Toward Knowledge-Enriched Conversational Recommendation Systems

Tong Zhang[1], Yong Liu[2, 3], Boyang Li[1, 2], Peixiang Zhong[2, 3],
Chen Zhang[4], Hao Wang[4] and Chunyan Miao[1, 2, 3]

[1]School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
[2]Alibaba-NTU Singapore Joint Research Institute, NTU, Singapore
[3]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, NTU, Singapore
[4]Alibaba Group, China
{tong24, stephenliu, boyang.li, peixiang001, ascymiao}@ntu.edu.sg,
zhangchen010295@163.com, cashenry@126.com

Abstract

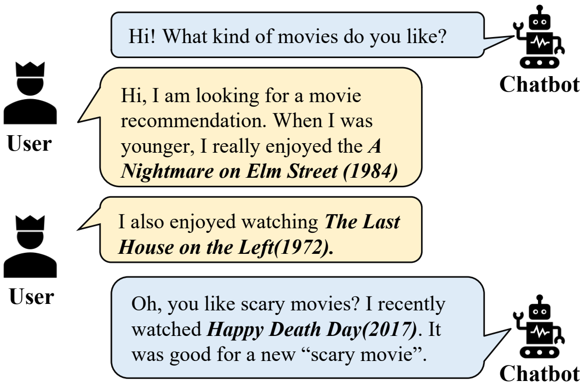
Conversational Recommendation Systems recommend items through language based interactions with users. In order to generate naturalistic conversations and effectively utilize knowledge graphs (KGs) containing background information, we propose a novel Bag-of-Entities loss, which encourages the generated utterances to mention concepts related to the item being recommended, such as the genre or director of a movie. We also propose an alignment loss to further integrate KG entities into the response generation network. Experiments on the large-scale REDIAL dataset demonstrate that the proposed system consistently outperforms state-of-the-art baselines.

1 Introduction

Conversational recommendation systems (CRS) have received increasing attention from the Natural Language Processing community (Li et al., 2018; Chen et al., 2019; Zhou et al., 2020a; Sarkar et al., 2020; Liu et al., 2020; Zhou et al., 2020b; Hayati et al., 2020). CRS aims to recommend items, such as movies or songs, in naturalistic interactive conversations with the user. This interactive form allows the system to provide recommendations tailored to preferences provided by the user at the moment.

A crucial issue of CRS is to extract user preferences from the conversation, which often requires background information provided by knowledge graphs (KGs). As an example, in Figure 1, the user mentions two movies that belong to the horror genre. To this end, some existing studies (Chen et al., 2019; Zhou et al., 2020a) leverage knowledge graphs to understand user intentions.

We observe that when humans recommend items to friends, they usually describe attributes of the item. For example, to recommend a movie, they may mention the director or actors. Such information can be easily extracted from the knowledge graph, but has not been well utilized by existing approaches. To emulate naturalistic conversations, we propose a Bag-of-Entities (BOE) loss, which encourages the generated utterances to mention concepts related to the item. Moreover, we propose an alignment loss that ties the word embeddings to the entity embeddings.



Hi! What kind of movies do you like?

Hi, I am looking for a movie recommendation. When I was younger, I really enjoyed the ***A Nightmare on Elm Street (1984)***

I also enjoyed watching ***The Last House on the Left(1972).***

Oh, you like scary movies? I recently watched ***Happy Death Day(2017)***. It was good for a new "scary movie".

Figure 1: An example of a conversation between a user and the Chatbot for movie recommendation.

Experiments demonstrate that the proposed two losses improve model performance. The proposed the Knowledge-Enriched Conversational Recommendation System (KECRS) consistently outperforms state-of-the-art CRSs on the large-scale REDIAL dataset (Li et al., 2018).

