C^2-CRS: Coarse-to-Fine Contrastive Learning for Conversational Recommender System

Yuanhang Zhou[1†], Kun Zhou[1†], Wayne Xin Zhao[2,3*], Cheng Wang[4], Peng Jiang[4], He Hu[1]

[1]School of Information, Renmin University of China, China

[2]Gaoling School of Artificial Intelligence, Renmin University of China, China

[3]Beijing Key Laboratory of Big Data Management and Analysis Methods, China

[4]Kuaishou Inc., China

sdzyh002@gmail.com, francis_kun_zhou@163.com, batmanfly@gmail.com,

wangcheng03@kuaishou.com, jp2006@139.com, hehu@ruc.edu.cn

ABSTRACT

Conversational recommender systems (CRS) aim to recommend suitable items to users through natural language conversations. For developing effective CRSs, a major technical issue is how to accurately infer user preference from very limited conversation context. To address issue, a promising solution is to incorporate external data for enriching the context information. However, prior studies mainly focus on designing fusion models tailored for some specific type of external data, which is not general to model and utilize multi-type external data.

To effectively leverage multi-type external data, we propose a novel coarse-to-fine contrastive learning framework to improve data semantic fusion for CRS. In our approach, we first extract and represent multi-grained semantic units from different data signals, and then align the associated multi-type semantic units in a coarse-to-fine way. To implement this framework, we design both coarse-grained and fine-grained procedures for modeling user preference, where the former focuses on more general, coarse-grained semantic fusion and the latter focuses on more specific, fine-grained semantic fusion. Such an approach can be extended to incorporate more kinds of external data. Extensive experiments on two public CRS datasets have demonstrated the effectiveness of our approach in both recommendation and conversation tasks.

CCS CONCEPTS

• **Information systems** → **Users and interactive retrieval**; **Recommender systems**.

KEYWORDS

Conversational Recommender System; Contrastive Learning

†Equal contribution.

*Corresponding author.

1 INTRODUCTION

With the rapid development of intelligent agents in e-commerce platforms, conversational recommender system (CRS) [4, 13, 24] has become an emerging research topic, which aims to provide effective recommendations to users through natural language conversations. In general, a CRS consists of a conversation module to converse with users and a recommender module to make recommendations.

To improve user satisfaction, a major objective of the CRS is to accomplish the recommendation task within as few conversation turns as possible. Therefore, it is important for a CRS to accurately capture user preferences based on several initial turns of conversational utterances, which contain very limited context information to understand user needs. Considering this issue, existing works [2, 16, 32] have introduced external data sources to enrich the contextual information. One line of research [2, 36] utilizes structured external data (*i.e.,* knowledge graphs) to enhance the representations of entities and words occurring in the conversational context. Another line of research [18] introduces unstructured external data (*e.g.,* item reviews) to improve the item representations for recommendation and help generate informative responses.

Due to the data heterogeneity, it is difficult to directly external data for improving CRS, because conversation data and external data usually correspond to very different information forms (*e.g.,* conversation utterances *v.s.* knowledge graphs) or semantic content (*e.g.,* conversation utterance *v.s.* online reviews). There exists a natural semantic gap between conversation data and external data. The case becomes even more difficult when multi-type external data (also in different data forms) is available for utilization. Several efforts have been made to leverage external data for improving CRS [18, 36]. However, they mainly focus on designing specific semantic fusion model tailored to some type of external data, which cannot apply to multi-type external data. Therefore, it is essential to develop a general approach to bridging the semantic gap between different data signals for CRS.

To fuse multi-type context data, a major challenge is that they usually correspond to very different semantic spaces. And, it may
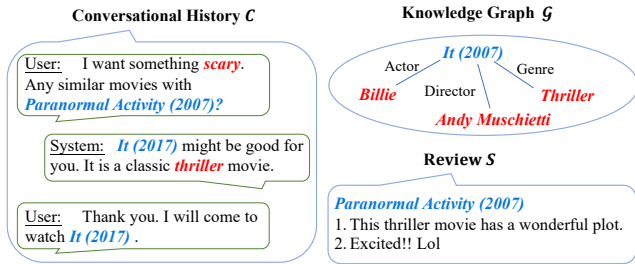
Figure 1: An illustrative example of a conversation on movie recommendation between a user and the system. The associated external structured data (knowledge graph) and unstructured data (reviews) of conversational context have been also presented. Items (movies) are in blue and entities (*e.g.,* actors) are in red.

hurt the original representation performance if we directly align their semantic space as previous studies [18, 36]. To address this issue, we are inspired by an important observation that context data itself (no matter external data or conversation data) is in a multi-grained form, *e.g.*, *entity* and *entity subgraph* in knowledge graphs, and *word* and *sentence* in utterances. Actually, user preference is also reflected in a multi-grained way [34]: a fine-grained semantic unit reflects some specific tastes (*e.g.,* an actor entity), while a coarse-grained semantic unit reflects some general tastes (*e.g.,* a set of comments for fiction movies). Given a conversation scenario, the semantic units from different data signals can be associated according to the reflected user preference. For example, in Figure 1, the word "*It*" in the fourth utterance corresponds to the entity "*It (2017)*" in Freebase, and the fourth utterance essentially corresponds to a subgraph spanned based on the entity "*It (2017)*". Such an example indicates that it needs a multi-grained semantic alignment in order to fuse different semantic spaces and better characterize user preference.

