The Journal of Supercomputing (2025) 81:153
https://doi.org/10.1007/s11227-024-06666-w

MC-CRS: enhanced conversational recommender system based on multi-contrastive learning

Xiaohong Li[1] · Jin Yao[1] · Peng liu[1] · Yang Han[1]

Accepted: 31 October 2024

Abstract
Accurately learning dynamic user preferences from limited conversations and generating responses with interpretations is crucial for conversational recommender systems (CRS). Existing research has mainly focused on enhancing the understanding of dialogue context with the help of specific types of external knowledge bases (especially knowledge graphs). This process often neglects the learning and modeling of the original dialogue and lacks the utilization and integration of multi-type data. To this end, we propose a new multi-contrastive learning approach for conversational recommender systems, called (MC-CRS), which first obtains two representations of a contextual information through a text encoder and a 'perturb' encoder, and utilizes contrastive learning to mine the deep semantic information hidden in the contextual information. Second, we use structured knowledge graphs and personalized multi-reviews to pre-train the recommendation module, which uses contrastive learning to bridge the semantic gap between multi-types of data to achieve diverse recommendations. We conduct a large number of experiments on two public CRS datasets, and the final results demonstrate the effectiveness of our approach in recommendation and conversation generation tasks.

Keywords Conversational recommender system · Contrastive learning · Knowledge graph · Contextual information · Data augmentation

✉ Xiaohong Li
xiaohongli@nwnu.edu.cn

Jin Yao
2022222228@nwnu.edu.cn

Peng liu
2023222136@nwnu.edu.cn

Yang Han
2023222132@nwnu.edu.cn

[1] College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, People's Republic of China

Published online: 11 November 2024

1 Introduction

Conversational recommender systems (CRS) aim to capture users' dynamic preferences through dialogues in order to deliver high-quality recommendations [1]. Typically, CRS consists of a recommendation module that learns user preferences based on the content of dialogues and provides appropriate recommendations to the user, and a dialog module that combines the results of the recommendation module to generate a natural and fluent response [2]. Unlike traditional static recommendations that require users to search actively, CRS is widely used in various real-world scenarios such as e-commerce platform customer service and intelligent voice assistant Siri due to its attributes of 'dynamic interaction' and 'real-time feedback,' and its more natural and humanized dialogue improves user experience and user engagement [3].

As a research hotspot in the field of natural language processing, numerous researchers have made significant efforts to develop CRS. Some CRS models [4] are based on reinforcement learning to achieve a balance between exploration and exploitation, and some methods [5] employ more complex encoders for dialogue representation learning. Due to the limited number of CSR dialogue turns, inferring user preferences from contextual information alone is challenging. Consequently, several studies focus on modeling user preferences in CRS using external data. Zhou et al. [6] introduce structured knowledge graphs (KG) to enhance entity and word representation in dialogues, while Lu et al. [7] incorporate unstructured reviews to capture users' diverse impressions of items. However, most existing studies concentrate on utilizing and modeling external data, while ignoring the original data of CRS, namely the dialogue context information. Simultaneously, considering the heterogeneity among data, there exist substantial disparities in the manner of information representation between dialogue data and external data (such as KG and multi-reviews of items), which hinder the ability to augment CRS from external data effectively [8]. Essentially, there is a semantic gap between the multi-types of data mentioned above (conversations and external data sources), especially when faced with multi-types of external data in different formats, existing studies only consider learning user preferences from specific data types while neglecting the utilization and integration of multi-types of data [6, 7].

Inspired by the fact that contrastive learning [9, 10] can ignore noise interference in the representation learning process and then learn the advantages of higher dimensions and more essential semantic information. We hope to solve the above problems without increasing additional costs with the help of contrastive learning, which is first used to model contextual information. In fact, it is not easy to establish contrastive learning for learning context information representation in CRS. Traditional text contrastive learning [11–13] may lead to semantic changes when augmentation (such as translation, deletion, insertion, and transposition) is not appropriate. For example, when exploring whether 'I like my parents' and 'my parents like me' are similar, the machine may struggle to identify semantic changes accurately. In text contrastive learning, it is essential to construct

new examples while keeping the semantics of the original examples unchanged. Second, leveraging external data for data augmentation can enhance CRS performance. External structured KGs, such as ConceptNet [14] and DBpedia [15], provide word-level rich relationships and structured facts of item attributes, and external diversity reviews enrich the personalized features of item attributes. The existing methods [6, 7] only consider the fusion of the above specific types of external data in constructing item characteristics and capturing user preferences and do not fully mine and utilize all external data. At the same time, multi-type data correspond to different semantic spaces, and direct alignment causes the loss of effective information.

To this end, we propose a new multi-contrastive learning framework called MC-CRS. First, we introduce a contrastive learning framework without text augmentation and try to improve it at the coding level by treating the text encoder and its perturbed version as two encoders, obtaining two related sentence-level representations for contrastive learning, and mining the hidden semantic information of the dialogue. The semantics of the original and generated sentences in this process are identical, only the generated Embedding is different. Second, we pre-train the recommendation module with structured KG and diversity reviews to comprehensively understand the items' characteristics. First, we extract word and entity knowledge from multi-reviews of the item and adopt a contrastive learning approach to bridge the semantic gap between the item's multi-reviews, word KG, and entity KG. Finally, we fine-tune the pre-training results of the recommendation module with rich dialogue data. The primary contributions of this paper can be summarized as follows:

- We emphasize the importance of contextual information by introducing a contrastive learning framework, SimT, without text augmentation that perturbs at the coding level to mine the hidden semantic information in dialogues.
- We propose an augmented conversation recommendation framework, MC-CRS, to provide users with accurate and diverse recommendation services by deeply mining conversation information and fusing multiple types of external data for data augmentation.
- We conduct extensive experiments to validate that MC-CRS outperforms other baseline models in recommendation and dialogue tasks.

2 Related work

Due to its ability to dynamically capture user preferences through dialogue, CRS holds significant research value. In this section, we review related work from three perspectives: CRS, enhance CRS by utilizing external data, and contrastive learning.

Conversational recommender system CRS has garnered extensive attention because of its capability to capture dynamic user preferences through user feedback in real-time interactions to achieve accurate recommendations. Existing work mainly addresses the dialogue recommendation problem from two perspectives: One is attribute-based CRS [16, 17], which aims to complete the recommendation task in

the shortest number of turns [4]. The research centers on developing dialogue strategy modules, typically trained through reinforcement learning [16, 18]. Another is open-ended CRS [3, 19], which focuses on incorporating item information into the response text, aiming to provide users with an accurate and smooth dialogue experience. It typically includes both a recommendation module and a dialogue module. Our work focuses on an extended study of the second recommendation approach, given that the contextual information contained in a dialogue is extremely limited [6], subsequent studies [7, 8, 20] introduced external data (e.g., KG and reviews) to augment the item representations and help CRS improve the performance of interactive recommendations.

Enhance CRS by utilizing external data The use of KG incorporates item knowledge and thus improves the quality of conversational recommendation. KBRD [2] introduced KG into a CRS and proposed a knowledge-based CRS for the first time. KGSF [6] simultaneously uses relational graph convolutional network (RGCN) [21] and graph convolutional network (GCN) [22] to learn the user's structural and semantic knowledge on DBpedia [15] and ConceptNet [14], respectively, to solve the semantic gap problem with the help of mutual information. The above knowledge-based CRS approach is limited because it can only model part of the structured relationships between item and its features and cannot capture item information comprehensively. External multi-reviews can reflect users' emotional tendencies toward items and help generate personalized user recommendations. RevCore [7] proposes a review augmentation framework to extract entity knowledge from collected item reviews to enrich item features. C^2-CRS [8] uses a multi-granularity learning framework to design a coarse-to-fine semantic alignment approach to model user preferences. However, matching items with only a single review makes learning users' personalized preferences difficult. Latter [3] proposes a framework for comprehensive learning of item features. However, its simple way of information fusion will cause loss to the original representation when multiple types of data are fused. Although the above studies utilize external data to improve the performance of CRS, they focus on utilizing external data, while conversation context information, as a core component of CRS, has not been adequately modeled and mined.