2 Related work

We briefly review work on conversational recommendation systems and conversational characters in e-commerce settings. A number of works on conversational recommendation systems focus solely on interactive recommendation rather than language understanding (Christakopoulou et al., 2016, 2018; Sun and Zhang, 2018; Zhang et al., 2018; Lei et al., 2020a,b; Zou et al., 2020; Xu et al., 2021; Zhang et al., 2022). In contrast, a second category of works aims to provide both accurate interactive recommendations and generate natural

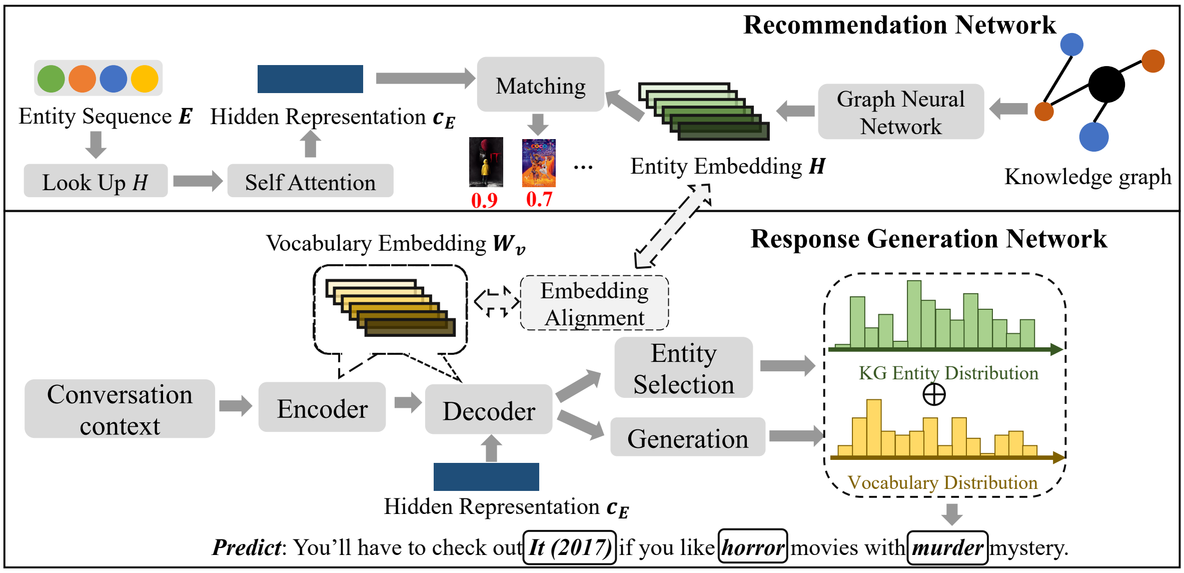


Figure 2: The overall framework of the proposed KECRS model.

and human-like responses (Li et al., 2018; Chen et al., 2019; Zhou et al., 2020a; Sarkar et al., 2020; Liu et al., 2020; Zhou et al., 2020b; Hayati et al., 2020). Finally, research on conversational characters for e-commerce has the broad goal of building a complete shopping assistant that can answer a variety of questions in addition to recommendation (Li et al., 2017; Yang et al., 2018; Fu et al., 2020; Song et al., 2021).

3 Approach

The overall goal of a conversational recommendation system is to identify an item (e.g., a movie, a song, or a piece of merchandise) that the user will likely interact with and suggest the item to the user in the form of natural language conversations.

Formally, we represent the historic conversation $X = \langle x_1, x_2, ..., x_n \rangle$ as a sequence of n utterances x_i. The knowledge graph $\mathcal{G} = \{(v_h, r, v_t)\}$ is a set of entities E and a set of relationships r between the head entity $v_h \in E$ and the tail entity $v_t \in E$.

The conversational recommendation task is to predict the next utterance x_{n+1} using the recommendation network $f(X, \mathcal{G})$ and the response generation network $g(X, \mathcal{G}, f(X, \mathcal{G}))$. $f(X, \mathcal{G})$ predicts the next item to recommend to the user, whereas $g(X, \mathcal{G}, f(X, \mathcal{G}))$ predicts the utterance x_{n+1} one word at a time.

