

Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion

Kun Zhou
Peking University
franciszhou@pku.edu.cn

Wayne Xin Zhao*†
Gaoling School of Artificial
Intelligence, Renmin University of
China
batmanfly@gmail.com

Shuqing Bian
School of Information, Renmin
University of China
bianshuqing@ruc.edu.cn

Yuanhang Zhou
School of Information, Renmin
University of China
sdzyh002@gmail.com

Ji-Rong Wen†
Gaoling School of Artificial
Intelligence, Renmin University of
China
jrwen@ruc.edu.cn

Jingsong Yu
Peking University
yjs@ss.pku.edu.cn

ABSTRACT

Conversational recommender systems (CRS) aim to recommend high-quality items to users through interactive conversations. Although several efforts have been made for CRS, two major issues still remain to be solved. First, the conversation data itself lacks of sufficient contextual information for accurately understanding users' preference. Second, there is a semantic gap between natural language expression and item-level user preference.

To address these issues, we incorporate both word-oriented and entity-oriented knowledge graphs (KG) to enhance the data representations in CRSs, and adopt Mutual Information Maximization to align the word-level and entity-level semantic spaces. Based on the aligned semantic representations, we further develop a KG-enhanced recommender component for making accurate recommendations, and a KG-enhanced dialog component that can generate informative keywords or entities in the response text. Extensive experiments have demonstrated the effectiveness of our approach in yielding better performance on both recommendation and conversation tasks.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

Conversational Recommender System; Knowledge Graph; Mutual Information Maximization

*Corresponding author.

†Also with Beijing Key Laboratory of Big Data Management and Analysis Methods.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '20, August 23–27, 2020, Virtual Event, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7998-4/20/08...\$15.00

<https://doi.org/10.1145/3394486.3403143>

ACM Reference Format:

Kun Zhou, Wayne Xin Zhao*†, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen†, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*, August 23–27, 2020, Virtual Event, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3394486.3403143>

1 INTRODUCTION

Recently, conversational recommender system (CRS) [4, 18, 19, 26, 34] has become an emerging research topic in seeking to provide high-quality recommendations through conversations with users. Different from traditional recommender systems, it emphasizes interactive clarification and explicit feedback in natural languages, and has a high impact on e-commerce.

In terms of methodology, CRS requires a seamless integration between a recommender component and a dialog component. On one hand, the dialog component clarifies user intents and replies to the previous utterance with suitable responses. On the other hand, the recommender component learns user preference and recommends high-quality items based on contextual utterances. To develop an effective CRS, several solutions have been proposed to integrate the two components, including belief tracker over semi-structured user queries [26, 34] and switching decoder for component selection [18, 19].

Although these studies have improved the performance of CRS to some extent, two major issues still remain to be solved. First, a conversation mainly consists of a few sentences, lack of sufficient contextual information for accurately understanding user preference. As shown in Table 1, a user is looking for scary movies similar to “*Paranormal Activity (2007)*”, where her/his preference is simply described by two short sentences. In order to capture the user's intent, we need to fully utilize and model the contextual information. In this example, it is important to understand the underlying semantics of the word “scary” and the movie “*Paranormal Activity (2007)*”. Apparently, it is difficult to obtain such fact information solely based on the utterance text. Second, utterances are represented in natural languages, while actual user preference is reflected over the items or entities (e.g., actor and genre). There is a natural semantic gap between the two kinds of data signals. We need an effective

Table 1: An illustrative example of a user-system conversation for movie recommendation. The mentioned movies and important context words are marked in italic blue font and red font, respectively.

User	I am looking for some movies
System	What kinds of movie do you like?
User	Today I'm in a mood for something <i>scary</i> . Any similar movies like <i>Paranormal Activity (2007)</i> ?
System	<i>It (2017)</i> might be good for you. It is a classic <i>thriller</i> movie with <i>good plot</i> .
User	Great! Thank you!

semantic fusion way to *understand* or *generate* the utterances. As shown in Table 1, the system has presented both the recommended movie and the reason for recommendation. Without bridging the semantic gap, it is infeasible to generate the text for explaining the recommendation, e.g., “*thriller movie with good plot*”.

For enriching the conversation information, external knowledge graph (KG) has been utilized in CRS [4]. They mainly focus on incorporating item knowledge, while the word-level enrichment (e.g., the relation between “*scary*” and “*thriller*”) has been somehow neglected. Furthermore, they have not considered the semantic gap between natural language and external knowledge. Therefore, the utilization of KG data is likely to be limited. In essence, the problem originates from the fact that the dialog component and the recommender component correspond to two different semantic spaces, namely word-level and entity-level semantic spaces. Our idea is to incorporate two special KGs for enhancing data representations of both components, and fuse the two semantic spaces by associating the two KGs.

To this end, in this paper, we propose a novel conversational recommendation approach via KG based semantic fusion. Specially, we incorporate a word-oriented KG (i.e., ConceptNet [23]) and an item-oriented KG (i.e., DBpedia [2]). ConceptNet provides the relations between words, such as the synonyms, antonyms and co-occurrence words of a word; DBpedia provides the structured facts regarding the attributes of items. We first apply graph neural networks to learn node embeddings over the two KGs separately, and then propose to apply the Mutual Information Maximization [25, 28, 33] method to bridge the semantic gap between the two KGs. The core idea is to force the representations of nodes in the two KGs to be close given the word-item co-occurrence in the conversation. In this way, we can unify the data representations in the two semantic spaces. Such a step is particularly useful to connect contextual words with items (including the mentioned entities) in conversations. Based on the aligned semantic representations, we further develop a KG-enhanced recommender component for making accurate recommendations, and a KG-enhanced dialog component that can generate informative keywords or items in the response text.

