CLICK: Contrastive Learning for Injecting Contextual Knowledge to Conversational Recommender System

Hyeongjun Yang, Heesoo Won, Youbin Ahn, and Kyong-Ho Lee
Department of Computer Science, Yonsei University
{edbm95, hswon97, ybahn, khlee98}@yonsei.ac.kr

Abstract

Conversational recommender systems (CRSs) capture a user preference through a conversation. However, the existing CRSs lack capturing comprehensive user preferences. This is because the items mentioned in a conversation are mainly regarded as a user preference. Thus, they have limitations in identifying a user preference from a dialogue context expressed without preferred items. Inspired by the characteristic of an online recommendation community where participants identify a context of a recommendation request and then comment with appropriate items, we exploit the Reddit data. Specifically, we propose a **C**ontrastive **L**earning approach for **I**njecting **C**ontextual **K**nowledge (**CLICK**) from the Reddit data to the CRS task, which facilitates the capture of a context-level user preference from a dialogue context, regardless of the existence of preferred item-entities. Moreover, we devise a relevance-enhanced contrastive learning loss to consider the fine-grained reflection of multiple recommendable items. We further develop a response generation module to generate a persuasive rationale for a recommendation. Extensive experiments on the benchmark CRS dataset show the effectiveness of CLICK, achieving significant improvements over state-of-the-art methods.

1 Introduction

Recommender systems help users find potential items of interest in the rapidly expanding pool of candidates. To do so, the recommender systems aim to accurately identify user preferences. Thus, traditional recommender systems (Guo et al., 2020) utilize click or purchase data to obtain user preferences. However, such traditional approaches lack dynamically modeling user preferences since the data has a temporal validity and does not offer a diverse clue to user information. Conversational recommender system (CRS) is a research area that overcomes the limitation of the traditional

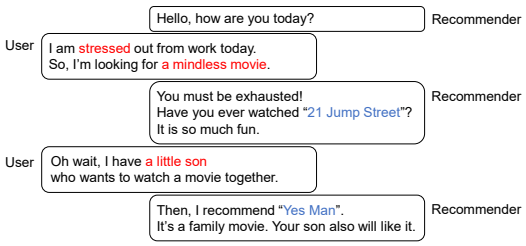

Figure 1: An example of a user-recommender conversation for movie recommendations. The important information for a recommendation is written in red, and the recommended items are written in blue.

recommender systems by capturing users' interests through natural conversation. The conventional CRSs (Chen et al., 2019; Ma et al., 2021; Zhou et al., 2020a) mainly derive a user preference from item-entities, which appear in a conversation and exist in a knowledge graph (KG). However, users often express their preferences without preferred items. For example, in Figure 1, the user depicts his/her requirement (*a mindless movie*), feeling (*stressed*), and situation (*with a little son*). In order for an accurate recommendation of a movie (*Yes Man*) to be derived, the recommender must capture that the user wants those characteristics of a movie, "*a movie to watch when stressed*", "*a mindless movie*" and "*a movie to watch with a little son*". Therefore, merely depending on item-entities for user preference identification is insufficient. We argue that a dialogue context also provides a user preference, regardless of the existence of item-entities.

Motivated by the characteristic of an online recommendation community where participants identify a context of a recommendation request and then comment with appropriate items, our method exploits the Reddit data to extract the contextual knowledge that enables inferring a preferred item in a KG from a recommendation request. Thus, we propose a Contrastive Learning approach for Injecting Contextual Knowledge (**CLICK**), which

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1875–1885
May 2-6, 2023 ©2023 Association for Computational Linguistics

captures not only the entity-level preference derived from mentioned entities but also identifies the context-level preference from a dialogue context by utilizing both a KG and Reddit data.

However, the different modality between recommendable item-entities in a KG and request texts in the Reddit data hinders CLICK from learning contextual knowledge. Thus, we alleviate the different modalities via contrastive learning to align the pairs of a recommended item-entity and a request text. Moreover, different from the conventional contrastive learning that classifies the pairs into only positive or negative, our distinction is to consider the relative relevance among positive pairs. For example, if a seeker asks a recommendation community for a movie to watch when feeling down, several movies can be recommended. While some movies are related to many participants, others may be referenced by a few people. Thus, to reflect differing relativity of multiple recommendable items from a request, we design a relevance-enhanced contrastive learning loss function. Consequently, leveraging the learned contextual knowledge, CLICK makes accurate recommendations based on the user preferences extracted from both entity- and context-level perspectives. In the response generation task, we develop a response generator that produces a suitable explanation with a recommended item based on a captured requirement from a dialogue context.