Figure 2 shows the overall structure of our proposed method, the Knowledge-Enriched Conversational Recommendation System (KECRS).

3.1 Recommendation Network

First, we exhaustively match each word in the conversational history X with the name of each entity in the KG. In this way, we identify K entities from the history and sequence them according to their original positions. Next, we apply a graph convolutional network, R-GCN (Schlichtkrull et al., 2017) to encode the entire KG and obtain embeddings for each KG entity node. The D-dimensional entity embeddings of the K entity form the matrix $\boldsymbol{H}_E \in \mathbb{R}^{K \times D}$. Subsequently, we apply an attention operation where the attention vector $\boldsymbol{\alpha}$ is computed by 2 fully connected (FC) layers.

$$\begin{aligned} \boldsymbol{\alpha} &= \mathrm{softmax}\big(\mathbf{W}_k \mathrm{tanh}(\mathbf{W}_q \mathbf{H}_E^\top)\big), \\ \mathbf{c}_E &= \boldsymbol{\alpha} \mathbf{H}_E, \end{aligned} \tag{1}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable parameters. The resulting $\mathbf{c}_E \in \mathbb{R}^D$ is a condensed representation of entities appearing in the conversational history.

The recommendation module classifies $\boldsymbol{c}_E$ directly into one of the items. We directly take the entity embedding $\boldsymbol{e}_i$ from the R-GCN network as the representation of the item. The probability of recommending item i is computed with softmax:

$$P_{rec}(i) \propto \exp(\boldsymbol{c}_E^\top \boldsymbol{e}_i). \tag{2}$$

The module is trained using the cross-entropy loss. To avoid the model recommending the same movie that the user might have just mentioned, we only consider as a ground-truth recommendation the movie that is first time to be mentioned by the recommender in the conversation.

3.2 Response Generation Network

The response generation module predicts the utterance to the user word by word. We use the classic encoder-decoder Transformer architecture (Vaswani et al., 2017), where the encoder encodes the entire conversational history word by word.

At decoding time step j, the output of the Transformer decoder $\boldsymbol{s}_j$ is concatenated with the entity representation $\boldsymbol{c}_E$ and goes through two fully connected layers before the softmax function. The probability distribution over the vocabulary is

$$P_{res} = \text{softmax}\big(\mathbf{W}_v \boldsymbol{W}_a [\mathbf{s}_j; \mathbf{c}_E] + \boldsymbol{b}\big), \qquad (3)$$

where $\mathbf{W}_v$ is the word embedding matrix shared with the encoder. $\mathbf{W}_a$ is a trainable linear projection to align the dimensions, and $\boldsymbol{b}$ is the bias. We train the module using cross-entropy at every decoder time step.

To separate movie names from other words in the conversation, for every movie name we create specialized tokens in the vocabulary. For example, the token for the movie name *It* is separate from the word token *it*. This is feasible as the dataset, REDIAL, has explicitly represented movie names with special strings.

3.3 Bag-of-Entities Loss

Although the response generation module trained using per-step cross-entropy is capable of recommending items, it rarely mentions concepts related to the recommended item. We postulate that mentioning related entities will produce natural conversations. For example, when recommending the movie *It*, one may want to mention that it is a horror movie based on a book by Stephen King.

For this purpose, we introduce the Bag-of-Entity (BOE) loss, which encourages the decoder state $[\mathbf{s}_j; \mathbf{c}_E]$ to contain additional information about first-order neighbors of the ground-truth recommendation on the KG.

First, at every time step, we compute a score $\boldsymbol{r}_j \in \mathbb{R}^M$ for all M entities in the knowledge graph,

$$\boldsymbol{r}_j = \mathbf{H}\mathbf{W}_b[\mathbf{s}_j; \mathbf{c}_E] + \mathbf{b}_{ent}, \qquad (4)$$

where $\mathbf{H}$ contains the embeddings of all KG entities, as produced by the R-GCN. $\mathbf{W}_b$ is a trainable matrix for dimension alignment and $\mathbf{b}_{ent}$ the bias.