Contrastive learning Contrastive learning allows learning a quality representation space from data [23]. The current mainstream practice is to construct sample pairs using data enhancement strategies to narrow the embeddings of positive sample pairs and increase the distance with negative samples. Contrast learning in the field of computer vision [24] generally corresponds to image clipping and deformation to form a positive sample, and in the field of natural language processing [25], researchers often de-phrase the text, re-translate using other languages, add Gaussian noise, and dropout to constitute positive examples. Unlike the computer vision domain, enhancement or perturbation of natural language can easily change the semantics, and False Positive has a significant impact on the model [26, 27]. How we can construct new examples by adding perturbations while keeping the semantics of the original example unchanged is a critical issue in contrastive learning.

In our research, we fully utilize contextual information and external data to learn the structural features of the item on the KG and the personalized features of the item in multi-reviews. We are internally introducing a contrastive learning

framework, SimT, that does not require augmentation to model contextual information, and externally eliminating semantic gaps between different data with the help of contrastive learning.

3 Preliminaries

The task of the CRS is to conduct multi-turn human–machine dialogue, and the agent learns user preferences according to contextual information so that users can get satisfactory item recommendations in at least T turns. The process mainly consists of two modules: (1) conversation module generates T turns of utterances, in each turn, the module can choose to directly recommend or continue to request, if the user successfully accepts the recommended item, then the conversation ends, otherwise it will continue to ask, until the maximum turn T is reached, and then automatically quit. (2) The recommendation module, according to the user's preference, selects a group of items that the user may be satisfied with for recommendation.

Dialogue contextual information The system mainly learns the user's preferences from the user's current dialogue, where user $u \in \mathcal{U}$ comes from set $\mathcal{U}$, item $i \in \mathcal{I}$ belongs to set $\mathcal{I}$, and word $w \in \mathcal{W}$ comes from vocabulary $\mathcal{W}$. The dialogue context C is composed of n utterance turns, represented as: $C = \{s_t\}_{t=1}^{n}$, s_t represents the tth turn of utterances. Each utterance is composed of a series of contextual words $s_t = \{w_j\}_{j=1}^{m}$.

Word knowledge graph In CRS, item i is mainly recommended by learning user preferences from the entities and words mentioned by the user in the current dialogues. Following previous work [6], we use the external lexical dataset ConceptNet [14] as a word-oriented KG $\mathcal{G}_v$. It stores semantic facts as triples $\langle w_1, r, w_2 \rangle$, where word $w \in \mathcal{W}$ reflects semantic knowledge in the dialogue, and r is a word relationship. Considering our CRS task, we extract words occurring in the dialogue corpus from ConceptNet.

Entity knowledge graph The structural relationship between items and entities is also significant. As shown in previous studies [2, 6], we introduce DBpedia [15] as an entity-oriented KG $\mathcal{G}_n$. In our definition, all items (such as movies) are also entities in $\mathcal{E}$. The triples in DBpedia are represented by $\langle e_1, r, e_2 \rangle$, where $e_1, e_2 \in \mathcal{E}$ is an item or entity from the entity set $\mathcal{E}$ and r is the relationship between entities. We extract the entities mentioned by users in the conversation with the help of entity join technology [2]. With these entities as the central node, we select the one-hop neighbors of these nodes.

Multi-reviews for items Following previous work [7], we adopt phrase selection to select N appropriate sentences from the collected brief introductions or explanations of items as multi-reviews $\mathcal{R}^i = \{r_1, r_2, \ldots, r_N\}$ for item i. All multi-reviews for items constitute the review dataset $\mathcal{R}$ and a review $r \in \mathcal{R}$ is composed of text $r = \{w_j\}_{j=1}^{m}$.

4 Methods

In order to get high-quality recommendations, we propose a new multi-contrastive learning framework for CRS, called MC-CRS, which encodes rich external data separately from three parts: contextual information coding module, word KG coding module and entity KG coding module. The contextual information coding module is a simple contrastive learning framework without text augmentation and introduces contrastive learning methods to make up for the semantic gap between multiple types of external data. We divide the recommendation module into two stages: pre-training and fine-tuning. First, in the pre-training stage, the multi-reviews of the item are taken as inputs, and the three-part encoder is used for coding learning to comprehensively understand the relationship between item and its features. Second, in the fine-tuning stage, the rich CRS dialogue data are taken as inputs, and the high-quality recommendations are realized after fine-tuning with the same framework as the pre-training. Next, we will introduce the structure of the recommendation module in the MC-CRS model of the MC-CRS model in detail, as shown in Fig. 1. It includes (1) contextual information encoder, (2) word/entity KG encoder, and (3) contrastive fusion stage. Considering that the recommendation module has the same architecture in the pre-training and fine-tuning stages, we take the conversation context C as input (fine-tuning phase) in the following introduction.

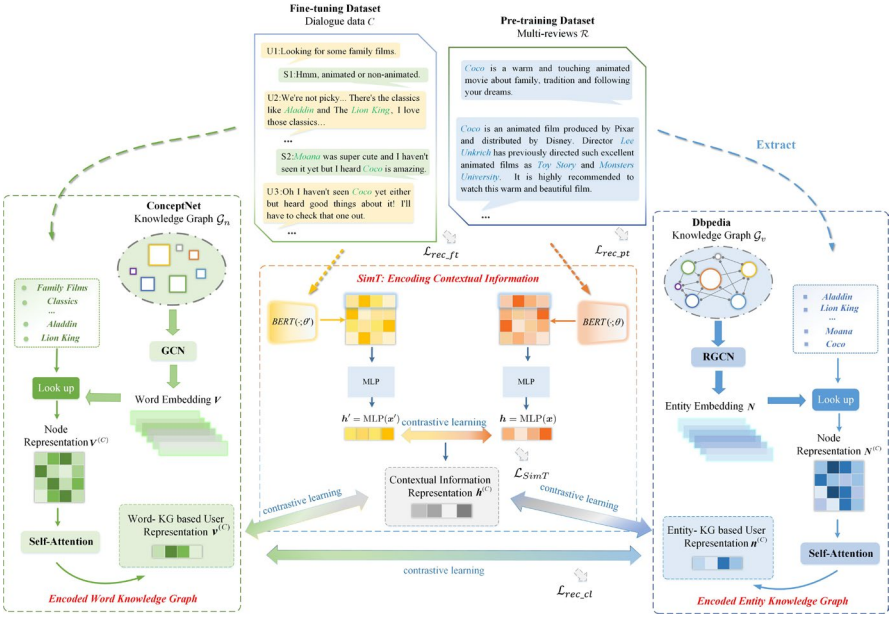


Fig. 1 Schematic overview in our recommendation module in MC-CRS

4.1 Encode the contextual information

4.1.1 Encode the dialogue contextual information

We use dialogue context $C = \{s_t\}_{t=1}^{n}$ to record the information interaction between the user and the agent. The utterance s_t in the tth turn is composed of a series of contextual words $s_t = \{w_j\}_{j=1}^{m}$. Usually, the utterance content in the same context C is closely related and represents the user's preference together. We splice the utterance in the dialogue context C into a word sequence $H^C = \{w_k^C\}_{k=1}^{L_c}$, where L_c represents the length of the word sequence of dialogue C. To obtain high-level semantic representation vectors in text, such as dialogue context, we utilize the BERT [28] model to encode contextual information. After work [3], we add a special token [CLS] before the text H^C and input it into BERT together and finally use the output vector $\boldsymbol{x}^{\mathrm{CLS}}$ corresponding to this symbol as the semantic representation $\boldsymbol{x}^{(C)}$ of the whole dialogue C, as shown below:

$$\left[\boldsymbol{x}^{\mathrm{CLS}}, \boldsymbol{x}^{w_1^C}, \boldsymbol{x}^{w_2^C}, \ldots, \boldsymbol{x}^{w_{L_c}^C}\right] = \mathrm{BERT}\left(\left[\mathrm{CLS}, w_1^C, w_2^C, \ldots, w_{L_c}^C\right]\right) \tag{1}$$

$$\boldsymbol{x}^{(C)} = \boldsymbol{x}^{\mathrm{CLS}} \tag{2}$$

where $\boldsymbol{x}^{\mathrm{CLS}} \in \mathbb{R}^{d_W}$ is the representation of the token [CLS] after inputting the dialogue context to the BERT, which can be used to understand and represent the semantic information of the whole input sentence, $\boldsymbol{x}^{(C)} \in \mathbb{R}^{d_W}$ is the sentence-level representation of dialogue C, and d_W is the dimension of the word vector.