To this end, in this paper, we propose a novel **C**oarse-to-fine **C**ontrastive learning approach for **C**onversational **R**ecommender **S**ystem, called C^2-**CRS**. The core idea is to first extract and represent associated multi-grained semantic units from different data signals, and then align the corresponding semantic units from different data signals in a coarse-to-fine way. To implement coarse-to-fine semantic fusion, we involve both *coarse-grained* and *fine-grained* pre-training procedures, where the former focuses on more general, coarse-grained user preference while the latter focuses on more specific, fine-grained user preference. Such a way allows the model to gradually fuse different semantic spaces in a multi-grained manner, which is likely to yield more coherent fusion representations. The optimization objectives for coarse- and fine-grained fusion are developed in a unified form of contrastive learning, where we pull semantically associated semantic units from different data signals together and push apart irrelevant ones in their representation spaces. Based on the pre-trained representations, we develop a recommender module and a conversation module to accomplish the corresponding tasks. Our approach is general to leverage various types of external data for CRS.

To our knowledge, it is the first time that multi-type external data has been leveraged for CRS. We address this task with a novel coarse-to-fine contrastive learning approach, which can better fuse semantic spaces from different data signals. Such an approach can be extended to incorporate more kinds of external data. Extensive experiments on two public CRS datasets have demonstrated the effectiveness of our approach in both recommendation and conversation tasks.

2 RELATED WORK

In this section, we review the related work from the following two perspectives, namely conversational recommender system and contrastive learning.

2.1 Conversational Recommender System

With the rapid development of dialog system [21, 28, 30, 39] in recent years, interactive conversation with users becomes an appealing approach to obtaining user dynamic intent and preference. Based on it, conversational recommender system [12, 13, 24] (CRS) has become an emerging research topic, which aims to provide high-quality recommendations to users through natural language conversations.

One category of CRSs utilized predefined actions to interact with users, such as item attributes [4, 37] and intent slots [24]. Most of these methods mainly focus on accomplishing the recommendation task within fewer turns [12, 24]. A surge of works adopt multi-armed bandit [4], reinforcement learning [12] and Thompson sampling [14] for interaction with users. However, this category of works do not pay much attention on generating high-quality natural language responses but usually rely on predefined dialog templates and rules to compose responses [12, 24].

Another category of CRSs develop natural language based approaches, focusing on both making accurate recommendation and generating human-like responses [15, 16, 26, 36], these methods incorporate a generation-based dialog component to converse with users. However, since a conversation usually contains a few sentences, lack of sufficient contextual information for accurately capturing user preference. Existing works leverage entity-oriented knowledge graph [2], word-oriented knowledge graph [36] and review information [18] to alleviate this issue.

This paper extends the second category of research by leveraging rich external data for improving CRSs. Our key novelty lies in the coarse-to-fine contrastive learning approach, which can effectively fuse multi-type heterogeneous data in a principled way.

2.2 Contrastive Learning

Contrastive learning has been long considered as effective in constructing meaningful representations [3, 6, 17], which learns the representations by making a comparison between different samples. Usually, it assumes a set of paired examples, where both examples are semantically related neighbors. Then the training objective is defined to pull the representations of neighbors together and push apart non-neighbors.

Recently, as a pre-training technique, contrastive learning has achieved remarkable success in computer vision [3, 8], natural language processing [5, 27] and information retrieval [1]. These

works utilize data augmentation strategies to construct semantic related example pairs based on original data (*e.g.,* random cropping and rotation of images [3]). Besides, contrastive learning has also been adopted to fuse multi-view information, such as image and text [19, 29, 33], text and graph [9]. It drives the representations of related information in different types to be similar, such that each kind of representations can be enhanced by each other.

In our method, we first extract and represent associated multi-grained semantic units from different data signals, and then align the corresponding semantic units from different data signals via a coarse-to-fine contrastive learning. Such a contrastive learning approach can effectively fuse the representations of these semantic units from heterogeneous signals.

3 PRELIMINARIES

The goal of CRS is to recommend appropriate items to a user via a multi-turn natural language conversation. Typically, a CRS consists of two core components, namely the recommender component and the conversation component, and the two components should be integrated seamlessly to fulfill the recommendation goal.

Notations for CRS. Formally, let u denote a user from user set $\mathcal{U}$, i denote an item from item set $\mathcal{I}$, and w denote a word from vocabulary $\mathcal{V}$. A conversation (or a conversation history) C consists of a list of utterances, denoted by $C = \{s_t\}_{t=1}^n$, in which each utterance s_t is a conversation sentence at the t-th turn and each utterance $s_t = \{w_j\}_{j=1}^m$ is composed by a sequence of words. As the conversation goes on, the conversation utterances are aggregated (called *conversation history* or *conversation context*), and the CRS estimates the user preference based on the conversation history. At the t-th turn, the recommender component selects a set of candidate items $\mathcal{I}_{t+1}$ from the entire item set $\mathcal{I}$ according to the estimated user preference, while the dialog component needs to produce the next utterance s_{t+1} to reply to previous utterances.

External Data. Besides the conversation history C, in the online platform, some external data is usually available to enhance the performance of CRS. In the literature, two kinds of external data are widely used for CRS, namely structured external data *knowledge graph* [2, 36] and unstructured external data *item reviews* [18]. We also consider leveraging the two kinds of external data in this work. For knowledge graph $\mathcal{G}$, it is composed by an entity set $\mathcal{N}$ and a relation set $\mathcal{R}$. The entity set contains all the items (*i.e.,* $\mathcal{I}$ is a subset of the entity set) and other item-related entities. Furthermore, for each user u, her/his interacted entities set (*i.e.,* entities occurring in the conversation history) is denoted by $\mathcal{N}_u$, which composes a subgraph $\mathcal{G}_u$ that reflect the user preference. For an online review, it can be considered as a document consisting of sentences $\mathcal{S} = \{\tilde{s}_j\}_{j=1}^n$. The two kinds of external data covers both structured and unstructured data, and reflect a multi-grained semantic form: entity *v.s.* subgraph for knowledge graph, and sentence *v.s.* document for online reviews. As will be shown later, our approach is general to incorporate multiple types of external data in CRS.