As we do not constrain exactly which word in the response should contain the information, we sum up the word-level scores and then apply the component-wise sigmoid function. The probability that entity m is mentioned in the response is thus

$$P_{\text{BOE}}(m) = \text{sigmoid}(\sum_{j=1}^{L} r_{jm}), \qquad (5)$$

where L is the length of the response and r_{jm} is the m^{th} component of $\boldsymbol{r}_j$.

We apply a binary cross-entropy loss for each KG entity. The ground-truth label is 1 if the entity is a first-order neighbor of the recommended item on the knowledge graph and 0 otherwise.

3.4 Aligning Word and Entity Embeddings

We create two types of tokens in the vocabulary V of the response generation network. The first type corresponds to a plain word appearing in the conversation text. The second type represents an entity that appears in the conversation and in the knowledge graph.

To tie the token embeddings of the second type to the R-GCN encoding of the knowledge graph, we propose the alignment loss. For a conversation, we use the entity representation $\mathbf{c}_E$ in Eq. (1) to represent all entities in the conversation and calculate the similarity score between $\mathbf{c}_E$ and each word embedding,

$$\mathbf{s} = \mathbf{W}_{v[E]}\mathbf{W}_c\mathbf{c}_E + \mathbf{b}_{align}, \qquad (6)$$

where $\mathbf{W}_{v[E]}$ is the matrix resulting from selecting the rows of $\mathbf{W}_v$ corresponding to entity tokens only. $\mathbf{W}_c$ is a trainable matrix and $\mathbf{b}_{align}$ is the bias. The alignment loss is the mean square error between the $\mathbf{s}$ and an indicator vector $\mathbf{q} \in \{0,1\}^{|E|}$.

$$L_{align} = ||\boldsymbol{s} - \boldsymbol{q}||^2 \qquad (7)$$

Specifically, if an entity e exists in the conversation, the corresponding component of $\mathbf{q}$ is set to 1. Otherwise, the component is 0.

Finally, to learn the parameters of generation module, we minimize the following objective function:

$$L_{total} = L_{gen} + \lambda_1 L_{\text{BOE}} + \lambda_2 L_{align}, \qquad (8)$$

where λ_1 and λ_2 are two hyperparameters. In the testing procedure, the probability distribution over the vocabulary at time step j is calculated as follows,

$$P_{all} = P_{res} + \lambda_3 P_{boe}, \qquad (9)$$

where λ_3 is a hyperparameter.

Model	Automatic			Human		
	Dist-2	Dist-3	Dist-4	Fluency	Relevancy	Informativeness
HRED-CRS	0.10	0.18	0.24	1.92	1.62	1.05
Transformer	0.15	0.31	0.46	2.03	1.73	1.36
KBRD	0.31	0.38	0.52	2.10	1.72	1.32
KGSF	0.38	0.61	0.73	2.32	2.11	1.56
KECRS(Ours)	**0.48***	**0.91***	**1.23***	**2.56***	**2.29***	**2.18***

Table 1: Automatic and human evaluation results of the response generation achieved by different methods. Human evaluation scores are from 0-3. Dist-2,3,4 is short for Distinct-2,3,4. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*

4 Experiments

4.1 Dataset

We use the REDIAL dataset (Li et al., 2018), which includes 10,006 conversations and 182,150 utterances related to 51,699 movies. Following (Li et al., 2018; Chen et al., 2019; Zhou et al., 2020a), we split REDIAL into training, validation, and testing sets with the ratio 8:1:1. We build the knowledge graph, TMDKG, from The Movie Database[1], which contains 15822 entities and 15 types of relations.

4.2 Evaluation Metrics

Following (Chen et al., 2019; Zhou et al., 2020a), we use Distinct n-gram (n=2, 3, 4) to measure the diversity of generated responses. To better evaluate the performance of generated responses, we adopt human evaluation. We randomly sample 100 multi-turn conversations from the test set and invite three annotators to score responses generated by different models from the following aspects: 1) **Fluency**: whether responses are fluent;2) **Relevancy**: whether responses are correlated with contexts;3) **Informativeness**: whether responses contain rich information of recommended items. Each aspect is rated in $[0, 3]$, and final scores are the average of all annotators. For all evaluation metrics, the higher value indicates better performances.