4.1.2 Simple text contrastive learning framework

Contrastive learning can help representation learning to ignore noise interference and obtain deeper semantic information. We want to use contrastive learning to model the original conversation information without adding additional cost. Traditional text contrastive learning may lead to semantic changes when inappropriately augmented.

Receiving inspiration from the research conducted by Xia et al. [29] on graph node contrastive learning framework, we introduce a simple text contrastive learning framework, SimT, without text augmentation. After using different encoders to perturb the embedding representation of text data, we compare the semantic similarity between two perturbed encoders. We achieve contrastive learning by narrowing the vector representations of the same semantics in the vector space. First, we construct the encoder $\mathrm{BERT}(\cdot;\theta)$ and the perturbed encoder $\mathrm{BERT}(\cdot;\theta')$, performing operation 4.1.1. We take the embedded output of [CLS] token serves as the representation of the whole sentence and obtain two sentence-level representations $\boldsymbol{x}$ and $\boldsymbol{x}'$ in the same dialogue context C, respectively:

$$\boldsymbol{x} = \mathrm{BERT}(C;\theta), \boldsymbol{x}' = \mathrm{BERT}(C;\theta') \tag{3}$$

$$\theta_l' = \theta_l + \eta \cdot \Delta\theta_l; \Delta\theta_l \sim \mathcal{N}\left(0, \sigma_l^2\right) \tag{4}$$

where the parameter θ' of the perturbed encoder is obtained by using Eq. 4, θ_l and θ'_l are the lth weight parameters of the encoder and the encoder after perturbation, respectively. η adjusts the degree of perturbation, and $\Delta\theta_l$ represents the perturbation sampled from the normal distribution.

Chen et al.'s study [30] demonstrates that using a projection head can ensure that the embedding vectors of two encoders can be in the same space, improving the quality of representation. Here, we use a two-layer perceptron (MLP) to project to obtain $\boldsymbol{h}$ and $\boldsymbol{h}'$:

$$\boldsymbol{h} = \mathrm{MLP}(\boldsymbol{x}), \boldsymbol{h}' = \mathrm{MLP}(\boldsymbol{x}') \tag{5}$$

We randomly select N groups of dialogues for training, and consider embeddings $(\boldsymbol{h}_n, \boldsymbol{h}'_n)$ obtained from the same dialogues under two different perturbations as 'positive pair,' and embeddings $(\boldsymbol{h}_n, \boldsymbol{h}_{i\neq n})$ from different dialogues under the same perturbation view as negative sample pairs. The cosine similarity function, denoted as $\mathrm{sim}(\boldsymbol{h}, \boldsymbol{h}') = \boldsymbol{h}^\top \boldsymbol{h}' / \|\boldsymbol{h}\| \|\boldsymbol{h}'\|$, measures the extent of similarity between two instances. The contrast loss of the n th group of dialogues is defined as follows:

$$\mathcal{L}_{\mathrm{SimT}} = -\log \frac{\exp\left(\mathrm{sim}(\boldsymbol{h}_n, \boldsymbol{h}'_n)\right)/\tau)}{\sum_{i=1, i\neq n}^{N} \exp\left(\mathrm{sim}(\boldsymbol{h}_n, \boldsymbol{h}_i)\right)/\tau)} \tag{6}$$

where τ is the temperature parameter, $\boldsymbol{h}_n$ and $\boldsymbol{h}'_n$ distribution represents the two embedding vector representations of the nth dialogue. After contrastive learning, we take the output h of the original encoder as the representation of the final contextual information $\boldsymbol{h}^{(C)}$.

4.2 Encoding external knowledge graph information

4.2.1 Encoding entity-oriented knowledge graph

Referring to previous work [2, 6, 31], we learn structured knowledge from DBpedia [15] to enrich the dialogue content. DBpedia KG links all entities appearing in dialogue C to obtain structured information between item entities. We use RGCN [21], to encode the characteristics of each node in the graph to capture rich relationship information between entities. Formally, the embedding $\boldsymbol{n}_e^{(l+1)}$ of node e in the l+1th graph layer is calculated as follows:

$$\boldsymbol{n}_e^{(l+1)} = \sigma\Big(\sum_{r\in\mathcal{R}e'\in\mathcal{N}_e^r} \frac{1}{\mathrm{Z}_{e,r}} \boldsymbol{W}_r^{(l)} \boldsymbol{n}_{e'}^{(l)} + \boldsymbol{W}^{(l)} \boldsymbol{n}_e^{(l)}\Big) \tag{7}$$

where, $\boldsymbol{n}_e^{(l)} \in \mathbb{R}^{d_E}$ is the node representation of entity e at the lth layer, $\mathcal{N}_e^r$ represents the adjacent node set with e which there exists a relationship r, $\boldsymbol{W}_r^{(l)}$ and $\boldsymbol{W}^{(l)}$ are the learnable parameters at the lth layer, and $\mathrm{Z}_{e,r}$ is the normalizing factor.

We take the representation $\boldsymbol{n}_e^l$ at the lth layer as the final representation $\boldsymbol{n}_e$ of entity e, the computation of all entity embeddings forms the entity matrix $\boldsymbol{N}_e$. Using

the self-attention mechanism as shown in Eq. 8, after weighting the entities according to their importance, we get an entity knowledge representation $\boldsymbol{n}^{(C)}$ of the user:

$$\boldsymbol{n}^{(C)} = \boldsymbol{N}_e \cdot \boldsymbol{\alpha} \tag{8}$$

$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{b}^{\top} \cdot \tanh\left(\boldsymbol{W}_{\alpha}\boldsymbol{N}_e\right)) \tag{9}$$

where $\boldsymbol{\alpha}$ represents the attention weight vector employed for calculating the importance between entities, $\boldsymbol{W}_{\alpha}$ and $\boldsymbol{b}$ represent learnable parameters.

4.2.2 Encoding word-oriented knowledge graph

Learning structured knowledge in conversations helps to reflect user preferences, but it has shortcomings in generalization ability. We incorporate ConceptNet KG to facilitate the learning semantic relationships between words to address the problem above. Since ConceptNet contains a large amount of relational information that has little impact on recommendations, we do not consider relational information and choose to use GCN to capture rich semantic information at word-level. Specifically, given the dialogue C, we first use GCN to learn word embeddings on the knowledge graph $\mathcal{G}_v$:

$$\boldsymbol{V}^{(l)} = \text{ReLU}(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{V}^{(l-1)}\boldsymbol{W}^{(l)}) \tag{10}$$

where, $\boldsymbol{V}^{(l)} \in \mathbb{R}^{V \times d_w}$ is the representation matrix of nodes, $\boldsymbol{A}$ is the adjacency matrix of the graph, $\boldsymbol{D}$ is the diagonal matrix, and $\boldsymbol{W}^{(l)}$ is the learnable matrix of the l layer. Then, the user's knowledge semantic representation $\boldsymbol{v}^{(C)}$ is calculated by self-attention similar to that in Eq. 8.