To our knowledge, it is the first time that the integration of dialog and recommender systems has been addressed by using KG-enhanced semantic fusion. Our model utilizes two different KGs to enhance the semantics of words and items, respectively, and unifies their representation spaces. Extensive experiments on a public CRS

dataset have demonstrated the effectiveness of our approach in both recommendation and conversation tasks.

2 RELATED WORK

Conversational recommendation system (CRS) contains two major modules, namely the recommender component and the dialog component. We first introduce the related work in the two aspects.

Recommender systems aim to identify a subset of items that meet the user’s interest from the item pool. Traditional methods are highly based on the historical user-item interaction (e.g., click and purchase) [3, 15]. However, user-item interaction data is usually sparse. To tackle the data sparsity problem, many techniques have been developed by utilizing the side information of items, such as review and taxonomy data [1, 9, 30]. As a comparison, CRS mainly focuses on the recommendation setting through conversation instead of historical interaction data. Especially, knowledge graphs have been widely adopted to enhance the recommendation performance and explainability [10, 31, 37].

Conversation systems aim to generate proper responses given multi-turn contextual utterances. Existing works can be categorized into retrieval-based methods [11, 40] and generation-based methods [17, 21]. The first category of approaches try to find the most reasonable response from a large repository of historical conversations [11, 32], and the generation-based methods utilize learnable models to produce the response text. Based on the attentive seq2seq architecture [29], various extensions have been made to tackle the “safe response” problem and generate informative responses [17, 22, 38, 39].

Early conversational or interactive recommendation systems mainly utilized predefined actions to interact with users [5]. Recently, several studies started to integrate the two components for understanding users’ needs and recommend the right items through natural language conversation [16, 26, 35]. Overall, these methods emphasize the precise recommendation, while the conversation component is implemented by simple or heuristic solutions. Specially, a standard CRS dataset has been released in [18], and a hierarchical RNN model was proposed for utterance generation. Furthermore, follow-up studies [4, 19] incorporated external KG to improve the CRS, where their focus was to mainly enhance the item representations.

Based on previous studies, we design a novel conversational recommendation approach by incorporating and fusing word-level and entity-level knowledge graphs. Via KG fusion, our model is able to learn better data representations in both recommender and dialog components, which leads to better performance on item recommendation and utterance generation.

3 PRELIMINARIES

Conversational recommendation systems (CRS) aim to recommend proper items to a user through a multi-turn conversation. In the conversation, a chat agent analyzes and learns the preference of the user according to contextual conversation history. Then, it either generates appropriate recommendations or starts a new round of conversation for further clarification. The process ends until the task succeeds or the user leaves. In a CRS, there are two major components to develop, namely the recommender component and

the dialog component. The two components should be integrated seamlessly, and successful recommendation is considered as the final goal.

Formally, let u denotes a user from user set \mathcal{U} , i denotes an item from item set \mathcal{I} , and w denotes a word from vocabulary \mathcal{V} . A conversation (or a conversation history) C consists of a list of utterances, denoted by $C = \{s_t\}_{t=1}^n$, in which each utterance s_t is a conversation sentence at the t -th turn. At the t -th turn, the recommender component selects a set of candidate items \mathcal{I}_t from the entire item set \mathcal{I} according to some strategy, while the dialog component needs to produce the next utterance s_t to reply to previous utterances. Note that \mathcal{I}_t can be equal to \emptyset when there is no need for recommendation. In such a case, the dialog component may raise a clarification question or generate a chit-chat response. Given a n -turn conversation, the goal of a CRS is to generate the *response utterance* to the user, including both the recommendation set \mathcal{I}_{n+1} and the reply utterance s_{n+1} .

4 APPROACH

In this section, we present the KG-based Semantic Fusion approach to the CRS task, named *KGSF*. We first introduce how to encode both word-oriented and item-oriented KGs, and then fuse the semantics of the two KGs. Based on the fused KGs, we finally describe our solutions for both recommendation and conversation tasks. The overview illustration of the proposed model is presented in Fig. 1.

4.1 Encoding External Knowledge Graphs

As shown in Table 1, it is difficult to fully understand user preference and generate a suitable response. We identify the two basic semantic units in dialog and recommender systems, namely *word* and *item*, respectively. Here, we utilize two separate KGs to enhance the representations of the basic semantic units on both sides.

4.1.1 Encoding Word-oriented KG. We adopt the widely used ConceptNet [23] as the word-oriented KG. It stores a semantic fact as a triple $\langle w_1, r, w_2 \rangle$, where $w_1, w_2 \in \mathcal{V}$ are words and r is a word relation. Not all the words in ConceptNet are useful for our task. Hence, we only consider words that appear in our corpus, and extract their related triples from ConceptNet. We also remove words with very few triples related to the words in our corpus.

To encode the word-oriented KG, we adopt graph convolutional neural network [6, 14] (GCN) for capturing the semantic relations between word nodes. At each update, GCN receives the information from the one-hop neighborhood in the graph and performs the aggregation operation as:

$$\mathbf{V}^{(l)} = \text{ReLU}(\mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \mathbf{V}^{(l-1)} \mathbf{W}^{(l)}) \quad (1)$$

where $\mathbf{V}^{(l)} \in \mathbb{R}^{V \times d_W}$ are the representations of nodes and $\mathbf{W}^{(l)}$ is a learnable matrix at the l -th layer, \mathbf{A} is the adjacency matrix of the graph and \mathbf{D} is a diagonal degree matrix with entries $\mathbf{D}[i, i] = \sum_j \mathbf{A}[i, j]$. By stacking multiple convolutions, node information can be propagated along with the graph structure. When the algorithm ends, we can obtain a d_W -dimensional representation \mathbf{n}_w for a word w . Here, we do not incorporate the relation information, because the number of relations is large and many relations are not directly useful to the recommendation task.