4.3 Baseline Methods

We compare KECRS with the following baseline methods: 1) **HRED-CRS** (Li et al., 2018): This is a basic CRS based on HRED(Serban et al., 2016); 2) **Transformer** (Vaswani et al., 2017): This is a basic transformer model that generates responses only from utterance text and does not contain a separate recommendation module; 3) **KBRD** (Chen et al., 2019):This is a knowledge-based CRS that employs DBpedia to understand the user's intentions and leverage KG information as a bias for generation; 4) **KGSF** (Zhou et al., 2020a): This method exploits both entity-oriented and word-oriented KGs to enrich the data representations. It adopts two KG-enriched decoder layers for the generation.

Model	Dist-2	Dist-3	Dist-4
KGSF	0.38	0.61	0.73
KECRS$_{\text{w/o BOE}}$	0.31	0.64	0.87
KECRS$_{\text{w/o align}}$	0.36	0.69	0.95
KECRS	**0.48***	**0.91***	**1.23***

Table 2: Response generation performances of KGSF and different variants of KECRS. * indicates that the improvement over the best baseline method is statistically significant with $p < 0.01$ using student *t-test*

4.4 Results and Discussion

The automatic and human evaluation results of different methods are shown in Table 1. We note that Transformer performs better than HRED-CRS, which demonstrates that Transformer is powerful to understand and generate natural language. KBRD performs better than Transformer, because it adds a vocabulary bias to fuse knowledge from KG into the generated responses. Among all the baseline models, KGSF generates the most diverse responses, by exploiting both TMDKG and ConceptNet (Speer et al., 2017). The potential reason is that KGSF employs two additional KG-based attention layers to make the generative model focus more on items and relevant entities in TMDKG and ConceptNet. Moreover, the proposed KECRS model outperforms all baseline methods with a large margin in terms of all evaluation metrics. This demonstrates that the proposed BOE loss and alignment loss can work jointly to better leverage KG and generate more diverse and informative responses.

[1] https://www.themoviedb.org/

For human evaluation, we note that *Fluency* is relatively higher compared to *Informativeness* and *Relevancy* for all models. This indicates that responses generated by these models are fluent and can be understood by human judges. However, responses generated by baseline models are more likely to be generic responses (*e.g.,* “I haven’t seen that one”). By including additional supervision signals and aligning embeddings of word and entities, the proposed KECRS model alleviates this issue. Overall, KECRS can understand the dialogue context and generate fluent, relevant, and informative responses.

4.5 Ablation Study

To better understand effectiveness of each component in KECRS, we study the performances of following two variants of KECRS: 1) **KECRS$_{\text{w/o BOE}}$**, which removes the BOE loss, and 2) **KECRS$_{\text{w/o align}}$**, which removes the infusion loss.

Table 2 summarizes the response generation performance in terms of Distinct n-gram (n=2,3,4). Distinct n-gram measures the diversity of sentences by calculating the number of distinct n-gram in generated responses. KECRS outperforms KECRS$_{\text{w/o BOE}}$, which indicates the proposed BOE loss can help the model learn to generate responses not only from conversations but also from the knowledge graph. Moreover, KECRS$_{\text{w/o align}}$ performs poorer than KECRS. This indicates that aligning the the word embeddings and entity embeddings also helps improve the model performances. Compared with KGSF, both ablated versions of KECRS can achieve better performances in terms of most metrics. This again demonstrates that encouraging model to mention concept related to the recommended items and aligning word embeddings with KG entity embeddings both can help model generate more diverse responses.

4.6 Case Study

In Table 3, we present a qualitative comparison of the responses generated by different models. The conversation is selected from the dataset REDIAL. When the user expresses preferences on "*There’s Something About Mary*", KECRS infers the user may like romantic comedy movie. Thus, KECRS recommends another romantic comedy movie "*Meet the Parents*" and provides an informative and natural response *"It ’s a classic. It’s a little older , but still funny and romantic"*.