4.3 Contrastive fusion based on data enhancement

In order to fully model user preferences, we previously captured context information from different perspectives, including (1) conversational context representation $\boldsymbol{h}^{(C)}$ (2) entity knowledge representation $\boldsymbol{n}^{(C)}$ (3) knowledge semantic representation $\boldsymbol{v}^{(C)}$. However, there exists a disparity in meaning between these sets of information. In order to ensure the semantic space of multi-type context data remains consistent, we need to fuse these three information representations effectively. As contrastive learning can identify common features among similar instances and differentiate the differences between different instances, we employ contrastive learning to learn the intrinsic characteristics of preference distribution across multi-types of data and their inherent correlations. We regard the three information representations of the same user u's preference for $(\boldsymbol{n}^{(C)}, \boldsymbol{v}^{(C)})$, $(\boldsymbol{v}^{(C)}, \boldsymbol{h}^{(C)})$ and $(\boldsymbol{h}^{(C)}, \boldsymbol{n}^{(C)})$ as positive examples and require that their representations should be more similar than those of other users u'. At the same time, we regard the representations of different users u' in the same batch as negative examples of u. Given a batch of examples $\mathcal{F}$ we calculate the sum of contrastive learning losses as follows:

$$\mathcal{L}_{rec\_cl} = \mathcal{L}_{CL}(\boldsymbol{n}^{(C)}, \boldsymbol{v}^{(C)}) + \mathcal{L}_{CL}(\boldsymbol{v}^{(C)}, \boldsymbol{h}^{(C)}) + \mathcal{L}_{CL}(\boldsymbol{h}^{(C)}, \boldsymbol{n}^{(C)}) \tag{11}$$

$$\mathcal{L}_{CL}(\boldsymbol{q}_u, \boldsymbol{q}_u^+) = -\log \frac{\exp\left(sim(\boldsymbol{q}_u, \boldsymbol{q}_u^+)/\tau\right)}{\sum_{u,u' \in \mathcal{F}; u \neq u'} \exp\left(sim(\boldsymbol{q}_u, \boldsymbol{q}_{u'}^-)/\tau\right)} \tag{12}$$

where $sim(\cdot, \cdot)$ is a cosine similarity function to measure the correlation between two representations, τ is a temperature hyperparameter, and u' represents all different users in a batch of samples $\mathcal{F}$ except u.

4.4 Recommendation module

4.4.1 Pre-training recommendation module

In the previous chapter, we learned rich data representations on different external data using the corresponding encoders. In the recommendation task, we aim to suggest the top-K items to CRS users based on these representations. We divide the recommendation module into two parts: pre-training and fine-tuning. First, we conduct pre-training by combining structured knowledge graph and diversity reviews to understand the characteristics of items comprehensively. Specifically, we take multi-reviews of an item as input, and for a certain item i, we extract text and knowledge information from its multi-reviews $\mathcal{R}^i = \{r_1, r_2, \ldots, r_N\}$. A review $r \in \mathbb{R}^i$ consists of text $H^r = \{w_k\}_{k=1}^{L_R}$, and extracts the KG entities $E^r = \{e_j\}_{j=1}^{L_E}$ and KG words $V^r = \{v_i\}_{i=1}^{L_V}$ involved in review r. The matching score between item i and its corresponding review r_i is higher than that between other items and r_i mark. First, the text contrastive Learning SimT encoder encodes the review r as $\boldsymbol{h}^{(r)}$, and RGCN and GCN, respectively, encode the KG entities and words extracted from the review as $\boldsymbol{n}^{(r)}$ and $\boldsymbol{v}^{(r)}$. Subsequently, we use the contrastive learning method to bridge the semantic gap between multi-reviews of an item, word knowledge graph and entity knowledge graph. At this stage, the entity representation has assimilated personalized features of items along with structured semantic knowledge from both the review context and word KG. We take the entity representation $\boldsymbol{n}^{(r)}$ after contrast fusion as the final fusion representation $\boldsymbol{e}^{(r)}$. Following this, we compute the matching score between item i and the fusion representation $\boldsymbol{e}^{(r)}$ of review r, and set the cross-entropy loss as:

$$P(i|r) = \text{softmax}(\boldsymbol{e}^{(r)} \cdot \boldsymbol{n}_i) \tag{13}$$

$$\mathcal{L}_{\text{rec\_pt}} = -\sum_{i=1}^{N} \sum_{r=1}^{M} p_{ir} \cdot \log\big(\text{P}(i|r)\big) + \lambda_{CL\_pt} \cdot \sum_{(\boldsymbol{q}_r, \boldsymbol{q}_r^+)} \mathcal{L}_{CL}(\boldsymbol{q}_r, \boldsymbol{q}_r^+) \tag{14}$$

where $\boldsymbol{n}_i$ is the embedding of item i, N is the number of items in the conversation corpus, i is the index of the item, M is the number of multi-reviews corresponding to the item, r is the review index, and $\mathcal{L}_{CL}(\boldsymbol{q}_r, \boldsymbol{q}_r^+)$ is the loss of multi-type data alignment.

Algorithm 1 The pre-training recommendation module algorithm of MC-CRS model.

Input: Reviews dataset $\mathcal{R}$, multi-reviews $\mathcal{R}^N = \{r_1, r_2, \cdots, r_M\}$ of item N, M is the number of multi-reviews, entity-oriented KG $\mathcal{G}_n$, and word-oriented KG $\mathcal{G}_v$.
Output: Model parameters Θ^s, Θ^f and Θ^r of the pre-training recommendation modul.

Randomly initialize all node embeddings and model parameters in graphs $\mathcal{G}_n$, $\mathcal{G}_v$.
for i=1 to N **do**
  for r=1 to M **do**
    Acquire contextual information representations $\boldsymbol{h}^{(r)}$ from review r by SimT using Eq. 3 to Eq. 5.
    Perform gradient descent on Eq. 6 w.r.t. Θ^s.
    Acquire entitys' and words' representations $\boldsymbol{N}^{(r)}$ and $\boldsymbol{V}^{(r)}$ from $\mathcal{G}_n$, $\mathcal{G}_v$ by Eq. 7 and Eq. 10, respectively.
    Acquire $\boldsymbol{n}^{(r)}$ and $\boldsymbol{v}^{(r)}$ by self-attention using Eq. 8, respectively.
    Acquire $\boldsymbol{e}^{(r)}$ using Eq. 11 to fuse multi-types of data by contrastive learning.
    Perform gradient descent on Eq. 12 w.r.t. Θ^f.
    Compute $P(i|r)$ using Eq. 13.
    Perform gradient descent on Eq. 14 w.r.t. Θ^r.
  end for
end for
return Θ^s, Θ^f and Θ^r.

In the pre-training recommendation module, we need to learn three sets of parameters, namely SimT module, contrastive fusion module and recommendation module, which are Θ^s, Θ^f and Θ^r, respectively. We present the pre-trained recommendation module algorithm for the MC-CRS model in Algorithm 1.

4.4.2 Fine-tuning the recommendation module

After the pre-training, the model has a certain understanding of the relationship between items and their features. Then, we fine-tune the pre-training results of the recommendation module through rich conversation data to achieve high-quality recommendations. Similar to the fine-tuning step, we take the conversation C in the CRS dataset as input and obtain (1) the conversation context representation $\boldsymbol{h}^{(C)}$ (2) the entity knowledge representation $\boldsymbol{n}^{(C)}$ (3) the knowledge semantic representation $\boldsymbol{v}^{(C)}$. Then, through contrastive learning, we fuse the inherent features of the preference distribution between multiple data as well as their internal correlations. The final output is the entity knowledge representation $\boldsymbol{n}^{(C)}$ as the user preference representation $\boldsymbol{e}^{(u)}$. Similarly, the matching score of each item i with $\boldsymbol{e}^{(u)}$ is calculated as follows:

$$P(i|C) = \text{softmax}(\boldsymbol{e}^{(u)} \cdot \boldsymbol{n}_i) \tag{15}$$

$$\mathcal{L}_{\text{rec\_ft}} = -\sum_{j=1}^{K}\sum_{i=1}^{N} p_{ij} \cdot \log\left(P(i|j)\right) + \lambda_{\text{CL\_ft}} \cdot \sum_{(\boldsymbol{q}_u, \boldsymbol{q}_u^+)} \mathcal{L}_{\text{CL}}(\boldsymbol{q}_u, \boldsymbol{q}_u^+) \tag{16}$$

where K is the number of dialogues, j is the index of the dialogue, N represents the number of items, and i denotes their index, and $\mathcal{L}_{\text{CL}}(\boldsymbol{q}_u, \boldsymbol{q}_u^+)$ is the loss of multi-type data alignment.