Our contributions are summarized as follows: (1) We propose the knowledge injection method that facilitates the inference of a recommendable item based on a dialogue context in order to identify a comprehensive user preference. (2) We design the relevance-enhanced contrastive learning loss that promotes the fine-grained reflection of positive pairs according to the relevance. (3) We further develop a contextual knowledge-enhanced recommender module and a response generation module, which captures a comprehensive user preference and produces an explanation based on a user's needs, respectively. (4) The proposed model outperforms baselines, especially regardless of the existence of item-entities in a user utterance.

2 Related Work

2.1 Conversational Recommender Systems

CRS (Gao et al., 2021) allows obtaining user preferences through dynamic interaction, overcoming the limitations of traditional recommender systems (Guo et al., 2020) that are heavily dependent on static interaction history. The research area can be broadly categorized into two groups: template-based and natural language-based CRSs. Template-based methods (Deng et al., 2021; Lei et al., 2020; Zhou et al., 2020d) interact with a user, following a slot-filling approach. Although many template-based CRSs have been proposed, they suffer from producing inflexible response due to the inherent template scheme.

On the other hand, natural language-based CRS (Chen et al., 2021; Zhou et al., 2020c) allows users to depict their needs in a free text. Such approaches focus on how to extract knowledge from external data (e.g., historical data, KGs, and review data) and how to utilize the knowledge to capture a user preference and generate a persuasive response since the CRS benchmark dataset contains repetitive and limited utterances. Following the stream of CRS research, Zhou et al. (2020b) utilize historical interaction data to enhance a sequence of preferred items. Chen et al. (2019) and Zhou et al. (2020a) exploit KGs to capture user preferences with mentioned entities. Ma et al. (2021), Moon et al. (2019), and Zhou et al. (2021) leverage explicit reasoning on a KG based on the mentioned entities. Lu et al. (2021) and Zhou et al. (2022) utilize review data to enrich insufficient item information over a KG. However, despite the diverse use of external data, the current methods still lack capturing a user's needs depicted in a free text since they identify preferences mainly with mentioned item-entities that exist in a KG. To resolve the limitation, we propose a contrastive learning approach for injecting knowledge into the CRS task by utilizing textual recommendation data (Reddit) and a KG. Thus, our approach captures user preferences from both mentioned entities and a dialogue context, taking full advantage of CRS in which users can express their preferences freely in natural language.

2.2 Contrastive Learning

Another line that motivates our work is contrastive learning, which is a dominant approach in self-supervised learning. Especially in multi-modal scenarios, contrastive learning (Oord et al., 2018) plays a vital role in refining feature representation. It aims at forcing representations to have a smaller distance between positive pairs and a greater distance between negative pairs than the positive ones. Due to the efficacy of contrastive learning strategy,

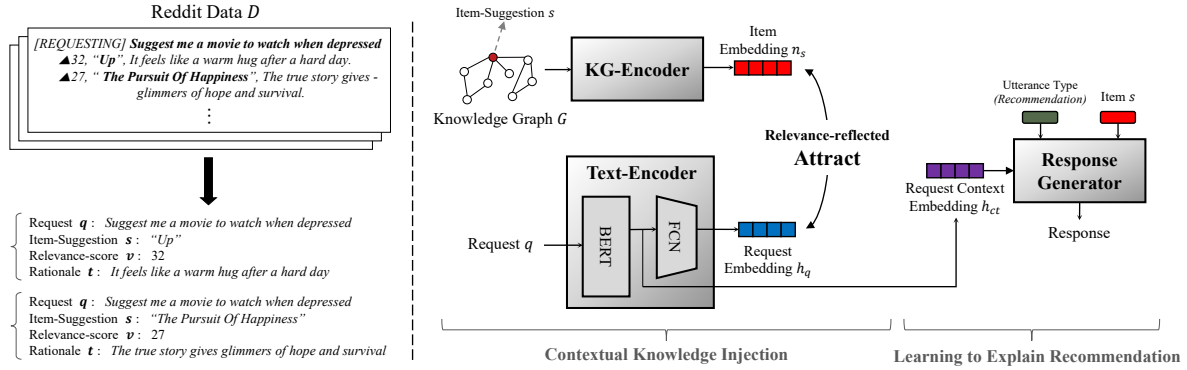

Figure 2: The pre-training stage of CLICK which consists of two encoders and a response generator.

it has been widely used in various research fields such as multilingual text-to-video search (Huang et al., 2021), visual pre-training (Yuan et al., 2021), and other domains (Xia et al., 2021; Zolfaghari et al., 2021). Especially, in recent research on CRS, Zhou et al. (2022) adopt contrastive learning to fuse multi-type external data. Likewise, we refine our multi-type external data representations by contrastive learning. However, different from the existing contrastive learning loss, we further devise relevance-enhanced contrastive learning loss to consider multiple recommendable items, of which each has a different relative relevance.