4.1.2 Encoding Item-oriented KG. Another kind of semantic units to consider are items and their related entities. Following [4], we utilize DBpedia [2] as item-oriented KG. Similar to ConceptNet, a triple in DBpedia is denoted by $\langle e_1, r, e_2 \rangle$, where $e_1, e_2 \in \mathcal{E}$ are items or entities from the entity set \mathcal{E} and r is entity relation from the relation set \mathcal{R} . To extract the entity subgraph, we collect all the entities appearing in our corpus by following the approach from [4]. Starting from these items and entities as seeds, we extract their one-hop triples on the DBpedia graph.

For item-oriented KG, relation semantics are important to consider. Different from ConceptNet, we utilize R-GCN [20] to learn item representations on the extracted subgraph. Formally, the representation of node e at $(l+1)$ -th layer is calculated as:

$$\mathbf{n}_e^{(l+1)} = \sigma \left(\sum_{r \in \mathcal{R}} \sum_{e' \in \mathcal{E}_e^r} \frac{1}{Z_{e,r}} \mathbf{W}_r^{(l)} \mathbf{n}_{e'}^{(l)} + \mathbf{W}^{(l)} \mathbf{n}_e^{(l)} \right) \quad (2)$$

where $\mathbf{n}_e^{(l)} \in \mathbb{R}^{d_E}$ is the node representation of e at the l -th layer, \mathcal{E}_e^r denotes the set of neighboring nodes for e under the relation r , $\mathbf{W}_r^{(l)}$ is a learnable relation-specific transformation matrix for the embeddings from neighboring nodes with relation r , $\mathbf{W}^{(l)}$ is a learnable matrix for transforming the representations of nodes at the l -th layer and $Z_{e,r}$ is a normalization factor.

4.2 KG Fusion via Mutual Information Maximization

Above, we obtain the node representations for the word-oriented KG and item-oriented KG, denoted by two embedding matrices \mathbf{V} (\mathbf{v}_w for word w) and \mathbf{N} (\mathbf{n}_e for item e), respectively.

In order to bridge the semantic gap between words and items, we propose to use Mutual Information Maximization technique [25, 28, 33], called *MIM*. MIM has been used to mutually improve the data representations of two coupled signals (e.g., input and output). Its core idea is based on the concept of Mutual Information (MI). Given two variables X and Y , their MI is defined as:

$$MI(X, Y) = D_{KL}(p(X, Y) || p(X)p(Y)), \quad (3)$$

where D_{KL} is the Kullback-Leibler (KL) divergence between the joint distribution $p(X, Y)$ and the product of marginals $p(X)p(Y)$. Usually, MI is difficult to compute. MIM tries to maximize it instead of seeking the precise value via the following formula:

$$MI(X, Y) \geq \mathbb{E}_P[g(x, y)] - \mathbb{E}_N[g(x', y')], \quad (4)$$

where \mathbb{E}_P and \mathbb{E}_N denote the expectation over positive and negative samples respectively, and $g(\cdot)$ is the binary classification function that outputs a real number which can be modeled by a neural network.

In our setting, we have two kinds of semantic units (namely words and entities), and would like to align their semantic representation spaces. Given a conversation, we first collect words (non-stopwords) and entities (including items) from the utterance text. For an entity-word pair $\langle e, w \rangle$ that co-occur in a conversation, we pull their representations close through a transformation matrix:

$$g(e, w) = \sigma(\mathbf{n}_e^\top \cdot \mathbf{T} \cdot \mathbf{v}_w), \quad (5)$$