Task Definition. Based on these notations, the task of this paper is defined as: given the multi-type context data (*i.e.,* conversation

history C, knowledge graph $\mathcal{G}$ and reviews $\mathcal{S}$), we aim to (1) accurately recommend items $\mathcal{I}_{t+1}$ and (2) generate proper response s_{t+1} to the user u at the $(t + 1)$-th turn of a conversation.

4 APPROACH

In this section, we present the proposed **C**oarse-to-fine **C**ontrastive learning for **C**onversational recommender system, called $\mathbf{C}^2$-**CRS**. Since we consider utilizing multi-type context data for CRS, we first study how to separately encode these context data. Then, we introduce the coarse-to-fine contrastive learning approach to gradually fusing context data for pre-training effective data representations. Based on the learned representations, we finally describe our solutions for both recommendation and conversation tasks. The overview illustration of the proposed model is presented in Figure 2.

4.1 Encoding Multi-type Context Data

Given the multi-type context data, we adopt corresponding encoding modules to generate the data representations in separate semantic spaces. Next, we present the encoding modules for conversation history, knowledge graph and reviews, respectively.

4.1.1 Encoding Conversation History. Conversation history C consists of utterances $\{s_t\}_{t=1}^n$ generated in a session between the CRS and the user, which depicts the user-system interaction about the information needs. Since the utterances, which are usually short sentences, are closely relevant, following previous works [2, 36], we concatenate these utterances $\{s_t\}_{t=1}^n$ in a chronological order to form a long sentence $s_{1:n}$. To obtain the representations of the conversation history, we adopt a standard Transformer [25] to encode $s_{1:n}$ as following:

$$F = \text{Transformer}(s_{1:n}), \tag{1}$$

where contextual word representations F are obtained on the top layer of Transformer.

Then we utilize a self-attentive layer to produce the representation of the conversation history $\mathbf{e}_C$:

$$\mathbf{e}_C = F \cdot \text{softmax}(\boldsymbol{b}^\top \tanh(\mathbf{W}_{sa}F)), \tag{2}$$

where $\mathbf{W}_{sa}$ and $\boldsymbol{b}$ are learnable parameter matrix and vector.

4.1.2 Encoding Knowledge Graph. The knowledge graph $\mathcal{G}$ consists of an entity set $\mathcal{N}$ and a relation set $\mathcal{R}$. It stores semantic facts in the form of a triple $\langle n_1, r, n_2 \rangle$.

Consider that the edges connected the nodes may own useful relation information, we utilize R-GCN [20] to encode $\mathcal{G}$. Formally, the representation of node e at $(l + 1)$-th layer is calculated as:

$$\boldsymbol{n}_e^{(l+1)} = \sigma\Big(\sum_{r \in \mathcal{R}} \sum_{e' \in \mathcal{E}_e^r} \frac{1}{Z_{e,r}} \mathbf{W}_r^{(l)} \boldsymbol{n}_{e'}^{(l)} + \mathbf{W}^{(l)} \boldsymbol{n}_e^{(l)} \Big) \tag{3}$$

where $\boldsymbol{n}_e^{(l)} \in \mathbb{R}^{d_E}$ is the node representations of e at the l-th layer, $\mathcal{E}_e^r$ denotes the set of neighboring nodes for e under the relation r, $\mathbf{W}_r^{(l)}$ and $\mathbf{W}^{(l)}$ are learnable matrices and $Z_{e,r}$ is a normalization factor. After aggregating the graph information, we obtain the node representations matrix $\boldsymbol{N}$ on the top R-GCN layer. As shown in previous studies [2, 36], it is particularly important to consider the historically interacted entities $\mathcal{N}_u$ for modeling the user preference.
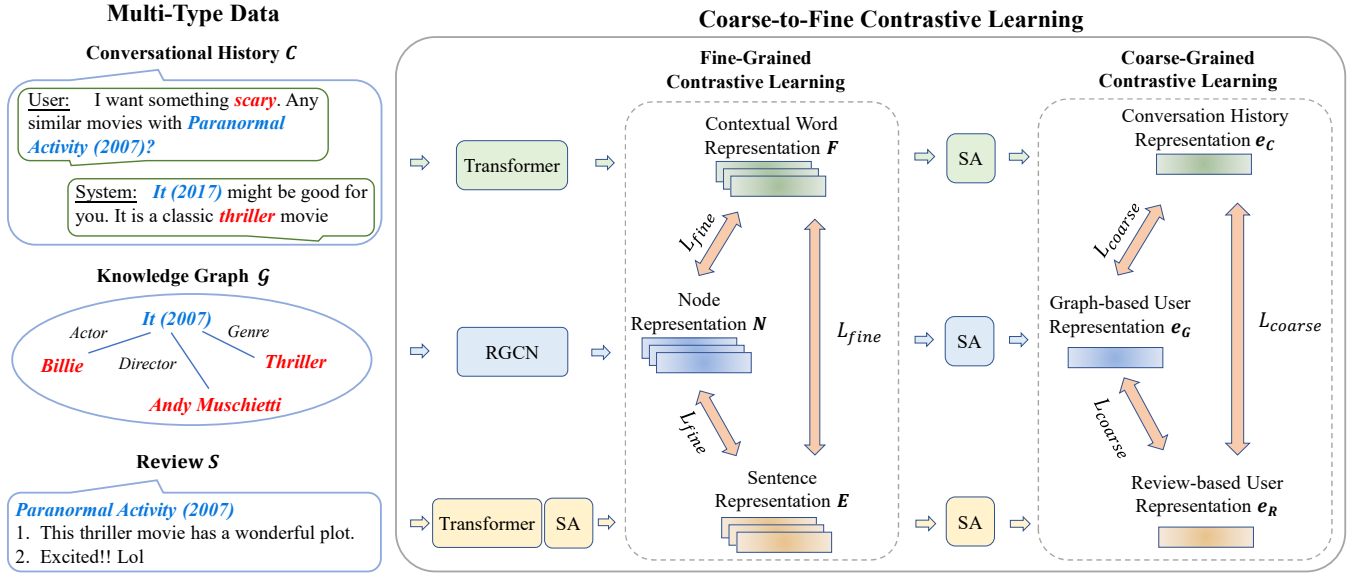
Figure 2: The overview of our model in a movie recommendation scenario, where "SA" denotes the self-attention layer. In the coarse-to-fine contrastive learning, we first conduct coarse-grained contrastive learning using coarse-grained features, and then conduct fine-grained contrastive learning using fine-grained features.