In the subsequent experiments, we will also demonstrate that employing pre-training loss as a regularizer in the subsequent experiments, we will also demonstrate that employing pre-training loss as a regularizer. The loss L_{rec} of the final training target of the recommendation module is expressed as follows:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{rec\_ft}} + \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec\_pt}} \tag{17}$$

4.5 Conversation module

We will introduce how to generate smooth reply discourse in dialogue tasks. Inspired by KGSF [6], we establish an advanced response module using various types of external data, incorporating multiple cross-attention layers into the standard transformer decoder, which helps to effectively inject useful information from context and reviews into the reply. The following transformation chain can describe the implementation steps: generated word → dialogue context → multi-reviews. The equation can be abbreviated as follows:

$$\boldsymbol{Y}^l = \text{Decoder}(\boldsymbol{Y}^{l-1}, C, R) \tag{18}$$

where $\boldsymbol{Y}^l$ is the output representation matrix of the lth layer of the decoder.

At the same time, we also need to include relevant entity information and keywords in the generated response. We employ a replication mechanism to augment the representation of a specific type of token, as follows:

$$Pr(y_t|y_1, \ldots, y_{t-1}) = Pr_1(y_t|\boldsymbol{Y}_t) + Pr_2(y_t|\boldsymbol{Y}_t, C) + Pr_3(y_t|\boldsymbol{Y}_t, R) \tag{19}$$

where $Pr_1(\cdot)$ represents the generation probability function on the vocabulary with the decoder output $\boldsymbol{Y}_t$ as input, while $Pr_2(\cdot)$ and $Pr_3(\cdot)$ represent the replication probabilities implemented in the original context and review, respectively, following the copy mechanism [32]. Then, we train the conversation module using cross-entropy loss:

$$L_{gen} = -\frac{1}{K}\sum_{t=1}^{K} \log(\Pr\left(s_t|s_1, \ldots, s_{t-1}\right)) \tag{20}$$

where K is the number of conversation turns, s_t represents the tth turn of discourse in the conversation.

5 Experiments

To evaluate the efficacy of MC-CRS, a set of experiments were carried out to address the following pivotal research inquiries (RQs):

- (RQ1) How does the performance of our model in conversational and recommendation tasks compare to that of all baselines?
- (RQ2) How does various components of MC-CRS affect its performance?
- (RQ3) How does different hyperparameter configurations impact the performance of MC-CRS?
- (RQ4) How does the MC-CRS's multi-contrastive modules affect CRS performance?

5.1 Experimental settings

In this section, we will introduce the details of our experiments, encompassing our datasets, baseline models, and evaluation metrics.

5.1.1 Datasets

We comprehensively evaluated our model on two widely used public real-world datasets, in Chinese (TG-ReDial [33]) and English (ReDial [5]). The ReDial dataset focuses on conversations in the movie domain, where one party asks and the other party provides a recommendation. The dataset contains 1006 conversations involving 504 users and 51,699 movies. TG-ReDial is a topic-guided conversational recommendation dataset, where conversations not only involve movie recommendations but also cover multiple related topics, focusing on guiding the natural transition of conversations from non-recommendation scenarios to recommendation scenarios. It contains 10,000 conversations involving 1482 users and 33,834 movies. For both datasets, we experimented with a training, validation, and test sample ratio of approximately 8:1:1. To obtain review data for each movie, we scraped the top 15 reviews for each movie from two well-known movie review websites, IMDB (for the ReDial dataset) and douban (for the TG-ReDial dataset).

5.1.2 Evaluation metrics

We evaluated the recommendation task separately from the conversation task. In the recommendation task, we used *Recall@k* ($R@k$, $k = 1, 10, 50$) [5] to calculate the ratio between the top-K items recommended by the system and the ground truth recommendation. In the conversation task, we combined automatic and human evaluation. The automatic evaluation metric was the Distinct n-gram ($n = 2, 3, 4$) [34], which measures the diversity of generated discourse at the sentence level. The main purpose of our conversation component was to mimic real responses while providing a successful recommendation, so we invited three annotators to perform human

evaluation to evaluate our model more comprehensively. We mainly focused on the information, fluency, and relevance of the responses produced by the model and the baseline experiments [3]. Information evaluated whether the response contained rich entity knowledge, and relevance evaluated whether the response contained context-related features. The evaluation score of each annotator ranged from 0 to 2, and we took the average score of the three people as the human evaluation result.

5.1.3 Implementation details

We implemented our model using PyTorch and CRSLab, with a word-level representation of 768 dimensions using the default settings of BERT, and an entity-level representation of 128 dimensions. Seven reviews were selected for each item in the experiment, and KG entities were extracted from the review text or conversation context using entity-linking techniques. In addition, the maximum length of the conversation and the review was set to 128. The number of layers for both R-GCN and GCN was set to 1 layer, and the normalization constant $Z_{e,r}$ of R-GCN was set to 1. We used the default parameter settings of the Adam optimizer. During the experiment, the batch size, learning rate, and the $\mathcal{L}_{\text{rec}}$ were set to 128, 0.001, and 0.7, respectively. The temperature of contrast learning was set to 0.07. The baseline model followed its own implementation to set hyperparameters for optimal performance and to ensure fair comparisons.

5.1.4 Baselines

We chose representative recommendation or conversation models as benchmarks to evaluate the superiority of MC-CRS:

Transformer [35]: Generates conversational responses using a transformer-based encoder and decoder approach.

TextCNN [36]: It encodes the current conversation and extracts user preferences from the dialogue context via a convolutional neural network (CNN)-based model.

ReDial [5]: This model collects the ReDial dataset and uses autoencoder learning to make recommendations in cold start settings.

KBRD [2]: This model enhances user preference representation to improve the performance of the recommendation system by introducing an external DBpedia KG.

KGSF [6]: It combines word-oriented and entity-oriented KG to enhance data representation and employs maximizing mutual information to align the semantic space of various elements within the ongoing discussion.

RevCore [7]: This model enhances the recommendation module and conversation generation module with user reviews of movies.

C^2-CRS [8]: It designs a new coarse-to-fine contrastive learning framework to model user preferences.

LATTE [3]: It leverages multi-reviews and KGs to transform CRS into an expert in learning item characteristics, achieving high-quality recommendations.

Among these baselines, TextCNN is a classical recommendation method, and the rest of the CRS methods only consider the current conversation without historical interaction records. We named our proposed model MC-CRS.

5.2 Overall performance (RQ1)

5.2.1 Evaluation on recommendation task

We conduct comparative experiments on the ReDial and TG-ReDial datasets to explore the effectiveness of MC-CRS in recommendation tasks, the results are presented in Table 1. First, the conversational recommendation method shows obvious advantages compared with non-CRS methods (e.g., TextCNN). This is because traditional recommendation methods [37, 38] such as TextCNN often ignore the integration of two core functions in CRS, the conversation module and the recommendation module, while CRS can better combine these two modules, thus providing users with more accurate and personalized recommendations. Second, the method using external data (i.e., KBRD, KGSF, RevCore, C^2-CRS, and LATTE) is superior to the method without external data (i.e., TextCNN, ReDial). This is because the available information in a dialogue context is often sparse and difficult to extract and utilize fully. With the help of external KG or reviews and other data, we can more comprehensively predict users' preferences. Third, in most cases, the method using both KG and reviews (i.e., RevCore, C^2-CRS, LATTE) is superior to the method using only KG (i.e., KBRD and KGSF), which further demonstrates that the effective combination of external reviews enriches the dialogue content and enhance personalized representations of item features, thus improving the accuracy of recommendations. Finally, and most importantly, our MC-CRS method is significantly better than all baseline methods on the recommendation task. This indicates that MC-CRS uses multi-types of external information for data augmentation and builds a better user representation. The model uses multi-contrastive learning frameworks to enhance its effectiveness. On the one hand, method model can learn contextual information better, while the other enables the model to effectively integrate multi-types of external data, accurately modeling the diverse preferences of users.