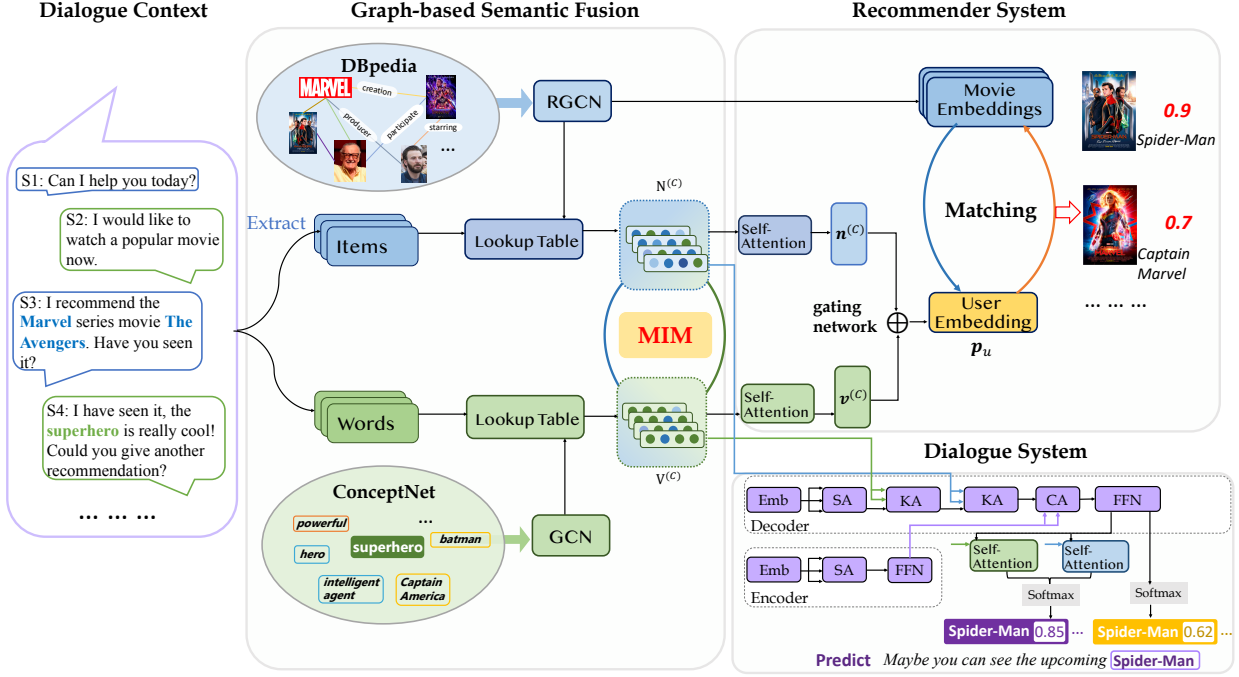


Figure 1: The overview of our model with a movie recommendation scenario. Here, “SA”, “KA”, and “CA” denotes *self-attention*, *KG-based attention* and *context-based attention*, respectively.

where \mathbf{n}_e and \mathbf{v}_w are the learned node representations for entity e and word w via KGs, respectively, $\mathbf{T} \in \mathbb{R}^{d_E \times d_W}$ is the transformation matrix that aligns two semantic spaces, and $\sigma(\cdot)$ is the sigmoid function. To apply the MIM method, we can consider all the word-entity pairs co-occurring in a conversation as *positive*, while random word-entity pairs are considered as *negative*. By integrating Eq. 5 into Eq. 4, we can derive the objective loss over all the conversations and minimize the loss with an optimization algorithm.

However, a conversation usually contains a number of contextual words, and it is time-consuming to enumerate all the word-entity pairs. Besides, some of these words are noisy, which is likely to affect the final performance. Here, we add a super token \tilde{w} for a conversation, assuming that it is able to represent the overall semantics of the contextual words. Instead of considering all the word-entity pairs, we only model the relation between each entity and the super token \tilde{w} using the $g(\cdot)$ function. We utilize self-attention mechanism for learning the representation of \tilde{w} :

$$\begin{aligned} \mathbf{v}_{\tilde{w}} &= \mathbf{V}^{(C)} \cdot \boldsymbol{\alpha}, \\ \boldsymbol{\alpha} &= \text{softmax}(\mathbf{b}^\top \cdot \tanh(\mathbf{W}_\alpha \mathbf{V}^{(C)})), \end{aligned} \quad (6)$$

where $\mathbf{V}^{(C)}$ is the matrix consisting of the embeddings of all the contextual words in a conversation C , $\boldsymbol{\alpha}$ is an attention weight vector reflecting the importance of each word, and \mathbf{W}_α and \mathbf{b} are parameter matrix and vector to learn. Using such a super token, we can significantly improve efficiency and identify more important semantic information from the entire conversation.

In order to effectively align the semantic space of the two KGs, we adopt the MIM loss for pre-training the parameters of the GNN models in Section 4.1.1 and 4.1.2, which forces the two semantic spaces to be close at the beginning. During the fine-tuning stage, we treat the MIM loss as a regularization constraint for GNN to prevent overfitting.

With the representations of the fused KGs, we next describe how to make recommendations and generate utterances in Section 4.3 and 4.4, respectively.

4.3 KG-enhanced Recommender Module

Given the learned word and item representations, we study how to generate a set of items for recommendation in CRS.

A key point for recommendation is to learn a good representation of user preference. Different from traditional recommender systems, following [4, 18], we assume no previous interaction records are available. We can only utilize the conversation data to infer user preference.

First, we collect all the words that appear in a conversation c . By using a simple lookup operation, we can obtain the word or item embeddings learned through the graph neural networks in Section 4.2. We concatenate the word embeddings into a matrix $\mathbf{V}^{(C)}$. Similarly, we can derive an item embedding matrix $\mathbf{N}^{(C)}$ by combining the embeddings of items.

Next, we apply the similar self-attentive mechanism in Eq. 6 to learn a single word vector $\mathbf{v}^{(C)}$ for $\mathbf{V}^{(C)}$ and a single item vector $\mathbf{n}^{(C)}$ for $\mathbf{N}^{(C)}$. In order to combine the two parts of information, we

apply the gate mechanism to derive the preference representation \mathbf{p}_u of the user u :

$$\begin{aligned}\mathbf{p}_u &= \beta \cdot \mathbf{v}^{(C)} + (1 - \beta) \cdot \mathbf{n}^{(C)}, \\ \beta &= \sigma(\mathbf{W}_{\text{gate}}[\mathbf{v}^{(C)}; \mathbf{n}^{(C)}]),\end{aligned}\quad (7)$$

Given the learned user preference, we can compute the probability that recommends an item i from the item set to a user u :

$$\text{Pr}_{rec}(i) = \text{softmax}(\mathbf{p}_u^\top \cdot \mathbf{n}_i), \quad (8)$$

where \mathbf{n}_i is the learned item embedding for item i . We can utilize Eq. 8 to rank all the items and generate a recommendation set to a user. To learn the parameters, we set a cross-entropy loss as:

$$\begin{aligned}L_{rec} = - \sum_{j=1}^N \sum_{i=1}^M & \left[- (1 - y_{ij}) \cdot \log (1 - \text{Pr}_{rec}^{(j)}(i)) \right. \\ & \left. + y_{ij} \cdot \log (\text{Pr}_{rec}^{(j)}(i)) \right] + \lambda * L_{MIM},\end{aligned}\quad (9)$$

where j is the index of a conversation, i is the index of an item, L_{MIM} is the Mutual Information Maximization loss, and λ is a weighted parameter. Here, we loop the entire collection and compute the cross-entropy loss. We only present the case that a conversation has a ground-truth recommendation. While, it is straightforward to extend the above loss to the case with multiple ground-truth items.