Therefore, we collect the node representations of user interacted entities as N_u, and utilize self-attentive mechanism as Eq. 2 to generate the graph-based user representation e_G. The basic idea is to automatically learn the importance of each interacted entity for user u, so that we can derive the user preference considering the levels of entity importance.

4.1.3 Encoding Reviews. The review text S_i of an item is a set of sentences $S_i = \{\tilde{s}_j\}_{j=1}^n$ written by online users about the item i. Similar as the conversation history, we utilize the standard Transformer model to encode each sentence and leverage a self-attention layer to obtain the sentence representation matrix E. Note that the parameters of this Transformer and self-attention layer are different from those in Section 4.1.1. Besides, consider that the reviews are usually noisy and may be irrelevant to user preference [18], we further utilize a sentence-level self-attention layer to select the useful information within the sentence representations. Finally, we produce the review-based user representation e_R.

After the above encoding, we can obtain the corresponding representations for conversation history, knowledge graph and review text. The three kinds of context data are represented in different semantic spaces. Next, we study how to fuse them in order to derive shared data semantics.

4.2 Coarse-to-Fine Contrastive Learning

In order to utilize context data, we need to effectively fuse their representation spaces, so that their information can be shared and leveraged across different data signals. In this section, based on the multi-grained correlations between multi-type context data, we propose a coarse-to-fine contrastive learning method to fuse the multi-type information to enhance the data representations.

The main idea is to represent each type of data from both coarse-grained and fine-grained views and associate the corresponding representations at different granularities in a coarse-to-fine manner. We will introduce the coarse-grained and fine-grained pre-training approaches in the following.

4.2.1 Coarse-Grained Contrastive Learning. Multi-type context data in a coarse-grained form mainly reflects the overall user preference. As in Section 4.1, we obtain the coarse-grained representations of the user from the views of the conversation data e_C (*i.e.,* conversation history), knowledge graph e_G (*i.e.,* subgraph) and review text e_R (*i.e.,* review document). There exists a natural semantic gap between these data representations. To bridge this gap, we propose a contrastive learning based pre-training approach to aligning and fusing the three types of context data.

Contrastive learning is a widely adopted pre-training technique which learns the representations by pulling semantically close representations together and pushing apart non-related ones. In our setting, we have three representations e_C, e_G and e_R depicting the same user preference from different views, which are considered as semantically close ones. Therefore, based on the three representations, we take the pairs (e_C, e_G), (e_C, e_R), and (e_G, e_R) as positive examples, while the representations from different users in the same batch are considered as negative examples. Thus, for a mini-batch with b pairs, the coarse-grained pre-training objective is the sum of the triple contrast learning loss, which can be formulated as:

$$L_{coarse} = L_{CL}(e_C, e_G) + L_{CL}(e_C, e_R) + L_{CL}(e_G, e_R), \quad (4)$$

$$L_{CL}(h, h^+) = \log \frac{e^{\text{sim}(h,h^+)/\tau}}{\sum_{h_i^- \in \{h^-\}} e^{\text{sim}(h,h_i^-)/\tau}}, \quad (5)$$

where h and h^+ are two types of coarse-grained representations of a user, $\{h^-\}$ is the negative example set for positive examples (h, h^+),

τ is a temperature hyperparameter and $\text{sim}(h, h^+)$ is the cosine similarity $\frac{h^\top h^+}{||h|| \cdot ||h^+||}$. Since the representations of the same users under different views are pulled together, this objective aligns the semantic space of the three types of context data. The three types of representations capture user preference from different views, which are complementary to each other. By semantic alignment and fusion using the contrastive learning objective, these representations can be also mutually improved by each other.

4.2.2 Fine-Grained Contrastive Learning. The above coarse-grained contrastive learning fuses the semantic space at the overall level. However, the corresponding semantic associations between fine-grained semantic units (*e.g.,* words and entities) are neglected. Fine-grained context data captures specific user preference about fine-grained characteristics. Therefore, we further propose to conduct fine-grained contrastive learning for better fusing the representation spaces. For the three types of context data, the fine-grained preferences are encoded in the contextual word representation (F_j) of a word from conversation history, the node representation (N_e) of an entity from the knowledge graph, and the representation (E_k) of a sentence from the review document, respectively.