Table 1 Comparison of MC-CRS and 7 competitors on the recommendation task, a number marked with* indicates the model's superior performance

	ReDial			TG-ReDial		
Models	*R*@1	*R*@10	*R*@50	*R*@1	*R*@10	*R*@50
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	0.183	0.005	0.030	0.074
RevCore	0.044	0.205	0.394	0.004	0.029	0.074
C^2-CRS	0.052	0.216	0.407	0.007	0.032	0.078
LATTE	0.050	0.208	0.397	0.005	0.031	0.076
MC-CRS	0.057*	0.221*	0.412*	0.009*	0.036*	0.082*

5.2.2 Evaluation on conversation task

To verify the effectiveness of our proposed method on the conversation generation task, we conducted automatic and human evaluations on the ReDial and TG-ReDial datasets, respectively.

Automatic evaluation The results are shown in Table 2. Our model MC-CRS outperforms all baseline methods in all evaluation metrics compared to the baseline can effectively generate diverse discourse. First, the relatively poor performance of ReDial among all CRS methods further validates that the introduction of external knowledge graphs of entities and semantic similarity information can enhance contextual entity and item representations. At the same time, this entity information can also be used to guide the word probability distribution of the conversation module, thus contributing to the generation of high-quality replies. Second, the methods that use KG and reviews at the same time (i.e., RevCore, C^2-CRS, LATTE) generally score higher on the distinct *n*-gram metric than the methods that do not use reviews. This result is because the information from reviews, as a richer and more easily accessible external resource, can effectively improve the diversity of replies generated by the system. Finally, RevCore exhibits poor performance compared to other methods utilizing knowledge graphs and reviews simultaneously. One possible reason is that it selects several reviews for each movie to guarantee the number without considering that the reviews may contain noise.

Human evaluation We mainly focus on three aspects of human evaluation: informative, fluency, and relevance to evaluate the generated answers. The experimental results in Table 3, clearly show that MC-CRS shows significant advantages in these three aspects, generating more fluent, contextually relevant, and informative discourse. One of the main reasons is that by integrating movie reviews and external KG, MC-CRS can fully learn and understand the context information during the training process, and accurately capture the user's interests and preferences, to integrate more relevant information when generating responses.

Table 2 MC-CRS automatically evaluates the results of a conversation generation task with 7 competitors, a number marked with* indicates the model's superior performance

	ReDial			TG-ReDial		
Models	Dist-2	Dist-3	Dist-4	Dist-2	Dist-3	Dist-4
Transformer	0.138	0.142	0.214	0.102	0.220	0.204
ReDial	0.215	0.238	0.351	0.185	0.228	0.347
KBRD	0.256	0.307	0.485	0.198	0.336	0.428
KGSF	0.273	0.364	0.523	0.242	0.386	0.495
RevCore	0.340	0.436	0.556	0.290	0.416	0.546
C^2-CRS	0.417	0.559	0.607	0.429	0.592	0.683
LATTE	0.445	0.571	0.635	0.437	0.653	0.761
MC-CRS	0.463*	0.647*	0.705*	0.441*	0.672*	0.805*

Table 3 MC-CRS competes with 7 competitors on the dialogue generation task for the results of human evaluation, a number marked with* indicates the model's superior performance

Model	Informativeness	Fluency	Relevance
Transformer	0.95	0.90	0.23
ReDial	1.34	1.14	0.15
KBRD	1.02	1.29	0.26
KGSF	1.45	1.35	0.40
RevCore	1.36	1.46	0.37
C^2-CRS	1.41	1.53	0.62
LATTE	1.45	1.48	1.09
MC-CRS	1.47*	1.56*	1.21*

5.3 Ablation studies (RQ2)

To evaluate the effectiveness of each part of MC-CRS, we also conducted ablation studies based on different variants of the model, including (1) MC-CRS(no-C-F) to remove the multiple contrast learning module (contextual information contrast and fusion contrast); (2) MC-CRS(no-C) to remove the contextual information contrast learning module; (3) MC-CRS(no-F) to remove the fusion multi-type data contrast learning module; (4) MC-CRS(no-PT) to remove the multi-review pre-training part; (5) MC-CRS(no-CID) to remove the contextual information coding part; (6) MC-CRS(no-E-KG) to remove the structured data (i.e., entity knowledge graph); (7) MC-CRS(no-W-KG) to remove the structured data (i.e., word knowledge graph);

In Fig. 2, we used three different metrics to evaluate the recommendation accuracy of seven variants. The results indicate that deleting any type of data representation will decrease recommendation results, suggesting that multi-types of external data are beneficial for enhancing data representation. First, the results show a significant decrease upon removing the multi-contrastive learning module, and the multi-type data fusion module in multi-contrastive is essential. This also confirms that using contrastive learning to fuse multi-type external data can effectively bridge the semantic divide among various data signals. Second, the pre-training part of removing multi-review produces a performance worse than our method. This indicates that our model is effective in pre-training through multi-review and extracted KG entities. Contextual information coding module also significantly impacts our model, indicating that the original dialogue context remains a crucial source for user preference acquisition in conversation recommendation. In the upcoming experiments, we will also substitute the multi-contrastive learning module (used for learning contextual information and fusing multi-type data) with the existing method to validate the efficacy of our model.

5.4 Parameter sensitive analysis (RQ3)

5.4.1 Choose the number of reviews

Item reviews represent the personalized impression of users on items. Our method realizes data augmentation with the help of users' reviews on items and then learns

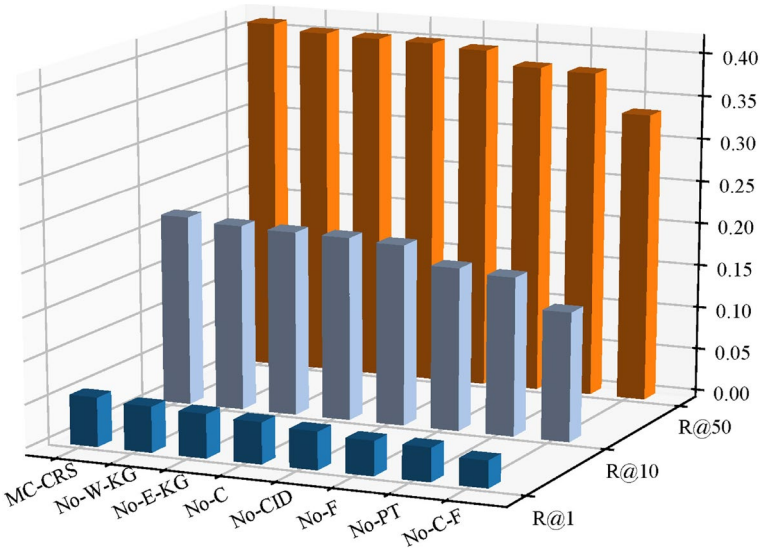
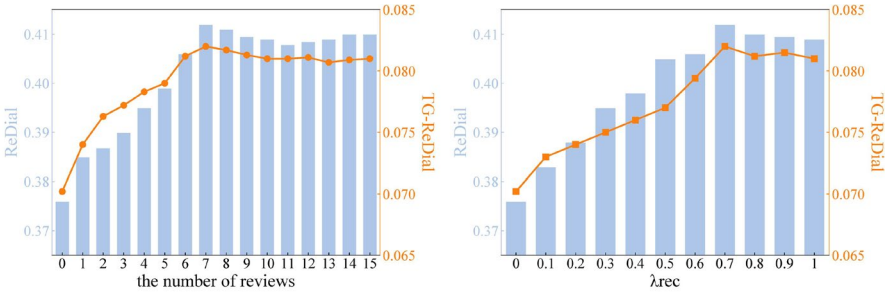


Fig. 2 The results of the ablation study for the recommendation task evaluation. C and F refer to the context information conversation module and the fusion conversation module, respectively; PT is the pre-training part of the model; CID refers to the contextual information coding part; E-KG AND W-KG refer to the external structured data entity KG and word KG, respectively

the diversity features of items. Therefore, we need to select multi-reviews to pre-train the recommendation module and use the KG and multi-reviews to capture the structural and personalized features of items. Here, we explore how the number of reviews affects the recommendation accuracy through experiments. The goal is to minimize the necessary reviews for the project while maintaining model efficiency. The number of reviews we select ranges from 0 to 15. The results of the experiment are displayed in Fig. 3a. In the process of the number of reviews gradually increasing from 0 to 7, the model gradually understands the relationship between items and item features and the recommendation accuracy also shows a steady increase. After the number of reviews reaches 7, the recommendation accuracy tends to be stable, and there is almost no noticeable change. From the perspective of memory resource utilization and time efficiency, seven reviews are enough for the model to comprehensively model the relationship between items and item features, and achieve a satisfactory recommendation effect for users.

