they often initiate queries by describing the sought products in broader terms, e.g., category or brand name (Fu et al., 2020). As the dialogue goes on, CRS gradually grasps the users' specific requirements and preferences referring to the relevant products to be recommended. Naturally, we form them into three tasks: **User Preference Recognition** focuses on eliciting as many user preferences as possible; **Dialogue Management** tends to estimate the dialogue policy at per conversation step. Whereas **Personalized Recommendation** focuses on decision making tailored to each individual user's interest.

User Preference Recognition. The preference words in the query can depict the user's preference for an item. And then, customer service staffs match appropriate items catering to user needs and recommend them to the user. Therefore, recognizing the preference words is an indispensable part of the conversation recommendation.

Users' utterances
你们家有没有小女生用的那种口红? (Do you have the lipstick that girls wear in your shop?)
我不喜欢暗红色的。 (I don't like dark red.)
比之前的那个颜色淡吗? (Is it lighter than the previous one?)
请问有比Iphone 性能好的安卓手机吗? (Is there any mobile phone with an Android system that is better performance than iphone.)
我不要黑色，我要亮色的。 (I don't want a black one, I want a bright color one.)
麻烦你重新推荐一个低于300的手机，不要贵的。(Please recommend another mobile phone with a price lower than 300. I don't want a too expensive one.)

Table 4: Examples for user preference words annotation: descriptive words, category words, comparative words and negative words.

We hire three crowd-sourced workers who are familiar with E-commerce customer services, to help us annotate the data. The annotators are requested to tag the user preference words from queries. We provide the product knowledge base and category lexicon for workers as references for the annotators. Generally, product attributes and values, usage scenarios, user groups, brands, and categories are considered as preference words. We define four kinds of preference words, including descriptive preference words, category words, negative preference words (i.e. *don't want black*), and comparative preference words (i.e. *less than 300*). Table 4 illustrates a detailed example from the annotation. After the data annotation, we merge the instances from three crowd-sourced workers to obtain a diverse and high-quality preference words corpus. We assign 9k sessions to each crowd-sourced worker and collect 1k cross-annotated sessions. We follow the previous works (Bowman et al., 2015; Chen et al., 2019b) to employ the Fleiss Kappa (Fleiss, 1971) as an indicator, where Fleiss $\mathcal{K} = \frac{\bar{p}_c - \bar{p}_e}{1-\bar{p}_e}$ is calculated from the observed agreement $\bar{p}_c$ and the agreement by chance $\bar{p}_e$. We obtain a Fleiss $\mathcal{K} = 0.87$, which indicates strong inter-annotator agreement. We acquire 67k+ descriptive words, 137k+ category words, 133 comparative words and 2k+ negative words. In the 25k+ sessions, 91.91% of sessions contain descriptive words, 99.97% include category words, 0.51% involve comparative words and 7.05% have negative words.

Dialogue Management. Empowered by real-time interactions, CRS can directly acquire users' needs. After gathering users' preferences well enough, the system should make the proper recommendation at the golden time, otherwise, users will lose their patience. Accurately predicting recommendation timing can greatly improve the user experience. Thus for the dialogue management, we specifically focus on the task of recommendation timing prediction (Timing to ask user's preference proactively is also effective dialogue policy, however, we leave it into future work due to space limitation). According to statistics, there are totally 23,932 sessions which contain utterances with the positive recommendation timing. In addition, the intention of user can be helpful for determining whether or not to recommend items to the user. To facilitate the research in the future, we also label the intention for each query in the dialogues with a in-house intent classifier of E-commerce domain. The classifier contains five intents, and it's trained with BERT (Devlin et al., 2018). Table 5 shows the distribution of 5 intentions (The other intention includes Time for shipping, Invoice Policy, Usage Consultation, etc.). The classification accuracy reaches 92% on the test set of intention dataset, indicating the quality of the classifier.