Similar to the coarse-grained pre-training, we also adopt the contrastive learning method for fine-grained pre-training. To construct the semantic-closed example pairs, we consider capturing the correlations among these fine-grained semantic units, *i.e.,* words from conversation history, entities from knowledge graph and sentences from review text. We also consider inter-type data associations. For the word-entity correlation, we utilize the entity linking method [2] to match the entity from the knowledge graph with its corresponding word in the conversation history. For the entity-sentence correlation, we match the entity with its most related sentence, which may be the most highly-rate review or descriptive text of the entity. In this way, we can construct the semantically-consistent *word-entity-sentence* representation triples (F_j, N_e, E_k) to depict the same fine-grained user preference. Then, we split the triples into pairs (F_j, N_e), (F_j, E_k) and (N_e, E_k) to construct paired positive examples for contrastive learning. Similarly, the representations from other users in the same batch compose a negative example set.

Therefore, our fine-grained pre-training objective can be formulated as:

$$L_{fine} = L_{CL}(F_j, N_e) + L_{CL}(F_j, E_k) + L_{CL}(N_e, E_k), \quad (6)$$

where $L_{CL}(\cdot)$ is defined according to Eq. 5. This optimization objective enforces the alignment among semantic spaces of words, nodes and sentences.

To further preserve the enhancement effect of coarse-grained contrastive learning, we integrate its optimization objective with a weight λ. Then, the final fine-grained pre-training objective L is given as follows:

$$L = L_{fine} + \lambda \cdot L_{coarse}. \quad (7)$$

Instead of directly fusing different semantic spaces, we design a coarse-to-fine fusion way, so the semantic space will be gradually pulled close and finally fused.

4.3 Fine-tuning the Conversation Recommender System

Based on the pre-trained representations, we introduce our approach to fine-tuning the data representations and network architectures to accomplish the conversational recommendation task. In the following, we will describe the architectures of our recommendation module and response generation module, and detail how to fine-tune them.

4.3.1 Fine-tuning Recommendation Module. Given the pre-trained item representations, we study how to fine-tune them for the recommendation task. We first generate the user representation, and then fine-tune it on item recommendation task.

Note that after semantic fusion, the learned entity representations have learned useful information from other types of context data. Therefore, we only adopt the entities $\mathcal{N}_u$ occurring in the conversation history for learning the user representation. Specifically, we stack the entities' representations of $\mathcal{N}_u$ into a matrix N_u, and then utilize self-attentive mechanism as in Eq. 2 to produce the user representation e_u. Finally, we compute the probability that recommends an item i from the item set to a user u:

$$P_{rec}(i) = \text{softmax}(e_u^\top n_i), \quad (8)$$

where n_i is the learned item embedding for item i. We can utilize Eq. 8 to rank all the items and generate a recommendation set to a user. To fine-tune the user representation e_u and item embedding n_i, we apply a cross-entropy loss as the optimization objective:

$$L_{rec} = -\sum_{j=1}^{N} \sum_{i=1}^{M} y_{ij} \cdot \log\left(P_{rec}^{(j)}(i)\right), \quad (9)$$

where N is the number of conversations, j is the index of a conversation, M is the number of items, and i is the index of an item. Here, we consider the case with a single ground-truth recommendation. It will be easy to extend the above loss to the case with multiple ground-truth items.

4.3.2 Fine-tuning Response Generation Module. Here, we study how to fine-tune the pre-trained representations for the conversation task. Following KGSF [36], we incorporate multiple cross-attention layers in a standard Transformer [25] decoder to fuse the pre-trained representations as following:

$$\mathbf{R}^l = \text{Decoder}(\mathbf{R}^{l-1}, N, E, F), \quad (10)$$

where $\mathbf{R}^l$ is the output representation matrix from the decoder at l-th layer. The above equation follows the similar transformation chain as KGSF [36]: *generated words $\rightarrow$ conversation history $\rightarrow$ knowledge graph $\rightarrow$ reviews*. We omit the equation details. Based on it, following existing works [36], we also design the copy network to enhance the generation of informative response.

To fine-tune this module for generating more informative responses, we devise an instance weighting enhanced cross-entropy loss as:

$$L_{gen} = -\frac{1}{m} \sum_{j=1}^{m} \log\left(\alpha_{w_j} P(w_j | w_1, \cdots, w_{j-1}))\right), \quad (11)$$

$$\alpha_{w_j} = \begin{cases} \max(\gamma, \frac{\beta}{f_{w_j}}) & \text{if } f_{w_j} \geq \beta \\ 1 & \text{otherwise} \end{cases}, \quad (12)$$

where m is the number of words in generated response, α_{w_j} is the weight considering the frequency of this token, β is a preset threshold, γ is to avoid punishing the high-frequency words too much, and f_{w_j} is the word frequency of w_j in corpus. With the above cross-entropy loss, we can punish the high-frequency tokens, and help generate more informative responses.

4.4 Discussion

To summarize, our approach provides a novel coarse-to-fine contrastive learning framework for leveraging multi-type context data to improve CRS. Next, we compare it with existing studies.

Conversation-centered approaches such as ReDial [13] and CRM [24] mainly focus on using the conversation history to accomplish the conversational recommendation task, with or without the external auxiliary data. Our work falls into this category and extends it by leveraging multi-type context data with a novel coarse-to-fine contrastive learning framework. As a comparison, most of existing studies focus on some specific type of external data, which is not general to leverage various heterogeneous external data.

Knowledge graph based approaches such as DeepCR [16], KBRD [2] and KGSF [36] fuse external KG and conversational context to help model user representations and generate more informative responses. This category of studies only considers structured external data, which neglects rich unstructured data as below.