4.4 KG-enhanced Response Generation Module

Here, we study how to generate a reply utterance in CRS. We adopt Transformer [27] to develop the encoder-decoder framework. Our encoder follows a standard Transformer architecture. We mainly introduce the KG-enhanced decoder.

To better generate responses at decoding, we incorporate KG-enhanced representations of context words and items. After the self-attention sub-layer, we conduct two KG-based attention layers to fuse the information from the two KGs:

$$\mathbf{A}_0^n = \text{MHA}(\mathbf{R}^{n-1}, \mathbf{R}^{n-1}, \mathbf{R}^{n-1}), \quad (10)$$

$$\mathbf{A}_1^n = \text{MHA}(\mathbf{A}_0^n, \mathbf{V}^{(C)}, \mathbf{V}^{(C)}), \quad (11)$$

$$\mathbf{A}_2^n = \text{MHA}(\mathbf{A}_1^n, \mathbf{N}^{(C)}, \mathbf{N}^{(C)}), \quad (12)$$

$$\mathbf{A}_3^n = \text{MHA}(\mathbf{A}_2^n, \mathbf{X}, \mathbf{X}), \quad (13)$$

$$\mathbf{R}^n = \text{FFN}(\mathbf{A}_3^n), \quad (14)$$

where $\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ defines the multi-head attention function [27] that takes a query matrix \mathbf{Q} , a key matrix \mathbf{K} , and a value matrix \mathbf{V} as input:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O, \quad (15)$$

$$\text{head}_i = \text{Attention}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V), \quad (16)$$

and $\text{FFN}(\mathbf{x})$ defines a fully connected feed-forward network, which consists of a linear transformation with a ReLU activation layer:

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{xW}_1 + b_1) \mathbf{W}_2 + b_2. \quad (17)$$

Above, \mathbf{X} is the embedding matrix output by the encoder, $\mathbf{V}^{(C)}$ and $\mathbf{N}^{(C)}$ are KG-enhanced representation matrices for words and items in a conversation c , respectively. And, \mathbf{A}_0^n , \mathbf{A}_1^n , \mathbf{A}_2^n and \mathbf{A}_3^n are the representations after self-attention, cross-attention with embeddings from ConceptNet, cross-attention with embeddings from DBpedia and cross-attention with encoder output, respectively. Finally, \mathbf{R}^n is the embedding matrix from the decoder at n -th layer.

Algorithm 1: The training algorithm for the KGFS model.

Input: The conversation recommendation dataset \mathcal{D} , item-oriented KG \mathcal{G}_1 , and word-oriented KG \mathcal{G}_2
Output: Model parameters Θ^g , Θ^r and Θ^d .
1 Randomly initialize Θ^g , Θ^r and Θ^d .
2 **Pre-train the Θ^g by minimizing the MIM loss in 4.2.**
3 **for** $t = 1 \rightarrow |\mathcal{D}|$ **do**
4 Acquire items' and words' representations from \mathcal{G}_1 and \mathcal{G}_2 by Eq. 1 and Eq. 2, respectively.
5 Acquire $\mathbf{v}^{(C)}$ and $\mathbf{v}^{(C)}$ by self-attention using Eq. 6.
6 Acquire \mathbf{p}_u by gate mechanism using Eq. 7.
7 Compute $\text{Pr}_{rec}(i)$ using Eq. 8.
8 Perform GD on Eq. 9 w.r.t. Θ^g and Θ^r .
9 **end**
10 **for** $i = 1 \rightarrow |\mathcal{D}|$ **do**
11 Acquire items' and words' representations from \mathcal{G}_1 and \mathcal{G}_2 by Eq. 1 and Eq. 2, respectively.
12 Acquire \mathbf{R}^n by KG-enhanced Transformer using Eq. 14.
13 Compute $\text{Pr}(y|y_1, \dots, y_{i-1})$ using Eq. 18.
14 Perform GD on Eq. 19 w.r.t. Θ^d .
15 **end**
16 **return** Θ^g , Θ^r and Θ^d .

Compared with a standard Transformer decoder, we have two additional steps in Eq. 11 and 12. The idea can be described using a transformation chain: *generated words* $\xrightarrow{\text{Eq. 11}}$ *word-oriented KG* $\xrightarrow{\text{Eq. 12}}$ *item-oriented KG* $\xrightarrow{\text{Eq. 13}}$ *context words*. Following such a chain, our decoder is able to gradually inject useful knowledge information from the two KGs in a sequential manner. The rationality for Eq. 12 lies in the fact that we have fused the two KGs as in Section 4.2.

Different from chit-chat models, the generated reply is expected to contain the recommended items, related entities and descriptive keywords. We further adopt the copy mechanism to enhance the generation of such tokens. Formally, given the predicted subsequence y_1, \dots, y_{i-1} , the probability of generating y_i as the next token is given as:

$$\text{Pr}(y_i|y_1, \dots, y_{i-1}) = \text{Pr}_1(y_i|\mathbf{R}_i) + \text{Pr}_2(y_i|\mathbf{R}_i, \mathcal{G}_1, \mathcal{G}_2), \quad (18)$$

where $\text{Pr}_1(\cdot)$ is the generative probability implemented as a softmax function over the vocabulary by taking the decoder output \mathbf{R}_i (Eq. 14) as input, $\text{Pr}_2(\cdot)$ is the copy probability implemented by following a standard copy mechanism [7] over the nodes of the two KGs, and $\mathcal{G}_1, \mathcal{G}_2$ denote the two KGs we have used. To learn the response generation module, we set the cross-entropy loss as:

$$L_{gen} = -\frac{1}{N} \sum_{t=1}^N \log (\text{Pr}(s_t|s_1, \dots, s_{t-1})), \quad (19)$$

where N is the number of turns in a conversation C . We compute this loss for each utterance s_t from C .