Intent	Data distribution
Inform Preference	28.25 %
Request Attributes	22.35%
Don't care	6.45 %
Confirmation	10.35 %
Other	32.60 %

Table 5: Distribution of user's intents in E-ConvRec.

Personalized Recommendation. On the E-commerce application, we naturally assume that if the customer eventually purchases the item we recommend, it indicates a successful recommendation. Therefore, the task of conversational recommendation is designed to judge whether the user will buy a candidate product based on user profile, product KB and conversation content (context before the recommendation moment). Personalized recommendation is formulated as a ranking task. During data construction, we label the purchased product related to the conversation as positive sample (ground-truth). We also mine at most 30 hard negative samples, including the products which appeared within the conversation but not purchased by the user, or those of the same category and sold in the same online shop. Figure 1 illustrates an example from our data annotation. Utterances in the red and blue box indicate the query from the user and the answer from the customer service staff respectively. The answer annotated with *clock icon* represents a proper recommendation timing. Whereas the product marked in the pink color indicates the effective recommendation. In general, our E-ConvRec is a large-scale, natural, and informative dataset collected from the E-commerce platform, which also contains high-quality manual annotations.

4. Experiments

To evaluate the validity of our E-ConvRec dataset from multiple facets, in this section, we conduct extensive experiments for the three tasks mentioned in Section 3.2. Next, we introduce the experiment setup and experimental results for each task.

4.1. User Preference Recognition

User preference recognition can be formulated as a sequence labeling task, similar to the named entity recognition (NER). Compared with English NER, Chinese NER is more challenging as the mainstream approaches are based on characters and without word

segmentation. Recently, the lattice structure has been proved to be an effective structure as it utilizes the word boundary information and avoids the error propagation from pre-processed word segmentation (Zhang and Yang, 2018). Here, we verify the mainstream Chinese NER models enhanced by the auxiliary word information.

Data Preparation. We select 26k high-quality queries with annotated preference words from 25k dialogues in E-ConvRec. We divide the data into training, validation, and test set in the ratio of 8:1:1. To leverage auxiliary word information, we collect a lexicon (i.e. **E-comm dict**) in E-commerce domain with the vocabulary size of 722k based on the product KB mentioned in Section 3.1. As comparison, we also leverage the public CTB [3], an open-domain dictionary with 700k words, to examine the lexicon-enhanced sequence labeling methods.

Baselines. Hence, we select several lexicon-enhanced NER models as the baselines:

- **LSTM+CRF** (Huang et al., 2015) - It is a classical baseline for NER. We use the BMES schema as tag set and integrate extra word boundary information into the embedding layer. In this way, the character representation can be augmented with the embedding of its corresponding words.
- **Simple-Lexicon** (Ma et al., 2019) - This model integrates the lexical information with a soft lexicon mechanism. By categorizing the matched words and condensing the word sets, it captures the matched lexicon features.
- **Multi-Digraph** (Ding et al., 2019) - In this work, a neural multi-digraph model is proposed to learn how to combine the gazetteer information and resolve conflicting matches with context information.
- **FLAT** (Li et al., 2020) - This model converts the lattice structure into a flatten sequence. Equipped with Transformer and well-designed position encoding, FLAT can fully leverage the lattice information during sequence labelling and present an excellent parallelization performance.
- **LEBERT** (Liu et al., 2021a) - This work proposes Lexicon Enhanced BERT for Chinese sequence labeling, which directly injects lexicon information into Transformer layers in BERT with a Lexicon Adapter.

Experimental Results. We use F_1 score as evaluation metric and present the results in Table 6. Due to the various speaking habits in real-world dialogues, there are plenty of colloquial expressions involved in our dataset, making preference words recognition a challenging task. Table 6 demonstrates the performance of state-of-the-art NER models on this task. It's observed that LEBERT and FLAT outperform other models by utilizing the lexicon information in a more effective approach. Meanwhile, Table 6 shows the contribution of different dictionaries. Compared with CTB, the domain-specific dictionary **E-comm dict** helps most of the models obtain further improvements, including **LSTM + CRF**, **Multi-Digraph** and **FLAT**. And **FLAT** obtains the best performance in this task. It also suggests that how to construct a high-quality in-domain lexicon would be an important research topic in the future.