5.4.2 Fine-tune parameters

In the recommendation module, we adopted a two-stage training strategy of pre-training and fine-tuning. In these two stages, we used the loss of the pre-training stage as the regularization term of the fine-tuning stage to better guide the learning process of the model, so as to achieve the goal of recommending top-K items to users. In order to explore the influence of these two stages on the accuracy of recommendation, we adjusted the accuracy of the parameter λ_{rec}, whose value range is



(a) Performance of MC-CRS at different number of reviews

(b) Performance of MC-CRS at different λ_{rec}

Fig. 3 Parameter sensitivity analysis and performance evaluation of MC-CRS on $R@50$ with varied review counts and λ_{rec} parameters

between 0 and 1, and the increment is set to 0.1. The results of the experiment are displayed in Fig. 3b. When λ_{rec}=0, i.e., without considering the pre-training step, the performance of the entire recommendation is the worst. This result fully proves the importance of the pre-training step. By learning the relationship between items and their features according to reviews and KG, the pre-training stage can provide effective a priori knowledge for the model and improve the accuracy of the model recommendation. With the accuracy of λ_{rec} increasing from 0 to 0.7, the success rate of recommendation shows an increasing trend. This phenomenon suggests that the proper utilization of pre-training loss enhances fine-tuning accuracy, resulting in improved model performance. After 0.7, the success rate of recommendation gradually stabilizes and is slightly down. The above result highlights the potential limitations of relying too heavily on pre-training loss during the fine-tuning phase, which may impede the model's learning capacity, which cannot fully adapt to the needs of specific tasks, thus reducing the accuracy of recommendation.

5.5 Effectiveness of multiple contrastive learning modules (RQ4)

5.5.1 Contrastive fusion of contextual information

In CRS, the dialogue context carries important user information. Accurately understanding the user's emotions and intentions from the original natural language and obtaining an effective representation of the dialogue can improve the accuracy of recommendations and achieve long-term retention of users. In MC-CRS, we utilize the contrastive learning method SimT at the coding level to acquire and learn contextual information. In order to assess the efficacy of the module in the representation of context information, we replaced the module with two conventional methods for encoding context information and conducted a series of experiments.

As shown in Fig. 4, we used three measures on the ReDial dataset to evaluate recommendation accuracy in four modes, one of which does not consider the encoding of contextual information. The final experimental results show that the encoding of context information without consideration is the worst, which also confirms that

the learning and modeling of original dialogue are essential in CRS. The effect of our comparative method SimT is better than transformer [35] and BERT [28]. This shows that the method can mine the deep semantic information hidden in the dialogue and accurately model user preferences, thus effectively improving the performance of dialogue recommendation. In most cases, BERT is better than transformer. This phenomenon may be caused by the fact that BERT only uses the encoder part of the transformer as its model structure. The decoder part not only fails to contribute to learning but may also interfere with it. Training both parts at the same time will also cost more computing power and time.

5.5.2 Contrastive fusion of multiple types of information

As a result of the significant semantic variance across multiple data types, fusing the underlying semantics is the key to effectively utilizing information. In order to explore the effectiveness of contrastive learning in multi-information fusion in MC-CRS, we replace the information fusion module in this paper with the conventional information fusion mode in other CRS studies [1, 3]. These studies mainly design fusion models for specific types of external data, and the modeling and utilization of multi-type external data are not comprehensive. We apply it to multi-type data fusion with a slight change, including: (1) multi-head attention (MHA) [35], which takes the word representation in the current conversation (represented by $\boldsymbol{H}^{(c)}$) as a query to simultaneously focus on the representation of

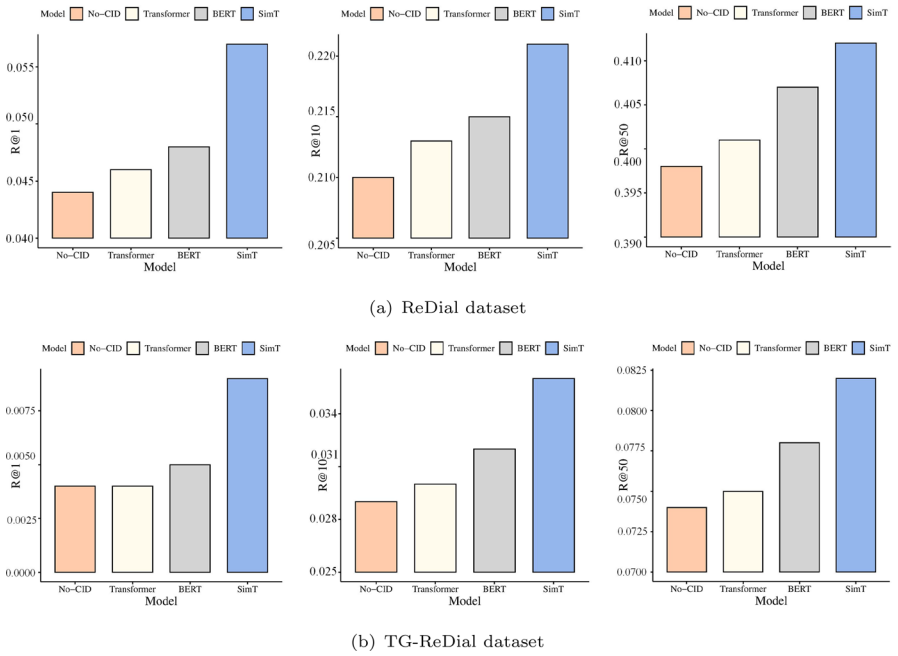


Fig. 4 Accuracy of recommendations for different methods of encoding contextual information methods in ReDial and TG-ReDial datasets

the external knowledge graph in the entity-level representation $\boldsymbol{n}^{(c)}$ and the word-level representation $\boldsymbol{v}^{(c)}$. Then, pooling layers (such as average pooling) are used to fuse the external user structure information and user conversation information. (2) Gate recurrent unit (GRU) [39], the entities $\boldsymbol{v}^{(c)}$ and $\boldsymbol{n}^{(c)}$ are reside in separate semantic spaces [6, 8, 34]. This necessitates the initial projection of word-level representations into the entity space, followed by using the GRU to combine both representations. Finally, user representations are obtained through multi-method representation learning. The results are depicted in Fig. 5, after conducting multiple experiments. The fusion performance of MHA is better than that of GRU. Using MHA can output the encoded representation information of the attention layer in different subspaces, thus improving the performance of the model. However, the word-level representation is projected into entity space before the processing of GRU, which may cause the loss of effective information. The effect of multi-type information fusion with contrast learning in our model MC-CRS is better than the other two. This is because contrast learning can effectively close the natural semantic gap between session data and external data, expand to include more types of external data, and model and utilize multi-type external data. However, the other two directly align the semantic space, which will damage the original representation performance.