4.5 Parameter Learning

Our parameters to learn are organized by three groups, namely the KG module, recommender module and conversation module, denoted by Θ^g , Θ^r and Θ^d respectively. Algorithm 1 presents the

Table 2: Results on recommendation task. Numbers marked with * indicate that the improvement is statistically significant compared with the best baseline (t-test with p-value < 0.05).

Test Models	All data			Cold start		
	R@1	R@10	R@50	R@1	R@10	R@50
Popularity	0.012	0.061	0.179	0.020	0.097	0.239
TextCNN	0.013	0.068	0.191	0.011	0.081	0.239
ReDial	0.024	0.140	0.320	0.021	0.075	0.201
KBRD	0.031	0.150	0.336	0.026	0.085	0.242
KGSF	0.039*	0.183*	0.378*	0.039*	0.174*	0.370*
-MIM	0.037	0.175	0.356	0.037	0.158	0.331
-DB	0.027	0.121	0.256	0.030	0.168	0.346

training algorithm for our KGSF model. The three components share parameters and affect each other.

To train the joint model, we pre-train the knowledge graph module Θ^g using Mutual Information Maximization loss. Next, we optimize the parameters in Θ^r and Θ^g . At each iteration, we first acquire words’ and items’ representations from the KG module. Then, we perform the self-attention and gate mechanism to derive user representations. Finally, we compute a cross-entropy loss by Eq. 9 with the MIM regularization, and perform gradient descent to update parameters Θ^r and Θ^g .

When the loss of the recommender component converges, we optimize the parameters in Θ^d . At each iteration, we first obtain words’ and items’ representations from the knowledge graph module. Then, we utilize KG-enhanced Transformer to derive the contextual representations. Finally, we compute the cross-entropy loss by Eq. 19, and perform gradient descent to update parameters in Θ^d .

5 EXPERIMENT

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

5.1 Experiment Setup

5.1.1 Dataset. We evaluate our model on the *REcommendations through DIALog* (REDIAL) dataset, which is a conversational recommendation dataset released by [18]. This dataset was constructed through Amazon Mechanical Turk (AMT). Following a set of comprehensive instructions, the AMT workers generated dialogs for recommendation on movies in a seeker-recommender pair. It contains 10,006 conversations consisting of 182,150 utterances related to 51,699 movies. This dataset is split into training, validation and test sets using a ratio of 8:1:1. For each conversation, we start from the first sentence one by one to generate reply utterances or recommendations by our model.

5.1.2 Baselines. In CRS, we consider two major tasks for evaluation, namely recommendation and conversation.

- *Popularity*: It ranks the items according to historical recommendation frequencies in the corpus.
- *TextCNN* [12]: It adopts a CNN-based model to extract textual features from contextual utterances as user embedding.

- *Transformer* [27]: It applies a Transformer-based encoder-decoder framework to generate proper responses without information from recommender module.

- *REDIAL* [18]: This model has been proposed in the same paper with our dataset [18]. It basically consists of a dialog generation module based on HRED [22], a recommender module based on auto-encoder [8] and a sentiment analysis module.

- *KBRD* [4]: This model utilizes DBpedia to enhance the semantics of contextual items or entities. The dialog generation module is based on the Transformer architecture, in which KG information serves as word bias for generation.

Among these baselines, *Popularity* and *TextCNN* [12] are recommendation methods, and *Transformer* [27] is the state-of-the-art text generation method. We do not include other recommendation models, since there are no historical user-item interaction records except the text of a single conversation. Besides, *REDIAL* [18] and *KBRD* [4] are conversation recommendation methods. We name our proposed model as *KGSF*.

5.1.3 Evaluation Metrics. In our experiments, we adopt different metrics to evaluate the two tasks. For recommendation task, following [4], we adopt Recall@ k ($k = 1, 10, 50$) for evaluation. Besides the standard setting, we consider a specific scenario in recommender systems, namely *cold start*.

In CRS, this problem can be alleviated to some extent, since we have conversation contexts. In order to simulate the cold-start scenario in CRS, we only consider the test cases without any mentioned items in context. This experiment aims to examine whether our fusion strategy is useful to learn user preference from word-based utterances. For the conversation task, the evaluation consists of automatic evaluation and human evaluation. Following [4], we use Distinct n -gram ($n = 2, 3, 4$) to measure the diversity at sentence level. For CRS, it is particularly important that the dialog system is able to generate informative replies related to items or entities. Hence, we introduce a new metric that calculates the ratio of items in the generated utterances. Different from traditional conversation tasks, we do not need to generate response resembling the ground-truth utterance. Instead, the final goal is to successfully make the recommendations. For this reason, we adopt human evaluation (on a random selection of 100 multi-turn dialogs from the test set) instead of using BLEU metrics. We invite three annotators to score the generated candidates in two aspects, namely *Fluency* and *Informativeness*. The range of score is 0 to 2. The final performance is calculated using the average scores of the three annotators.

5.1.4 Implementation Details. We follow the procedure in Algorithm 1 to implement our approach with Pytorch¹. The dimensionality of embeddings (including hidden vectors) is set to 300 and 128, respectively, for conversation and recommender modules. We initialize word embeddings via word2vec². In the KG module, we set the layer number to 1 for both GNN networks. We use Adam optimizer [13] with the default parameter setting. In experiments, the batch size is set to 32, the learning rate is 0.001, gradient clipping restricts the gradients within [0,0.1], and the normalization constant $Z_{v,r}$ of R-GCN in Eq. 2 is 1. During pre-training, we directly

¹<https://pytorch.org/>

²<https://radimrehurek.com/gensim/models/word2vec.html>

optimize the MIM loss as Section 4.2. While, during fine-tuning, the weight λ of the MIM loss in Eq. 9 is 0.025. Our code is publicly available via the link: <https://github.com/RUCAIBox/KGSF>.