Model	w/ CTB dict	w/ E-comm dict
LSTM + CRF	74.00	**75.29**
Simple-Lexicon	76.30	75.89
Multi-Digraph	76.37	**77.40**
FLAT	76.60	**79.24**
LEBERT	78.91	78.53

Table 6: Evaluation results of user preference recognition.

4.2. Dialogue Management

Data Preparation. We sample 87,270 turns of utterances for recommendation time prediction, in which 39,254 are positive samples. We randomly select at most two utterances within the same session as the negative samples. The training, validation, and test set are split in an 8:1:1 ratio.

Baselines. We formulate the recommendation timing prediction as a binary text classification task. Therefore, we select several representative methods for text classification as the baseline models.

- **TextCNN** (Kim, 2014) - In this model, the convolutional neural network (CNN) is applied to text classification. Multiple kernels of different sizes are used to extract the salient information in sentences.
- **TextRNN** (Liu et al., 2016) - This model adopts a recurrent neural network (RNN) for text classification. The structure of the model is flexible and can be replaced with various components.
- **TextRCNN** (Lai et al., 2015) - This model replaces the CNN module into TextCNN with an additional RNN layer to acquire context information and reduce noise. In addition, the maximum pooling layer is used to capture the important parts of the text.

Specifically, we train GloVe (Pennington et al., 2014) vectors with a large size of data collected from dialogues and item titles. The pre-trained character embedding from Glove serves as the initial representation for each character token in the sentence. We implement above baselines based on the open-source classification toolkit[4].

[3] https://ai.tencent.com/ailab/nlp/en/embedding.html

[4] https://github.com/Tencent/NeuralNLP-NeuralClassifier

Method	Precision	Recall	F1
TextRCNN	**72.61**	48.44	58.11
TextRNN	69.86	57.78	63.25
TextCNN	70.10	58.62	63.85
TextCNN+Intent	72.08	**59.73**	**65.33**

Table 7: Recommendation timing prediction evaluation results.

Model	Feature	AUC	T@1	T@5	T@10
DeepFM	BF	69.77	15.65	44.70	61.94
	BF+IF	80.19	33.56	63.85	76.28
	BF+CF	74.53	22.07	54.80	70.93
	BF+IF+CF	**83.17**	37.06	**70.20**	**81.46**
FGCNN	BF	70.73	16.87	45.50	62.81
	BF+IF	78.50	35.44	63.31	74.87
	BF+CF	73.63	22.02	53.83	69.39
	BF+IF+CF	80.82	**37.28**	68.24	78.74

Table 8: Conversational recommendation evaluation results."T@1" stands for Top@1.

Experimental Results. We adopt precision, recall and F_1 as evaluation metrics. Table 7 shows experimental results on four baseline methods. The three models present comparable results on our dataset. Intuitively, the intention of user should be an indicator to the recommendation timing. To investigate the contribution of user's intent information, we append the auxiliary intent feature mentioned in Section 3.2 into **TextCNN** and compare its performance with other methods. Experimental result shows that intent information can bring further improvement.

4.3. Personalized Recommendation

We formulate the personalized product recommendation as a click-through rate (CTR) prediction task. The mainstream approach is to extract various features and catch deep feature interactions with deep neural networks. Here, we first introduce the data preparation.

Data Preparation. This task requires rich information from users, products, and dialogues. Therefore, we integrate three different types of features for comparison. The basic features (**BF**) include user features and product attribute features. Out of the 20 types of user profiles, we select 14 most relevant features to the recommendation task. We also select top 30 high-frequency attributes from the product KB and cover at most 500 high-frequency attribute values as features.