3 Pre-training for Contextual Knowledge Injection

We introduce the training scheme of CLICK that follows a two-stage mechanism: 1) a pre-training stage for the contextual knowledge injection and 2) a fine-tuning stage for the conversational recommendation task.

In the two stages, we utilize DBpedia (Bizer et al., 2009) as a KG, following the existing CRSs (Chen et al., 2019; Zhou et al., 2020a; Ma et al., 2021; Zhou et al., 2022) that take advantage of KGs to capture user preferences based on factual information such as objects, concepts, and relationships among them. The KG is defined as $G = \{(e1, r, e2)|e1, e2 \in E, r \in R\}$, where a fact $(e1, r, e2)$ consists of relation r from entity $e1$ to entity $e2$ from entity set E and relation set R.

The Reddit data (Penha and Hauff, 2020) is leveraged to learn the contextual knowledge that facilitates the inference of a preferred item from a user utterance. The Reddit data is denoted as $D = \{(q, s, v, t)\}$, where q is a seeker's request text to receive a recommendation, s is an item-suggestion by a recommender, v is a relevance score of the item-suggestion, and t is an optional rationale for the item-suggestion. We describe the details of the Reddit dataset in Section 5.1.

In the pre-training stage, our approach infuses the contextual knowledge from Reddit data into a KG-encoder and a text-encoder by training them via a devised contrastive learning loss. Then, a response generator is pre-trained to equip the ability to generate an explanation for a recommendation based on the recognized needs. The pre-training stage of CLICK is described in Figure 2.

3.1 Contextual Knowledge Injection

Following previous studies (Zhou et al., 2020a; Ma et al., 2021), we encode the KG to obtain the representations of entities via R-GCN (Schlichtkrull et al., 2018) that considers neighborhoods under the specific relation type and direction. The representation of item-entity s from the KG-encoder is denoted as n_s. Then, we employ BERT (Devlin et al., 2019) and a fully-connected layer FCN to encode a request text q that contains the seeker's needs (more details in Appendix A.1). The request embedding h_q is obtained as:

$$h_q = FCN(BERT(q)). \quad (1)$$

Next, the above two encoders are pre-trained via a contrastive learning loss, which promotes alignment of the multi-modality originating from the KG and text. That way, the text encoder is encouraged to effectively infer a preferred item from a request. Moreover, different from a conventional contrastive learning loss that classifies the pairs into only positive or negative, we build the relevance-enhanced contrastive learning to reflect multiple recommendable items with a relative relevance as

follows:

$$L_{rel} = -log \sum_{s \in S} \frac{ex^{(sim_{q,s} - v_{q,s})}}{ex^{(sim_{q,s} - v_{q,s})} + \sum_{S'} ex^{(sim_{q,s'})}}, \quad (2)$$

where $v_{q,s}$ indicates the relevance score corresponding to the number of "Upvote" about the item suggestion s from the request q in positive set S, S' is a negative set, sim means cosine similarity between the item representation vector n_s and the request representation vector h_q that the above two encoders produce respectively, and $ex^{(x)}$ is exponential of x. The proposed loss function is able to attract positive pairs in proportion to the relevance score $v_{q,s}$ while repelling the negative pairs. Then the pre-trained encoders are utilized to capture a comprehensive user preference from a user utterance depicted regardless of the presence of entities.

3.2 Learning to Explain Recommendation

At this stage, the response generator is pre-trained to generate a recommendation response of the three utterance types (*question, recommendation, chit-chat*). The recommendation response should contain a recommended item and a rationale for the recommendation based on a captured user need. Thus, we adopt GPT-2 (Radford et al., 2019) as a response generator and feed a suggested item s and the utterance type (*recommendation*) to a token embedding layer of GPT-2. The first layer of the decoder A^1 then conducts self-attention with the token embeddings. Next, having obtained the request context embedding h_{ct} from the text-encoder (only $BERT$), cross-attention is conducted in the next decoder block to generate the explanation for the item-suggestion based on the needs extracted from the request of a seeker. The cross-attention layer of the decoder can be summarized as:

$$A^2 = FFN(MultiHead(O_{A^1}, h_{ct}, h_{ct})), \quad (3)$$

where the multi-head attention layer $MultiHead(Q, K, V)$ takes query Q, key K and value matrix V as input, and O_{A^1} is the output of the first decoder layer A^1. The response generator is pre-trained to generate the rationale t for the item-suggestion s conditioned on the request q and the utterance type ut via maximum likelihood estimator (MLE) generation loss:

$$L_{exr} = -\sum_{j=0}^{J} \log p(t_j | t_{<j}, q, s, ut). \quad (4)$$

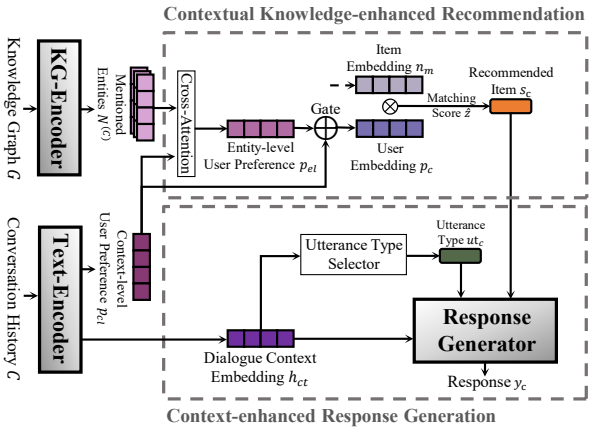

Figure 3: The fine-tuning stage of CLICK utilizing the pre-trained encoders and response generator for the conversational recommendation task.

4 Fine-tuning for Conversational Recommendation Task

In this section, the fine-tuning strategy of CLICK for the CRS task is introduced. Figure 3 illustrates the working flow of CLICK, leveraging the two pre-trained encoders and response generator from the pre-training stage. CLICK consists of two key components: (1) contextual knowledge-enhanced recommender module, which even understands user's needs from a dialogue context, and (2) context-enhanced response generation module, which generates a persuasive response based on a captured user preference. Prior to introducing the details of the fine-tuning scheme, we formulate a conversation history $C = \{x_1, y_1, ..., x_c\}$, where x_i is a user utterance, and y_i is a system utterance at turn-i. The goal of CLICK is to recommend a proper item s_c and generate the system response y_c, understanding the conversation history C.

4.1 Contextual Knowledge-enhanced Recommendation

The pre-trained text-encoder is leveraged to identify context-level user preference (e.g., *a movie to watch when feeling down*) that could not be captured with mentioned entities. In detail, the conversation history C is fed to the text-encoder, and we obtain the context-level user preference as $p_{cl} = FCN(BERT(C))$.

Next, we utilize mentioned entities to identify entity-level user preference (e.g., *a movie like Yes-Man*). In our implementation, we conduct cross-attention with a context-level user preference and entities, being different from the previous ap-

proaches that ignore the relative importance among mentioned entities based on a dialogue context. Thus, the cross-attention score is denoted as:

$$\alpha = softmax(q \cdot k^{\mathsf{T}}/\sqrt{d_k}), \\ q = p_{cl} \cdot W_q,\ k = n_e \cdot W_k, \tag{5}$$

where W_q,W_k are weight metrices, n_e is an entity embedding from the mentioned entities $N^{(C)}$, and the context-level user preference p_{cl} is from the text-encoder. Then, we obtain the entity-level user preference p_{el} as follows:

$$p_{el} = \alpha \cdot N^{(C)}, \tag{6}$$

Then, we adopt a gate mechanism to obtain the final user preference p_c at c-th turn by combining the entity-level and context-level user preferences as follows:

$$p_c = \beta \cdot p_{el} + (1-\beta) \cdot p_{cl}, \\ \beta = \sigma(W_{gate} \cdot [p_{el}; p_{cl}]), \tag{7}$$

where W_{gate} is a weight matrix, and $[;]$ indicates the concatenation operator. Consequently, we conduct the inner product of user preference p_c and the representation n_m of item m to predict the matching score:

$$\hat{z}(c,m) = p_c \cdot n_m^{\top}. \tag{8}$$

The recommender module is fine-tuned with a cross-entropy loss as:

$$L_{rec} = -\sum_{c=1}^{C}\sum_{m=1}^{M}[z_{cm} \cdot \log \hat{z}(c,m) \\ - (1-z_{cm}) \cdot \log(1-\hat{z}(c,m))], \tag{9}$$

where z_{cm} is the true label.

4.2 Context-enhanced Response Generation

Prior to generating a response, at each turn, CLICK determines the type of an utterance (*question, recommendation, chit-chat*) given a dialogue context. Following Ma et al. (2021), we regard the determination of utterance's type as a 3-way classification problem:

$$Pr(y_c^{ut}|C) = softmax(W_{ut}^2 \cdot ReLU(W_{ut}^1 \cdot h_{ct})), \tag{10}$$

where h_{ct} is a dialogue context embedding from the BERT. We thus formulate the determination loss as:

$$L_{ut} = -\log Pr(y_c^{ut}|x_1, y_1, ..., x_c), \tag{11}$$

where y_c^{ut} is the true label of utterance type.