5.2 Evaluation on Recommendation Task

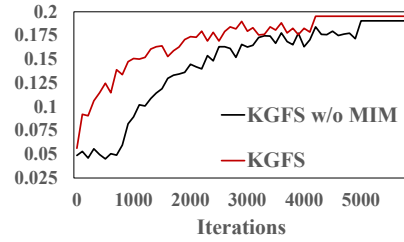
In this subsection, we conduct a series of experiments on the effectiveness of the proposed model for the recommendation task. Table 2 presents the performance of different methods in the two settings.

5.2.1 Evaluation on All Data Setting. We first consider the *all data* setting. As we can see, Popularity achieves a comparable performance with TextCNN. For our CRS task, utterance text is likely to be sparse and noisy, while Popularity utilizes global statistics for recommending popular items. Second, the two CRS models Redial and KBRD perform better than Popularity and TextCNN. Compared with TextCNN, Redial and KBRD only utilize the entities or items in context to make recommendations. Furthermore, KBRD performs better than Redial, since it incorporates external KG information. It indicates that KG data is useful to enhance the data representations and improves the performance of the CRS task. Finally, our model KGSF outperforms the baselines with a large margin. KGSF has incorporated both word-oriented and entity-oriented KGs, and further fuses the two KGs for enhancing the data representations.

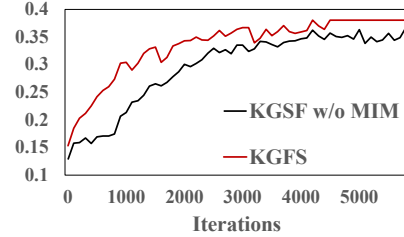
5.2.2 Evaluation on Cold Start Setting. For the *cold start* setting, first, the heuristic method Popularity performs very well, even better than TextCNN and ReDial in most cases. A possible reason is that in real-world recommender systems, a new user is likely to adopt a popular item. When there are no items or attributes mentioned in the context, the performance of KBRD and Redial has decreased substantially. As a comparison, our model KGSF still performs stably and achieves the best performance among all the methods. The major reason is that it not only utilizes item-oriented KG, but also uses word-oriented KG. By aligning the two semantic spaces, it can capture important evidence from utterance text to infer user preference in the cold start setting.

5.2.3 Ablation Study. In our model, we have incorporated two external KGs and adopted the Mutual Information Maximization method for KG fusion. Here, we would like to examine the contribution of each part. We incorporate two variants of our model for ablation analysis, namely *KGSF w/o MIM* and *KGSF w/o DB*, which remove the MIM loss and the DBpedia KG, respectively. Overall, we can see that both components contribute to the final performance. Besides, after removing the item-oriented KG, the performance decreases more significantly. It is because once it was removed, the corresponding fusion component with MIM is also removed.

5.2.4 The Effect of MIM Technique. We adopt the MIM technique to fuse semantic representations from two KGs. As shown in previous experiments, it is useful to improve the performance of our model on the recommendation task. Here, we would like to study whether its improvement is consistent and stable with the increase of the iteration number. We gradually increase the number of iterations for our model on the training set, and report the corresponding performance on the test set. Figure 2 shows how the performance of our model varies with the increase of iterations. We can see that



(a) Recall@10 in test set



(b) Recall@50 in test set

Figure 2: Performance (Recall@10 and Recall@50) comparison of KGSF *with* and *without* the MIM loss on test set.

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

Models	Dist-2	Dist-3	Dist-4	Item Ratio
Transformer	0.148	0.151	0.137	0.194
ReDial	0.225	0.236	0.228	0.158
KBRD	0.263	0.368	0.423	0.296
KGSF	0.289*	0.434*	0.519*	0.325*

with the MIM technique, our model can achieve a good result with fewer iterations compared to the variant without MIM. Overall, besides the performance improvement, the MIM technique is useful to improve the stability of the training process.

5.3 Evaluation on Conversation Task

In this subsection, we construct a series of experiments on the effectiveness of the proposed model on the conversation task.

5.3.1 Automatic Evaluation. We present the results of the automatic evaluation for different methods in Table 3. First, ReDial performs better than Transformer in Distinct-2/3/4, since it utilizes a pre-training RNN model to encode history utterances. While, Transformer performs better with the metric of *Item Ratio*. A possible reason is that Transformer architecture adopts the self-attention mechanism for capturing temporal pairwise interaction, which is more suitable to model the relations between words and items than RNN and CNN. Second, among the three baselines, KBRD generates the most diverse responses and achieves the highest item ratio, *i.e.*, containing more mentions of items in the generated text. This model

Table 4: Human evaluation results on the conversation task. Numbers marked with * indicate that the improvement is statistically significant compared with the best baseline (t-test with p-value < 0.05).

Models	Fluency	Informativeness
Transformer	0.92	1.08
ReDial	1.37	0.97
KBRD	1.18	1.18
KGSF	1.54*	1.40*

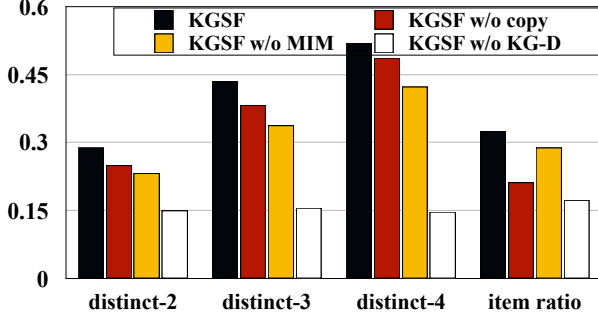


Figure 3: Ablation study on conversation task.

utilizes KG information to promote the predictive probability of entities and items. Compared with these baselines, our KGSF model is consistently better in all evaluation metrics. KGSF has utilized the KG information in two major steps, namely the knowledge-enhanced Transformer decoder and the copy mechanism, which enhances the informativeness of the generated text.