The product discussed in the current conversation should reflect the user's interest to some extent. Based on this motivation, three interactive features (**IF**) are involved between products mentioned in the dialogues and candidate products. The first feature is used to judge whether the candidate product appears in the conversation. If so, the corresponding feature is 1. The second interactive feature is the average attribute similarity of the product mentioned in the dialogue and the candidate product. (The similarity score is defined as the number of the same attribute values between two products divided by the number of all different attributes). The third interactive feature is the average Jaccard similarity (Jaccard, 1912) calculated from the title between products in the dialogue and the candidate product.

The context feature (**CF**) consists of two parts: we first calculate average cosine similarity between the preference words and the candidate product attribute values with Glove word embedding. Then we apply BERT (Devlin et al., 2018) to encode the utterance containing the preference words in the dialogue and the title of the candidate product and calculate average the cosine similarity between them. Both two features measure the similarity between dialogue context and product candidates.

We sample 1,073,216 data samples from the corpus. After filtering data without user profile and product KB, the number of total samples is adjusted to 876,335. The number of positive samples is 30,891, and all others are the negative samples (for each positive sample, there are nearly 30 negatives). We divide the data into training, validation, and test set on a scale of 8:1:1.

Baselines. We adopt follow two deep CTR models as our baselines:

- **DeepFM** (Guo et al., 2017) - This model combines a Factorization Machine (FM) with a neural network to learn both low-order and high-order feature interactions, which avoids artificial features being injected into the shallow part of the model.
- **FGCNN**(Liu et al., 2019) - This model consists of feature generation and deep classifier. Feature generation leverages CNN to generate local patterns and recombines them to generate new features. Deep classifier learns interactions between the raw features and new generated features and makes prediction.

Experimental Results. We adopt AUC (Area Under ROC) and Top@K as metrics to evaluate the model. Table 8 shows that the performances of both **DeepFM** and **FGCNN** are improved significantly after combining with more features, indicating the different kinds of features are complementary. Meanwhile, Table 8 also demonstrates the contribution of each kind of feature in a cumulative way. In general, **DeepFM** with BF, IF, CF features obtains the best performance in this task.

4.4. Case Study

To further explore the performance of different models for personalized recommendation, we present two cases on this task. As shown in Fig.5 (a), with only BF and CF, **DeepFM** gives a recommendation with high price which exceeds the user's expectation (the *average pay monthly* is 6,072 RMB in user profile). Whereas, given only basic features (BF) and interactive feature (IF), the second model recommends the items with the

User profile :	User level	User sex	User age	Avg pay monthly	Top3 purchased recently
	Golden Member	Male	26-35 years old	RMB: 6,072.98	watch/pen/earphone
Q:	您好。(Hello.)				
A:	您好，请问有什么可以帮助您的吗？(Hello, what can I do for you?)				
Q:	你们店里有什么手机，推荐一下？(I want to buy a mobile phone, what do you recommend?)				
A:	请问您对性能有什么需求么？(What are your requirements for performance?)				
Q:	没什么特别的需求，我不打游戏，最好是照相清晰的，内存要大于200G。(I don't have any special requirements. I don't play games. It's better to take clear pictures and the memory should be more than 200G.)				
A:	https://item.jd.com/10036452410557.html				
Q:	颜色我要深色的，不要金色的。(I would like a dark color, not golden.)				
Model	DeepFM+BF+CF	DeepFM+BF+IF	DeepFM+BF+CF+IF	Ground truth	
Recommend result					
	Price: 7488 RAM: 512G Color: Black Camera: 50 million pixels	Price: 6,488 RAM: 256G Color: Golden Camera: 50 million pixels	Price: 6,458 RAM: 256G Color: Black Camera: 50 million pixels	Price: 6,458 RAM: 256G Color: Black Camera: 50 million pixels	

(a)