User: Recommender: User:	Hi there, how are you? I’m doing great, how about you? Fine thanks. I ’d love to see something similar to ***There’s Something About Mary***. That movie always cracks me up !
Transformer:	I haven’t seen that one.
KBRD:	I haven’t seen that one. What about ***My Best Friend’s Wedding***?
KGSF:	I recommend ***The Other Woman***?
KECRS (Ours):	I love ***Meet the Parents***. It ’s a ***classic***. It’s a little ***older*** , but still ***funny*** and ***romantic***.

Table 3: Case Study.

5 Conclusions

In this paper, we propose a novel Knowledge-Enriched Conversational Recommendation System (KECRS). Specifically, we develop the Bag-of-Entity (BOE) loss and the alignment loss to improve the response generation performances. The experimental results on REDIAL demonstrate that the proposed BOE loss can guide the model to generate more knowledge-enriched responses by selecting entities in KG, and the alignment loss can ties the word embeddings to the entity embeddings. Overall, KECRS achieves superior response quality than state-of-the-art baselines.

For future work, we would like to investigate how to use keywords to conduct the conversation from chit-chat to the recommendation (Zhou et al., 2020b; Liu et al., 2020). Moreover, we are also interested in using external knowledge (*e.g.,* KG) to modify the REDIAL dataset and make responses more relevant to the recommended items (Jannach and Manzoor, 2020).

6 Acknowledgments

The authors gratefully acknowledge the support from the Alibaba-NTU Joint Research Institute, Nanyang Associate Professorship, and the National Research Foundation, Singapore (NRF-NRFF13-2021-0006). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of the funding agencies.

References

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. *arXiv preprint arXiv:1908.05391*.

Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H. Chi. 2018. Q&R: A two-stage approach toward interactive recommendation. In *SIGKDD*, pages 139–148.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *SIGKDD*, pages 815–824.

Min Fu, Jiwei Guan, Xi Zheng, Jie Zhou, Jianchao Lu, Tianyi Zhang, Shoujie Zhuo, Lijun Zhan, and Jian Yang. 2020. ICS-Assist: Intelligent customer inquiry resolution recommendation in online customer service for large E-commerce businesses. *arXiv preprint arXiv:2008.13534*.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. *arXiv preprint arXiv:2009.14306*.

Dietmar Jannach and Ahtsham Manzoor. 2020. End-to-End learning for conversational recommendation: A long way to go? In *RecSys*, pages 72–76.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020a. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *WSDM*, pages 304–312.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020b. Interactive path reasoning on graph for conversational recommendation. In *SIGKDD*, pages 2073–2083.

Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, and Lei Wang. 2017. Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *CIKM*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *NeurIPS*, pages 9725–9735.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. *arXiv preprint arXiv:2005.03954*.

Rajdeep Sarkar, Koustava Goswami, Mihael Arcan, and John Philip McCrae. 2020. Suggest me a movie for tonight: Leveraging knowledge graphs for conversational recommendation. In *COLING*, pages 4179–4189.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Shuangyong Song, Chao Wang, Haiqing Chen, and Huan Chen. 2021. An emotional comfort framework for improving user satisfaction in E-commerce customer service chatbots. In *NAACL*, pages 130–137.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *SIGIR*, pages 235–244.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. 2021. Adapting user preference to online feedback in multi-round conversational recommendation. In *WSDM*, page 364–372.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *SIGIR*, pages 245–254.

Yinan Zhang, Boyang Li, Yong Liu, Yuan You, and Chunyan Miao. 2022. Minimalist and high-performance conversational recommendation with uncertainty estimation for user preference.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *CIKM*, pages 177–186.

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 *SIGKDD*.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*.

Jie Zou, Yifan Chen, and Evangelos Kanoulas. 2020. Towards question-based recommender systems. *arXiv preprint arXiv:2005.14255*.