Review-enhanced approaches such as RevCore [18] introduce external reviews to CRS as the supplement of conversational context. These methods extract entities from reviews for recommender module and utilize copy mechanism for conversation module. However, reviews are inevitable to contain noise. In our approach, we leverage the coarse-to-fine pre-training approach to fuse the multi-type data, which can adaptively fuse useful information from data and prevent the noisy reviews to directly affect the modeling of user preference.

5 EXPERIMENT

In this section, we first set up the experiments, and then report the results and give a detailed analysis.

5.1 Experiment Setup

In this subsection, we provide an introduction to the details of our experiments, including dataset, baselines, evaluation metrics and implementation details.

5.1.1 Dataset. We evaluate our model on ReDial [13] and TG-ReDial [38] datasets. The ReDial dataset is an English conversational recommendation dataset constructed with Amazon Mechanical Turk (AMT). Following a set of comprehensive instructions, the AMT workers played the roles of seekers and recommenders to generate dialogue for recommendation on movies. It contains 10,006 conversations consisting of 182,150 utterances related to 51,699 movies. The TG-ReDial dataset is a Chinese conversational recommendation dataset, which emphasizes natural topic transitions from non-recommendation scenarios to the desired recommendation scenario. It is created in a semi-automatic way, hence human annotation is more reasonable and controllable. It contains 10,000 two-party dialogues consisting of 129,392 utterances related to 33,834 movies. For each conversation, we start from the first

Table 1: Results on the recommendation task. Numbers marked with * indicate that the improvement is statistically significant compared with the best baseline (t-test with p-value < 0.05).

Dataset	ReDial			TG-ReDial		
Models	R@1	R@10	R@50	R@1	R@10	R@50
Popularity	0.011	0.054	0.183	0.0004	0.003	0.014
TextCNN	0.013	0.068	0.191	0.003	0.010	0.024
ReDial	0.024	0.140	0.320	0.000	0.002	0.013
KBRD	0.031	0.150	0.336	0.005	0.032	0.077
KGSF	0.039	0.183	0.378	0.005	0.030	0.074
KECRS	0.021	0.143	0.340	0.002	0.026	0.069
RevCore	0.046	0.220	0.396	0.004	0.029	0.075
C^2-CRS	**0.053***	**0.233***	**0.407***	**0.007***	**0.032***	**0.078***

sentence one by one to generate reply to utterances or give recommendations. Since the above datasets do not contain the review data, we retrieve reviews for movies in ReDial and TG-ReDial from IMDB [1] and douban [2] respectively.

5.1.2 Baselines. In CRS, we consider two major tasks to evaluate our method, namely recommendation and conversation. Therefore, we not only compare our approach with existing CRS methods, but also select representative recommendation or conversation models as baselines.

• *Popularity*: It ranks the items according to recommendation frequencies in the training set of the corpus.

• *TextCNN* [10]: It applies a CNN-based model to extract user features from conversational context for ranking items.

• *Transformer* [25]: It utilizes a Transformer-based encoder-decoder method to generate conversational responses.

• *ReDial* [13]: This model is proposed in the same paper with the ReDial dataset. It consists of a dialog generation module based on HRED [22] and a recommender module based on auto-encoder [7].

• *KBRD* [2]: This model utilizes DBpedia to enhance the semantics of contextual items. The Transformer architecture is applied in the dialog generation module, in which KG information is used as word bias in generation.

• *KGSF* [36]: This model incorporates DBpedia and ConceptNet to enhance the semantic representations of items and words, and uses Mutual Information Maximization to align the semantic spaces of different components.

• *KECRS* [31]: It constructs a high-quality KG and develops the Bag-of-Entity loss and the infusion loss to better integrate KG with CRS for generation.

• *RevCore* [18]: It proposes a review-enhanced framework, in which reviews are selected by a sentiment-aware retrieval module and are utilized to enhance recommender module and dialogue generation module. For a fair comparison, we use the same review as our approach.

Among these baselines, *Popularity* and *TextCNN* [10] are recommendation methods, while *Transformer* [25] is the state-of-the-art text generation method. We do not include other recommendation models since there are no historical user-item interaction

---
[1]https://www.dbpedia.org/
[2]https://movie.douban.com/

records except dialogue utterances. Besides, *REDIAL* [13], *KBRD* [2], *KGSF* [36], *KECRS* [31] and *RevCore* [18] are conversational recommendation methods. We name our proposed model as C^2-*CRS*.

5.1.3 Evaluation Metrics. In our experiments, we adopt different metrics to evaluate on the two tasks. For the recommendation task, following [2], we adopt Recall@k (k = 1, 10, 50) for evaluation. For the conversation task, the evaluation consists of automatic evaluation and human evaluation. Following [2], we use Distinct *n*-gram (n = 2, 3, 4) to measure the diversity at sentence level. Besides, we invite three annotators to score the generated candidates in two aspects, namely *Fluency* and *Informativeness*. The range of score is 0 to 2. The final performance is calculated using the average score of three annotators.

5.1.4 Implementation Details. We implement our approach with Pytorch [3] and CRSLab [35] [4]. The dimensionality of embeddings (including hidden vectors) is set to 300 and 128, respectively, for conversation and recommender modules. In the structured encoder module, we set the layer number to 1 for GNN networks. We use Adam optimizer [11] with the default parameter setting. In experiments, the batch size is set to 256, the learning rate is 0.001, gradient clipping restricts the gradients within [0, 0.1], the temperature of contrastive learning is set to 0.07, and the normalization constant $Z_{v,r}$ of R-GCN in Eq. 3 is 1. During the coarse-grained pre-training stage, we directly optimize the loss as Section 4.2.1. While in the fine-grained pre-training stage, the weight λ of the coarse-grained contrastive learning loss in Eq. 7 is set to 0.2. The code and data are available at: https://github.com/RUCAIBox/WSDM2022-C2CRS.