After obtaining the dialogue context embedding, the utterance type, and the recommended item, we utilize such to generate a persuasive response that contains a recommended item under the controlled type of an utterance. Compared to an implicit way of response generation in CRSs (Chen et al., 2019; Zhou et al., 2021), which lacks the ability to include a recommended item in the generated response, the explicit way of response generation by Ma et al. (2021) feeds the recommended item and the utterance type as input into the response generator, which guarantees the inclusion of the item under the controlled response type. However, the explicit approach hinders generating a response based on a dialogue context. Thus, different from Ma et al. (2021), we further infuse the dialogue context implicitly into the response generator with the cross-attention mechanism to generate a response containing an identified dialogue context.

In the same manner as the pre-training process of the response generator, an utterance type and a recommended item is fed to the token embedding layer. The first layer of the decoder conducts self-attention with the token embeddings. Then, we enforce the module to generate a response based on the user's needs via the cross-attention mechanism with the dialogue context embedding h_{ct}, which is identical to the one in Equation 3. We formulate the response generation loss as:

$$L_{gen} = -\sum_{u=0}^{U}\log p(y_{u,c}|y_{<u,c}, C, s_c, ut_c), \tag{12}$$

where U is the length of a target response, $y_{u,c}$ is a generated u-th token of system response, C is a conversation history, and s_c is the recommended item at c-th turn. ut_c indicates the type of utterance. Additionally, the implementation details are in Appendix A.2.

5 Experimental Setup

5.1 Dataset

REDIAL is a benchmark CRS dataset released by Li et al. (2018). It was constructed through Amazon Mechanical Turk (AMT). The AMT workers play the role of a seeker/recommender. In daily talk, a seeker explains his/her tastes about movie and asks for movie suggestions. A recommender tries to understand the tastes and recommends movies. Accordingly, it consists of 10,006 conversations

Model	Standard Setting			No Mentioned Item Setting			Diversity
	R@1	R@10	R@50	R@1	R@10	R@50	
REDIAL	2.3	13.2	29.7	–	–	–	5.8
KBRD	3.0	15.8	33.8	1.5	5.1	12.2	11.2
KGSF	3.9	18.3	37.8	2.6	9.7	20.1	12.2
CR-Walker	4.0	18.7	37.6	2.4	10.2	19.8	14.9
C2CRS	5.2	22.9	40.7	2.1	8.9	18.3	13.9
GPT-2	2.3	14.5	31.5	1.6	7.9	16.3	9.9
GPT-2 (+*Reddit*)	2.5	15.2	32.1	1.9	8.5	17.1	11.1
BERT	3.0	15.3	34.4	1.9	8.7	17.5	10.3
BERT (+*Reddit*)	3.2	16.9	35.2	2.1	9.2	18.3	11.8
BART	3.3	17.1	36.3	2.1	9.0	18.0	11.4
BART (+*Reddit*)	3.6	18.3	37.5	2.3	9.5	18.9	12.7
CLICK	**5.5**	**23.8**	**43.2**	**3.5**	**14.9**	**28.9**	**18.3**
w/o relevance	5.3	23.1	41.1	3.3	13.2	26.7	16.9
w/o pre-training	4.2	19.2	38.1	2.5	10.0	20.3	14.7

Table 1: Overall performance comparison on the recommendation task.

containing 182,150 utterances and 51,699 movies. We set the training, validation, and test sets as a ratio of 8:1:1. **Reddit** is utilized as the textual recommendation data collected by Penha and Hauff (2020). Following the domain of a benchmark CRS dataset (REDIAL), we use the MovieSuggestions subreddit, which is a movie recommendation community. In the community, a seeker asks for a recommendation, and recommenders give suggestions with a rationale. Participants in the community consent to the suggested items with the "Upvote" button which is regarded as a relevance score. **Knowledge Graph** consists of 30,471 entities, 12 relations, and 392,682 triples extracted from DBpedia (Bizer et al., 2009) and MovieLens, following Ma et al. (2021). Additionally, the entities in utterances are linked to DBpedia nodes.

5.2 Baselines

We compared CLICK to the following five CRS baselines: (1) **REDIAL** (Li et al., 2018) is proposed with the CRS benchmark dataset. The model consists of an auto-encoder recommender and an RNN-based response generator. (2) **KBRD** (Chen et al., 2019) utilizes a KG to identify user preferences from mentioned entities. The transformer-based response generator uses a user representation as vocabulary bias. (3) **KGSF** (Zhou et al., 2020a) incorporates word-oriented and entity-oriented KGs to capture user preference. (4) **CR-Walker** (Ma et al., 2021) proposes tree-structured reasoning on a KG, which effectively utilizes background knowledge. It adopts GPT-2 to generate a response by feeding dialogue acts. (5) **C2CRS** (Zhou et al., 2022) leverages a KG and a review data for enriching the context information. It conducts data semantic fusion via contrastive learning. The response generator fuses the information from a KG and a review data.