User profile :	User level	User sex	User age	Avg pay monthly	Top3 purchased recently
	Golden Member	Female	36-45 years old	RMB: 7,872.98	seafood/shoes/milk
Q:	在吗？(Anyone here?)				
A:	在的！亲爱的客人，我有什么可以帮助您的吗？(Of course! What can I do for you? My dear.)				
Q:	快过年了，我需要一个新的冰箱，推荐一下。(Spring Festival is coming. I need a new refrigerator. Help me recommend one.)				
A:	我们有一款新出的冰箱，请您看一下(We have a new refrigerator, please take a look.)				
A:	https://item.jd.com/10023613092198.Html				
Q:	这款单开门的不错，但是这只适合小两口，我们过年一大家子人呢。(This single-door refrigerator is good, but this is only suitable for the couple, we have a big family.)				
A:	您的预算是多少呢？(What is your budget?)				
Q:	3000左右吧。(Around 3000.)				
Model	DeepFM+BF+CF	DeepFM+BF+IF	DeepFM+BF+CF+IF	Ground truth	
Recommend result					
	Price: 2,499 Capacity: 200-249L Color: Red Number of doors: 1	Price:2,099 Capacity: 100L Color: Green Number of doors: 1	Price:2,499 Capacity: 200-249L Color: White Number of doors: 1	Price:2,899 Capacity: 500-549L Color: Black Number of doors: 2	

(b)

Figure 5: Case study for the personalized recommendation. (Green box highlights the product mentioned in the conversation, and the orange one refers to ground truth.)

proper price but fails to capture the context features such as *dark color* in the conversation. In this case, user utilizes some words to explicitly express his/her needs (e.g. *take clear pictures*, *dark color*), which is challenging for system to understand. Whereas some other preference words (e.g. *memory should be more than 200G*) can be directly linked to the attributes in the product KB. Finally, benefiting from the combinations of basic feature, contextual feature, and interactive feature, the third model can make a correct recommendation. This illustrates that the bridging conversational corpus, user's portrait, and product KB is the key factor for successful recommendation.

Fig.5(b) presents a bad case where all the models fail to capture the user's intention for the recommendation. As illustrated from the example, though the user recognizes *single-door refrigerator* as a good candidate, it explicitly describes the demand for a larger size fridge with the phrase *we have a big family*. Thus all the models fail to understand the user's real intention. This case also indicates the complexity and variety for recommendation conversation presented in E-ConvRec.

5. Conclusions and Future Work

In this work, we contribute the Chinese conversational recommendation dataset which is large-scale, informative, and collected from the real scenario of E-commerce domain. To explore conversational recommendation in a real scene from multiple facets based on the dataset, we design three worth studying tasks which cover the critical problems of CRS. Extensive experiments are conducted and baselines are provided for these tasks. The experimental results indicate there is still a long way to go to solve the real scenario conversation recommendation problem. More in-depth researches on personalized preference recognition, multi-turn dialogue strategies, and response generation are needed in the future. Moreover, we will enrich the dataset annotations (e.g., emotions, richer intentions) in the future.

6. Acknowledgement

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600.

7. Reference

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Chen, Q., Lin, J., Zhang, Y., Ding, M., Cen, Y., Yang, H., and Tang, J. (2019a). Towards knowledge-based recommender dialog system. 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 1803–1813.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., and Wang, W. Y. (2019b). Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations (ICLR)*.

Christakopoulou, K., Radlinski, F., and Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD)*, pages 815–824.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Ding, R., Xie, P., Zhang, X., Lu, W., Li, L., and Si, L. (2019). A neural multi-digraph model for chinese ner with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1462–1467.

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., Szlam, A., and Weston, J. (2015). Evaluating prerequisite qualities for learning end-to-end dialog systems. In *4th International Conference on Learning Representations (ICLR)*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fu, Z., Xian, Y., Zhu, Y., Zhang, Y., and de Melo, G. (2020). Cookie: A dataset for conversational recommendation over knowledge graphs in e-commerce. *arXiv preprint arXiv:2008.09237*.