5.2 Evaluation on Recommendation Task

To verify the effectiveness of our proposed method on the recommendation task, we conduct a series of experiments and present the results in Table 1.

In general, conversational recommendation methods perform better than recommendation methods (*e.g.*, TextCNN and Popularity). Since these methods mainly focus on integrating the conversational module and the recommendation module, which are mutually beneficial for each other. For recommendation methods, we can see that TextCNN achieves better performance than Popularity. One reason is that Popularity only recommends the most popular items without considering the contextual information. In contrast, TextCNN is able to model personalized preference from contextual text for better recommendation.

For conversational recommendation methods, first, compared with the TG-REDIAL dataset, ReDial model performs better on the REDIAL dataset. One possible reason is that the contextual items in conversation from the TG-REDIAL dataset are much sparser than those in REDIAL dataset, but ReDial model relies heavily on the contextual items to generate the recommendation. Second, KBRD achieves better performance than ReDial. Since KBRD utilizes knowledge graph as external contextual information to improve user preference modeling, and then utilizes R-GCN and Transformer to accomplish the recommendation and conversation tasks, respectively. KGSF achieves better performance than KBRD and KECRS.

---

[3]https://pytorch.org/
[4]https://github.com/RUCAIBox/CRSLab

Table 2: Automatic evaluation results on the conversation task, where "Transf" denotes Transformer model. We abbreviate Distinct-2,3,4 as Dist-2,3,4. Numbers marked with * indicate that the improvement is statistically significant compared with the best baseline (t-test with p-value < 0.05).

Dataset	ReDial			TG-ReDial		
Models	Dist-2	Dist-3	Dist-4	Dist-2	Dist-3	Dist-4
Transf	0.067	0.139	0.227	0.053	0.121	0.204
ReDial	0.082	0.143	0.245	0.055	0.123	0.215
KBRD	0.086	0.153	0.265	0.045	0.096	0.233
KGSF	0.114	0.204	0.282	0.086	0.186	0.297
KECRS	0.040	0.090	0.149	0.047	0.114	0.193
RevCore	0.092	0.163	0.221	0.043	0.105	0.175
C^2-CRS	0.163*	0.291*	0.417*	0.189*	0.334*	0.424*

It improves the data representations by aligning the two semantic spaces between conversations and items via mutual information maximization. Finally, RevCore achieves better performance than other baselines. It incorporates external reviews to enhance the description of items, which help better capture user preference.

Our model C^2-CRS outperforms all the baselines, since C^2-CRS utilizes multi-type external information to help understand the conversational history, including conversational text, knowledge graph and reviews. To achieve it, our approach applies the coarse-to-fine contrastive learning to gradually fuse different types of information. Such a way can be beneficial to the recommender module by enhancing the data representations.

5.3 Evaluation on Conversation Task

In this subsection, we verify the effectiveness of the proposed model for the conversation task and report the results on automatic and human evaluation metrics.

5.3.1 Automatic Evaluation. We present the results of the automatic evaluation metrics for different methods in Table 2.

First, we can see that ReDial performs better than Transformer, since ReDial applies a pre-trained RNN model to produce better representations of historical conversation. Second, KBRD achieves better performance than ReDial in most settings. Since it enhances contextual entities and items by external KG and these entities are utilized to produce word probability bias for conversational module. Third, KGSF generates the most diverse response among these baselines. Since it not only aligns the 1 of conversational text and items, but also enhances their representations. Besides, multiple cross-attention layers are performed in the Transformer decoder to further interact the contextual information with the generated response. Finally, RevCore performs not well. One possible reason is that its involved reviews may contain noise.

Compared with these baselines, our model C^2-CRS performs consistently better in all the evaluation metrics, and it improves the Transformer decoder with enhanced multi-type data representations by coarse-to-fine contrastive learning. Such a gradual fusion approach is robust to the noise in the contextual information and can better capture useful semantics from conversational text, knowledge graphs, and reviews. Besides, we further design an instance

Table 3: Human evaluation results on the conversation task, where Transf refers to the Transformer model. Numbers marked with * indicate that the improvement is statistically significant compared with the best baseline (t-test with p-value < 0.05).

Models	Fluency	Informativeness
Transf	0.97	0.92
ReDial	1.35	1.04
KBRD	1.23	1.15
KGSF	1.48	1.37
KECRS	1.39	1.19
RevCore	1.52	1.34
C^2-CRS	**1.55***	**1.47***

Table 4: Results of ablation and variation study on the recommendation task. Coarse and Fine refer to coarse-grained and fine-grained contrastive learning, respectively. CH, SD and UD refer to conversational history, external structured data and external unstructured data, respectively.

Dataset	ReDial		
Models	R@1	R@10	R@50
C^2-CRS	**0.053***	**0.233***	**0.407***
C^2-CRS w/o Coarse-Fine	0.031	0.150	0.336
C^2-CRS w/o Coarse	0.049	0.220	0.394
C^2-CRS w/o Fine	0.048	0.221	0.400
C^2-CRS Multi-task	0.052	0.221	0.404
C^2-CRS w/o CH	0.048	0.223	0.403
C^2-CRS w/o SD	0.051	0.226	0.403
C^2-CRS w/o UD	0.051	0.222	0.395

weighting mechanism to help generate more informative responses. It shows that our approach can effectively improve the response generation task.

5.3.2 *Human Evaluation.* Table 3 presents the results of the human evaluation for the conversation task.