Additionally, three pre-trained language models (PLMs) were tested for a fair comparison on the same data. In table 1 and 2, (+*Reddit*) indicates a PLM pre-trained on the Reddit data. The three models are as follows: GPT-2(Radford et al., 2019), BERT(Devlin et al., 2019), and BART(Lewis et al., 2020). The training details of PLMs are in Appendix A.3.

6 Experimental Results

6.1 Evaluation on Recommendation

We adopted Recall@k to measure how many predicted items were included in the set of preferred items (Standard Setting). Moreover, we compared the performance under the setting where there were no mentioned item-entities in a dialogue context (No Mentioned Item Setting). This setting aims to evaluate how well the models identify a user preference from a dialogue context without mentioned item-entities. We also evaluated recommendation diversity (Diversity), defined as the proportion of distinct recommendations.

6.1.1 Performance Comparison

Table 1 shows the experimental results on recommendation task. **Standard Setting**: KBRD, KGSF, and CR-Walker performed better compared to REDIAL by capturing user preference based on entities in KGs. Compared to the baselines, C2CRS

Model	Automatic Evlauation				Human Evaluation		
	BLEU	Dist-2	Dist-3	Dist-4	Flu.	Rel.	Info.
REDIAL	0.229	0.215	0.247	0.234	2.03	1.97	1.67
KBRD	0.231	0.281	0.376	0.438	2.18	2.19	1.98
KGSF	0.239	0.320	0.432	0.519	2.31	2.21	2.12
CR-Walker	0.271	0.361	0.493	0.573	2.54	2.37	2.31
C2CRS	0.251	0.339	0.455	0.538	2.42	2.29	2.19
GPT-2	0.247	0.352	0.474	0.521	2.35	2.01	2.04
GPT-2 (+*Reddit*)	0.261	0.359	0.481	0.536	2.39	2.09	2.02
BART	0.249	0.354	0.486	0.528	2.37	2.07	2.12
BART (+*Reddit*)	0.262	0.357	0.492	0.539	2.38	2.17	2.11
CLICK (ours)	**0.308**	**0.380**	**0.521**	**0.598**	**2.60**	**2.63**	**2.58**
w/o gen. cross-att	0.296	0.371	0.510	0.589	2.58	2.51	2.47
w/o gen. pre-training	0.289	0.364	0.503	0.583	2.53	2.45	2.44
w/o pre-training	0.281	0.359	0.497	0.579	2.48	2.39	2.40

Table 2: Overall performance comparison on the response generation task

recommended appropriate items by incorporating a KG and a review data. In contrast, CLICK achieved the best results by effectively modeling comprehensive user preference from entity-level and context-level. Another observation is from the comparison with PLMs. Although the PLMs showed impressive performances, the improvement by utilizing the Reddit data was lower than ours. Thus, it proved that CLICK efficiently extracted contextual knowledge from Reddit data and exploited it for recommendation. **No Mentioned Item Setting**: To further demonstrate the CLICK's ability of capturing a context-level user preference, we compared the baselines on this setting. KGSF and CR-Walker performed relatively better than other baselines due to the utilized word-oriented KG allowing the recognition of diverse words in a conversation. We noted that CLICK showed best performances even when there were no mentioned item-entities. This is because the contextual knowledge extracted from the Reddit data provided a significant clue to a user's needs. More details are covered in Appendix A.4. **Diversity**: CLICK recommended diverse items compared to all the baselines. The improvement was derived from our strategy of modeling user preference in a fine-grained means, considering both mentioned entities and a dialogue context.

6.1.2 Ablation Study

Pre-training with the relevance-enhanced contrastive learning loss is an essential design in our method to infer a preferable item from a user utterance. Thus, we assessed the effects of CLICK with two variants, (1) **w/o relevance**: standard contrastive learning loss (Oord et al., 2018) was used, (2) **w/o pre-training**: CLICK was only trained on REDIAL dataset without being pre-trained with the Reddit data. As shown in Table 1, the performance of CLICK *w/o relevance* was lower since it did not consider the relative relevance that helped the model to learn a fine-grained representation. CLICK *w/o pre-training* provided the evidence for the pre-training stage of learning the contextual knowledge. Besides, it showed a decline in diversity since it did not capture a fine-grained user preference with the contextual knowledge that promotes various recommendations.

6.2 Evaluation on Response Generation

BLEU (Papineni et al., 2002) and Dist-N (Li et al., 2016) were adopted to measure the word-level correspondence and diversity of responses. We further conducted a human evaluation. Five human annotators evaluated 100 generated responses on fluency, relevancy, and informativeness ranging from 0 to 3. The average score of each metric is reported.

6.2.1 Performance Comparison

Table 2 shows the experimental results on response generation task. **Automatic Evaluation**: The result proved that our method generated a better quality response with diverse words. Such improvements are derived from the explicit feeding of a recommended item and an utterance type into a response generator and the implicit injection of a context via a cross-attention mechanism. **Human Evaluation**: The responses generated from C2CRS, KGSF, and KBRD are inclined to contain safe responses that are broadly adequate for any situation, such as *"I have not seen that one"*.