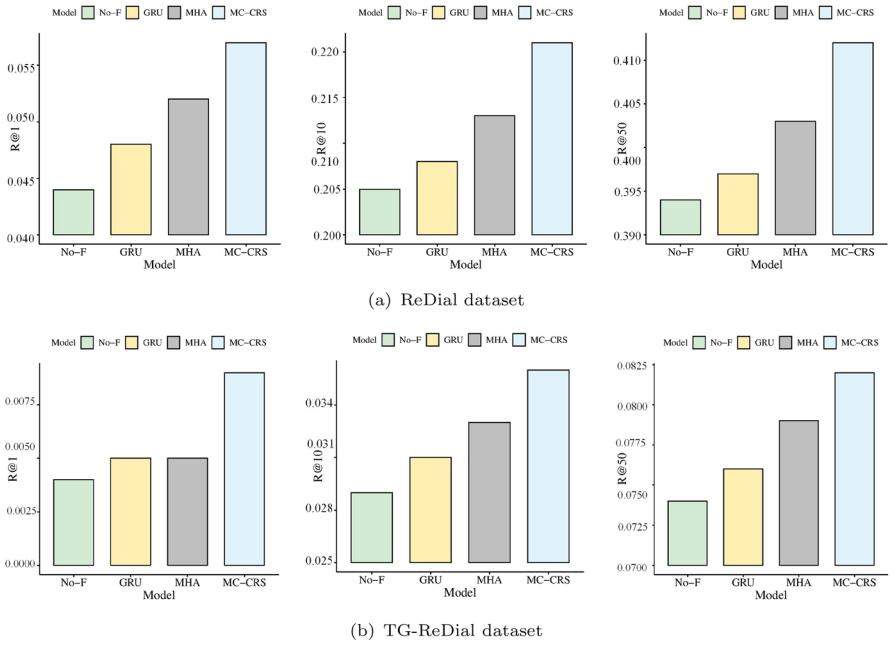


Fig. 5 Accuracy of recommendations for different multi-type data fusion methods in ReDial and TG-ReDial datasets

6 Conclusion and future work

In this paper, we propose an enhanced conversational recommendation approach founded on multi-contrastive learning. With the assistance of two external KGs and multi-reviews of the item, data augmentation is implemented, and contextual information is deeply excavated through internal and external contrastive learning mechanisms. Multi-types of external data are integrated to offer users precise and diverse recommendation services. The extensive experimental results fully validate the superiority of the proposed method on conversation and recommendation tasks.

In future work, the contrastive learning algorithm we introduce without data augmentation can be flexibly applied to representation learning on other tasks. Additionally, we can extend the multi-type external data fusion method to other domains of CRS. Specifically, we will investigate how to utilize multi-type external data, especially multi-review information, to enhance the interpretability of recommendation utterances in CRS, and explore how to efficiently enrich item reviews in the case of sparse data. Finally, we intend to integrate historical user-item interaction data to construct a more accurate model of user behavior.

Author Contributions X.L and J.Y conceptualized and devised the overarching research goals and ideas, developed models, implemented computer code along with supporting algorithms, synthesized study data, authored the initial draft, and provided critical commentary or revisions throughout both pre-and post-publication stages. P.L and Y.H were responsible for gathering experimental resources, provisioning study materials and laboratory samples, presenting data, and managing and maintaining research data.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability statement The datasets analyzed during the current study were all derived from the following public domain resources. [https://redialdata.github.io/; https://github.com/RUCAIBox/TG-ReDial].

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent to participate The authors declare that they agree to participate.

Consent for publication The authors declare that they agree to publish.

References

1. Li S, Xie R, Zhu Y, Ao X, Zhuang F, He Q (2022) User-centric conversational recommendation with multi-aspect user modeling. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 223–233
2. Chen Q, Lin J, Zhang Y, Ding M, Cen Y, Yang H, Tang J (2019) 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), pp 1803–1813

3. Kim T, Yu J, Shin W-Y, Lee H, Im J-h, Kim S-W (2023) Latte: A framework for learning item-features to make a domain-expert for effective conversational recommendation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp 1144–1153
4. Lei W, He X, Miao Y, Wu Q, Hong R, Kan M-Y, Chua T-S (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, pp 304–312
5. Li R, Ebrahimi Kahou S, Schulz H, Michalski V, Charlin L, Pal C (2018) Towards deep conversational recommendations. In: Advances in neural information processing systems, vol 31
6. Zhou K, Zhao W.X, Bian S, Zhou Y, Wen J-R, Yu J (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, pp 1006–1014
7. Lu Y, Bao J, Song Y, Ma Z, Cui S, Wu Y, He X (2021) RevCore: review-augmented conversational recommendation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP, pp 1161–1173
8. Zhou Y, Zhou K, Zhao W.X, Wang C, Jiang P, Hu H (2022) C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp 1488–1496
9. Hassani K, Khasahmadi AH (2020) Contrastive multi-view representation learning on graphs. In: International Conference on Machine Learning. PMLR, pp 4116–4126
10. Wu J, Wang X, Feng F, He X, Chen L, Lian J, Xie X (2021) Self-supervised graph learning for recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 726–735
11. Wu L, Li J, Wang Y, Meng Q, Qin T, Chen W, Zhang M, Liu T-Y (2021) R-drop: regularized dropout for neural networks. In: Advances in neural information processing systems, vol 34, pp 10890–10905
12. Gao T, Yao X, Chen D (2021) SimCSE: simple contrastive learning of sentence embeddings. In: 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. Association for Computational Linguistics (ACL), pp 6894–6910
13. Wu X, Gao C, Zang L, Han J, Wang Z, Hu S (2022) ESimCSE: enhanced sample building method for contrastive learning of unsupervised sentence embedding. In: Proceedings of the 29th International Conference on Computational Linguistics, pp 3898–3907
14. Speer RC (2017) 5.5: an open multilingual graph of general knowledge/r.speer'. In: J. Chin, C. Havasi//Thirty-First AAAI Conference on Artificial Intelligence
15. Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) DBpedia–a crystallization point for the web of data. J. Web Semantics 7(3):154–165
16. Zhang Y, Wu L, Shen Q, Pang Y, Wei Z, Xu F, Long B, Pei J (2022) Multiple choice questions based multi-interest policy learning for conversational recommendation. In: Proceedings of the ACM Web Conference 2022, pp 2153–2162
17. Zhou K, Zhao W.X, Wang H, Wang S, Zhang F, Wang Z, Wen J-R (2020) Leveraging historical interaction data for improving conversational recommender system. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp 2349–2352
18. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) Deep reinforcement learning: a brief survey. IEEE Signal Process Mag 34(6):26–38
19. Liang Z, Hu H, Xu C, Miao J, He Y, Chen Y, Geng X, Liang F, Jiang D (2021) Learning neural templates for recommender dialogue system. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 7821–7833
20. Ma W, Takanobu R, Huang M (2021) CR-Walker: tree-structured graph reasoning and dialog acts for conversational recommendation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 1839–1851
21. Schlichtkrull M, Kipf TN, Bloem P, Van Den Berg R, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15. Springer, pp 593–607
22. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations
23. Zhang Y, Zhu H, Wang Y, Xu N, Li X, Zhao B (2022) A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 4892–4903

24. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9729–9738
25. Gao T, Yao X, Chen D (2021) SimCSE: simple contrastive learning of sentence embeddings. In: 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. Association for Computational Linguistics (ACL), pp 6894–6910
26. Yan Y, Li R, Wang S, Zhang F, Wu W, Xu WC (2021) A contrastive framework for self-supervised sentence representation transfer. arXiv:2105.11741
27. Chuang C-Y, Robinson J, Lin Y-C, Torralba A, Jegelka S (2020) Debiased contrastive learning. In: Advances in neural information processing systems, vol 33, pp 8765–8775
28. Kenton JDM-WC, Toutanova LK (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp 4171–4186
29. Xia J, Wu L, Chen J, Hu B, Li SZ (2022) SimGRACE: a simple framework for graph contrastive learning without data augmentation. In: Proceedings of the ACM Web Conference 2022, pp 1070–1079
30. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp 1597–1607
31. Li W, Wei W, Qu X, Mao X-L, Yuan Y, Xie W, Chen D (2023) TREA: tree-structure reasoning schema for conversational recommendation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 2970–2982
32. Gu J, Lu Z, Li H, Li VO (2016) Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics
33. Zhou K, Zhou Y, Zhao WX, Wang X, Wen J-R (2020) Towards topic-guided conversational recommender system. In: Proceedings of the 28th International Conference on Computational Linguistics, pp 4128–4139
34. Wang X, Zhou K, Wen J-R, Zhao WX (2022) Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp 1929–1937
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30
36. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1746–1751
37. Wu S, Tang Y, Zhu Y, Wang L, Xie X, Tan T (2019) Session-based recommendation with graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 346–353
38. Wei W, Zhao S, Zou D (2023) Recommendation system: a survey and new perspectives. World Sci Annu Rev Artif Intell 1:2330001
39. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014