First, ReDial achieves better performance than Transformer, since it incorporates a pre-trained RNN encoder [23]. Second, KBRD achieves a comparable performance with ReDial, which indicates that the entity information from KG is beneficial to informativeness. Third, among these baselines, KGSF performs the best in term of *informativeness* since it aligns the conversational text and items via knowledge graph semantic fusion. Fourth, RevCore performs the best in term of *fluency*, since it utilizes external reviews to enhance the decoder for generating more fluent responses.

Finally, C^2-CRS performs the best in both metrics. By effectively leveraging and fusing multi-type data, our model is able to generate more informative words or entities, and meanwhile maintains the fluency of the generated text. Besides, the instance weighting mechanism in the fine-tuning stage also helps generate more fluent and informative tokens.

5.4 Ablation Study

We also conduct the ablation study based on different variants of our model, including: (1) C^2-*CRS w/o Coarse-Fine* removes the coarse-to-fine contrastive learning; (2) C^2-*CRS w/o Coarse* removes the coarse-grained contrastive learning; (3) C^2-*CRS w/o Fine* removes the fine-grained contrastive learning; (4) C^2-*CRS Multi-task* combines all the pre-training and fine-tuning tasks as a multi-task training; (5) C^2-*CRS w/o CH* removes the conversational history; (6) C^2-*CRS w/o SD* removes the structured data (*i.e.,* knowledge graph); and (7) C^2-*CRS w/o UD* removes the unstructured data (*i.e.,* reviews).

As shown in Table 4, firstly, we can see that removing coarse-to-fine contrastive learning leads to the largest performance decrease. Since there is a natural semantic gap in the multi-type data, it is key to fuse the underlying semantics for effective information utilization. Secondly, the direct multi-task setting (combining all the pre-training and fine-tuning tasks) yields a worse performance than our approach. The major reason is that it is difficult to fuse different types of data, which requires a more principled approach for semantic fusion. Finally, the variants that remove any kind of external data lead to a performance decrease. It shows that all kinds of external data are useful in our approach in enhancing the data representations.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel contrastive learning based coarse-to-fine pre-training approach for conversational recommender system. By utilizing the coarse-to-fine pre-training strategy, multi-type data representations can be effectively fused, such that the representations for limited conversation context are further enhanced, which finally improve the performance of CRS. By constructing extensive experiments, the effectiveness of our approach in both recommendation and conversation tasks has been demonstrated. It has shown that our approach is effective to bridge the semantic gap between different external data signals for CRS. Note that our approach is flexible to incorporate more kinds of external data, and is general to improve other tasks.

Currently, our focus is how to perform effective semantic fusion for incorporating external data for CRSs. As future work, we will consider designing a more general representation model that can be directly pretrained with various kinds of context data.

REFERENCES

[1] Shuqing Bian, Wayne Xin Zhao, Kun Zhou, Jing Cai, Yancheng He, Cunxiang Yin, and Ji-Rong Wen. 2021. Contrastive Curriculum Learning for Sequential User Behavior Modeling via Data Augmentation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3737–3746.

[2] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. (2019), 1803–1813.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[4] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *SIGKDD 2016*. 815–824.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[6] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[7] Junhua He, Hankz Hankui Zhuo, and Jarvan Law. 2017. Distributed-representation based hybrid recommender system with short item descriptions. *arXiv preprint arXiv:1703.04854* (2017).

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[9] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y Ng, and Ram Rajagopal. 2021. Learning Neighborhood Representation from Multi-Modal Multi-Graph: Image, Text, Mobility Graph and Beyond. *arXiv preprint arXiv:2105.02489* (2021).

[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP 2014*. 1746–1751.

[11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015*.

[12] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 304–312.

[13] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. (2018), 9748–9758.

[14] Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. 2020. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *arXiv preprint arXiv:2005.12979* (2020).

[15] Zujie Liang, Huang Hu, Can Xu, Jian Miao, Yingying He, Yining Chen, Xiubo Geng, Fan Liang, and Daxin Jiang. 2021. Learning Neural Templates for Recommender Dialogue System. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7821–7833.

[16] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep Conversational Recommender in Travel. *ArXiv* abs/1907.00710 (2019).

[17] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893* (2018).

[18] Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-augmented Conversational Recommendation. *arXiv preprint arXiv:2106.00957* (2021).

[19] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional Video Grounding with Dual Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2765–2775.

[20] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. *Lecture Notes in Computer Science* (2018), 593–607.

[21] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*.

[22] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues.. In *AAAI*.

[23] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *ICLR 2018*.

[24] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *SIGIR 2018*. 235–244.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. , 5998–6008 pages.

[26] Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. 2021. Finetuning Large-Scale Pre-trained Language Models for Conversational Recommendation with Knowledge Graph. *arXiv preprint arXiv:2110.07477* (2021).

[27] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466* (2020).

[28] Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural Response Generation with Meta-words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5416–5426.

[29] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 833–842.

[30] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243* (2018).

[31] Tong Zhang, Yong Liu, Peixiang Zhong, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. KECRS: Towards Knowledge-Enriched Conversational Recommendation System. *arXiv preprint arXiv:2105.08261* (2021).

[32] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *CIKM 2018*. 177–186.

[33] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747* (2020).

[34] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.

[35] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An Open-Source Toolkit for Building Conversational Recommender System. In *ACL 2021: System Demonstrations*. ACL, Online, 185–193. https://doi.org/10.18653/v1/2021.acl-demo.22

[36] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1006–1014.

[37] Kun Zhou, Wayne Xin Zhao, Hui Wang, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. Leveraging historical interaction data for improving conversational recommender system. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2349–2352.

[38] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4128–4139.

[39] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view Response Selection for Human-Computer Conversation. In *EMNLP*.