This issue was reflected as the lower score on relevancy and informativeness compared to CR-Walker. While CR-Walker generated a meaningful response with an explanation for a recommendation, CLICK achieved the best performance on all human evaluations. Especially, the better relevancy and informativeness demonstrated the efficacy of our response generator incorporating a dialogue context via the cross-attention.

6.2.2 Ablation Study

The ablation study wass conducted with the following three variants of CLICK, (1) **w/o gen. cross-att**: we removed the cross-attention mechanism incorporating a dialogue context. (2) **w/o gen. pre-training**: we only pre-trained the recommender module except for the response module, (3) **w/o pre-training**: we trained the CLICK on REDIAL dataset without pre-training. As shown in Table 2, we observed a significant decrease when the proposed strategies were removed. It proved that these strategies contributed to improving the performance of response generation. We noted that the cross-attention mechanism and the pre-training scheme of the response generator affected generating a relevant and informative response. Moreover, the *w/o pre-training* case brought a further decline in all metrics, indicating the usefulness of the pre-training strategy or generating informative responses using diverse words.

6.3 Case Study

A qualitative analysis of three cases is shown in Table 3. **Case 1**, there are no mentioned item-entities. Given the user utterance, KGSF recommended "*Black Panther*" that is suggested frequently in the CRS dataset regardless of a specific situation. Thus, it did not seem to capture the user's needs (*comforting movie*) precisely. Likewise, CR-Walker asked a question since it did not identify the needs exactly. In contrast, our model recommended a proper item with a rationale based on the captured needs. **Case 2** is a common case in the CRS dataset, which explicitly expresses the user's taste. Therefore, all baselines and CLICK generated a proper recommendation. Especially, CLICK provided a suitable explanation, understanding that the user wants comedy. **Case 3** includes multiple entities and a specific situation where the recommended movie should be enjoyable for little girls. We observed that all baselines suggested action movies, concentrating on the genre prevailing in the dialogue context. Here, the recommender should understand that the user is looking for a kids' movie, such as "*The Princess and The Frog*" besides action movies. In constrast, our model grasped the relatively important entity among mentioned entities based on the dialogue context. Consequently, our model's recommendation and rationale are well-grounded on users' needs.

[Case 1] **Recommender:** hello how are you today? **User:** I'm feeling pretty down tonight so I'd like to watch a comforting movie.
KGSF: Have you ever seen "Balck Panther"? **CR-Walker**: What type of movie do you like? **Ours**: What about "Mamma Mia!"? It's a comforting movie. **Human**: I recommend "The Secret Life of Walter Mitty", It's so satisfying.
[Case2] **Recommender:** hello what kind of movie do you like? **User:** I'm in the mood for a good comedy like "Get Smart"
KGSF: Have you ever seen "Jumanji"? **CR-Walker**: "Happy Gilmore" is a really good comedy. It's my favorite adam sandler movie. **Ours**: Have you ever seen "Men in Black"? It has funny story. **Human**: You should watch "Spy". I liked the humor of the actress.
[Case3] **Recommender:** Hello! Could you let me know any movies you liked? **User:** I like action movies like "John Wick " or "Die Hard 4.0", but today I'm looking for something similar to "The Princess and the Frog". I have little girls who want to watch a movie.
KGSF: Have you seen "Black Panther"? **CR-Walker**: I recommend "Taken". It's a really good movie with lots of action! **Ours**: What about "Tangled". It's a fairytale movie. Little girls will love it. **Human**: Then, I recommend "Frozen". Becasue littile girls like it.

Table 3: A qualitative analysis of three cases. The baselines, CLICK(ours), and human generate a response based on the given dialogue context. The important information for a recommendation is written in red, and the mentioned item-entities in a KG are written in blue.

7 Conclusions and Future Work

In this paper, we proposed the contextual knowledge injection via contrastive learning to capture a comprehensive user preference from a dialogue context. Utilizing the extracted contextual knowledge, CLICK not only precisely captures the preference on mentioned entities but also identifies the user's needs in a dialogue context. Moreover, the relevance-enhanced contrastive learning loss enables the contextual knowledge to reflect the acceptance degree of recommendations. We further developed the context-enhanced response generator to provide a persuasive explanation. Lastly, experimental results on the benchmark dataset confirmed the effectiveness of CLICK on providing appropriate recommendations, as well as generating high-quality responses.

In future work, we will explore a CRS with a recommendation dialogue strategy of actively asking questions to elicit a user taste instead of passively communicating with a user.