Gao, C., Lei, W., He, X., de Rijke, M., and Chua, T.-S. (2021). Advances and challenges in conversational recommender systems: A survey. *arXiv preprint arXiv:2101.09459*.

Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. (2017). Deepfm: a factorization-machine based neural network for ctr prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1725–1731.

Hayati, S. A., Kang, D., Zhu, Q., Shi, W., and Yu, Z. (2020). Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Jannach, D., Manzoor, A., Cai, W., and Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Kang, D., Balakrishnan, A., Shah, P., Crook, P., Boureau, Y.-L., and Weston, J. (2019). Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. 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 1951–1961.

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

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2267–2273.

Lei, W., He, X., Miao, Y., Wu, Q., Hong, R., Kan, M.-Y., and Chua, T.-S. (2020a). Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, pages 304–312.

Lei, W., Zhang, G., He, X., Miao, Y., Wang, X., Chen, L., and Chua, T.-S. (2020b). Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pages 2073–2083.

Li, R., Kahou, S., Schulz, H., Michalski, V., Charlin, L., and Pal, C. (2018). Towards deep conversational recommendation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9748–9758.

Li, X., Yan, H., Qiu, X., and Huang, X. (2020). Flat: Chinese ner using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6836–6842.

Liao, L., Long, L. H., Zhang, Z., Huang, M., and

Chua, T.-S. (2021). Mmconv: An environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 675–684.

Liu, P., Qiu, X., and Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2873–2879.

Liu, B., Tang, R., Chen, Y., Yu, J., Guo, H., and Zhang, Y. (2019). Feature generation by convolutional neural network for click-through rate prediction. In *The World Wide Web Conference (WWW)*, pages 1119–1129.

Liu, Z., Wang, H., Niu, Z.-Y., Wu, H., Che, W., and Liu, T. (2020). Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1036–1049.

Liu, W., Fu, X., Zhang, Y., and Xiao, W. (2021a). Lexicon enhanced chinese sequence labeling using bert adapter. 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)*, pages 5847–5858.

Liu, Z., Wang, H., Niu, Z.-Y., Wu, H., and Che, W. (2021b). Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4335–4347.

Ma, R., Peng, M., Zhang, Q., and Huang, X. (2019). Simplify the usage of lexicon in chinese ner. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5951–5960.

Massari, L. (2010). Analysis of myspace user profiles. *Information Systems Frontiers*, 12(4):361–367.

Moon, S., Shah, P., Kumar, A., and Subba, R. (2019). Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 845–854.

Ni, Y., Ou, D., Liu, S., Li, X., Ou, W., Zeng, A., and Si, L. (2018). Perceive your users in depth: Learning universal user representations from multiple e-commerce tasks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pages 596–605.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Radlinski, F., Balog, K., Byrne, B., and Krishnamoorthi, K. (2019). Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (SIGdial)*, pages 353–360.

Shah, P., Hakkani-Tur, D., Liu, B., and Tur, G. (2018). Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 41–51.

Sun, Y. and Zhang, Y. (2018). Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval (SIGIR)*, pages 235–244.

Xu, H., Moon, S., Liu, H., Liu, B., Shah, P., and Yu, P. S. (2020). User memory reasoning for conversational recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 5288–5308.

Zhang, Y. and Koren, J. (2007). Efficient bayesian hierarchical user modeling for recommendation system. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 47–54.

Zhang, Y. and Yang, J. (2018). Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1554–1564.

Zhou, K., Zhou, Y., Zhao, W. X., Wang, X., and Wen, J.-R. (2020). Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4128–4139.

Zhou, K., Wang, X., Zhou, Y., Shang, C., Cheng, Y., Zhao, W. X., Li, Y., and Wen, J.-R. (2021). Crslab: An open-source toolkit for building conversational recommender system. *arXiv preprint arXiv:2101.00939*.