Limitations

In this section, we summarized the limitations of our study as follows:

- The training and test experiments are conducted on the setting where the mentioned entities are aligned to a knowledge graph in the same manner as other CRSs. In a practical use, Named Entity Recognition (NER) task is needed, and entity linking is also required to align the entities to a KG.
- This work only conducts experiments on the movie domain and does not generalize to other domains since the benchmark CRS dataset is limited to the movie domain. Benchmark CRS datasets on various domains need to be published, and experiments on multiple domains are required to support a complete and objective study.

Ethics Statement

All authors of this paper acknowledged ACL Ethics Policy. We utilized public datasets, REDIAL dataset and Reddit dataset. REDIAL dataset was constructed by crowdsourcing with Amazon Mechanical Turk, which is conducted anonymously. And Reddit dataset was crawled from Reddit community site without personal information. Although we have not observed an inappropriate content in the generated response, we inform that pretrained GPT-2 is leveraged, which may produce an inappropriate response.

Acknowledgements

We appreciate all of the reviewers giving their sincere comments. And we would like to express our gratitude to Donghyun Kim and Wongyu Kim for their valuable discussion and motivation. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022R1A2B5B01001835). Kyong-Ho Lee is the corresponding author.

References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 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)*, pages 1803–1813.

Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2021. Towards explainable conversational recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2994–3000.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. *Unified Conversational Recommendation Policy Learning via Graph-Based Reinforcement Learning*, page 1431–1441.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander G Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *NAACL*, pages 2443–2459.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2073–2083.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. Revcore: Review-augmented conversational recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1161–1173.

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 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*, pages 1839–1851.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Fourteenth ACM Conference on Recommender Systems*, pages 388–397.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks.

Xin Xia, Hongzhi Yin, Junliang Yu, Qinyong Wang, Lizhen Cui, and Xiangliang Zhang. 2021. Self-supervised hypergraph convolutional networks for session-based recommendation. 35(5):4503–4511.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Jinfeng Zhou, Bo Wang, Ruifang He, and Yuexian Hou. 2021. Crfr: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4324–4334.

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

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

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020c. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139.

Sijin Zhou, Xinyi Dai, Haokun Chen, Weinan Zhang, Kan Ren, Ruiming Tang, Xiuqiang He, and Yong Yu. 2020d. Interactive recommender system via knowledge graph-enhanced reinforcement learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 179–188.

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 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*, pages 1488–1496.

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459.

A Appendix

A.1 The Pipeline between BERT and FCN in Text-Encdoer

The output of BERT is an embedding vector of a CLS token containing the meaning of the entire sentence. And the embedding vector is input into FCN.

A.2 Implementation Details

We implemented CLICK with Pytorch[1] and trained it on an NVIDIA RTX 3090. We set the embedding size of the KG-encoder (R-GCN)[2] to 128, and the depth of aggregating neighbors as 1. We used pre-trained BERT[3] and GPT-2[4] to intialize the models' parameters. The hidden size of BERT in the text-encoder is set to 768, and the hidden size of the response generation module (GPT-2) is identically set to 768. When obtaining the embedding of a context-level user preference from the text-encoder, we utilized a 128×768 fully-connected layer. We normalized the relevance scores from 0.25 to 0.5 for a stable convergence. We used Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-4 and a weight decay of 1e-3. The batch size of pre-training and fine-tuning is set to 32. The computational cost is 3 hours and 1 hour per 1 epoch for pre-training and fine-tuning, respectively.

A.3 PLMs Details

The PLMs with (*+Reddit*) are pre-trained on Reddit dataset and fine-tuned on REDIAL dataset. And the representation of the last token that a PLM generates is used for recommendation.

A.4 No Mentioned Item Setting

We utilized a special entity (*None entity*) in the same manner as other CRSs (KGSF, CR-Walker, and C2CRS) that utilize KGs. Thus, CLICK and other CRSs can be executed under any case. For example, at the first turn of the conversation in Figure 1, there is no mentioned entity. Thus, *None entity* is assigned as mentioned entities $N^{(C)}$ in Equation 6. Meanwhile, the text-encoder of CLICK is pre-trained with Reddit data to infer recommendable items (*e.g., 21JumpStreet, Mad-Max:FuryRoad, etc.*) from the request context where a user gets stressed or requires mindless movies. The gate mechanism in Equation 7 then emphasizes a context-level user preference p_{cl} obtained from the pre-trained text-encoder than an entity-level user preference p_{el} in the case of no mentioned entities. Thus, the user embedding concentrated on the context-level user preference p_{cl} enables CLICK to recommend *21JumpStreet* from the user's first utterance in Figure 1.

[1] https://pytorch.org/
[2] https://github.com/pyg-team/pytorch_geometric
[3] https://huggingface.co/bert-base-uncased
[4] https://huggingface.co/gpt2