5.3.2 Human Evaluation. Table 4 presents the result of human evaluation for the conversation task. First, among the three baselines, ReDial performs best in terms of the metric of *Fluency*, since it utilizes a pre-training encoder on multiple language tasks [24]. However, we find that it tends to generate short and repetitive responses. This is the so-called “safe response” issue in the dialog generation task. Without additional supervision signal, it is likely to overfit to the frequent utterances in the training set. Second, KBRD performs best in terms of *Informativeness* score among the three baselines. It utilizes KG data to promote the probability of low-frequency words. Finally, our proposed model KGSF is consistently better than all the baselines with a large margin. We carefully design a KG-enhanced Transformer decoder. Our model is able to utilize contextual information effectively, and generate fluent and informative responses.

5.3.3 Ablation Study. We also conduct the ablation study based on three variants of our complete model, include: (1) *KGSF w/o KG-D* by removing the KG-based attention layers from the Transformer decoder, (2) *KGSF w/o copy* by removing the copy mechanism, and (3) *KGSF w/o MIM* by removing the MIM loss. As shown in Figure 3, first, all the techniques are useful to improve the final performance. Besides, the KG-based attention layer seems to be more important in our task, yielding a significant decrease when removed. KG-based attention layers can effectively inject the fused KG information into

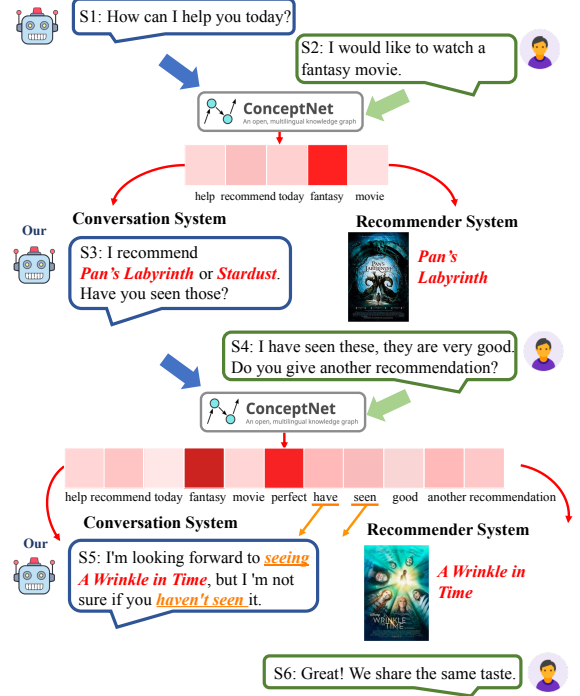


Figure 4: A sampled conversation with six-turn utterances between our CRS agent (recommender) and a real user (seeker). We use color bars to indicate attention weights of words in the recommendation component. The first-round recommendation is unsuccessful, while the second-round recommendation is successful.

the decoder by multi-head attention mechanism. For the MIM loss, besides its contribution to the recommendation task (See Table 2), it also improves the quality of the generated responses, which indicates its usefulness for KG-based semantic fusion.

5.4 Qualitative Analysis

In this part, we present a qualitative example to illustrate how our model works in practice.

In Fig. 4, a user requests the recommendations on *fantasy* movies, and our system accurately identifies the key word “*fantasy*” by assigning a larger attention weight. The attention weights are computed by the self-attentive mechanism in Eq. 6 based on the KG-enhanced word embeddings. With the focused preference of *fantasy*, our recommender component returns the candidate “*Pan’s Labyrinth*”. While, interestingly, our dialog component not only includes the mentions of the recommended movie, but also generates another related movie (“*Stardust*”) in the utterance. Receiving the response, the user rejects the recommendation since she/he has watched both movies before. Then, our recommender component updates the representation of user preference, and returns another recommendation. Although the words of “*have*” and “*seen*” received

a small attention weight by the recommender component, our dialog component also boosts their weights since they are helpful to generate a more informative reply.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel KG-based semantic fusion approach for CRS. By utilizing two external KGs, we enhanced the semantic representations of words and items, and used Mutual Information Maximization to align the semantic spaces for the two different components. Based on the aligned semantic representations, we developed a KG-enhanced recommendation component for making accurate recommendations, and a KG-enhanced dialog component that can generate informative keywords or entities in the utterance text. By constructing extensive experiments, our approach yielded better performance than several competitive baselines.

As future work, we will consider using more kinds of external information to improve the performance of CRS, e.g., user demographics [36]. Besides, we will investigate how to make the utterance more persuasive and explainable for the recommendation results. Finally, another interesting topic is how to incorporate historical user-item interaction data and start the conversation with a pre-learned user profile.

ACKNOWLEDGEMENT

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61872369 and 61832017, Beijing Academy of Artificial Intelligence (BAAI) under Grant No. BAAI2020ZJ0301, and Beijing Outstanding Young Scientist Program under Grant No. BJJWZYJH012019100020098, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China under Grant No.18XNLG22 and 19XNQ047. Xin Zhao is the corresponding author.

REFERENCES

- [1] Yang Bao, Hui Fang, and Jie Zhang. 2014. TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation. In *AAAI 2014*. 2–8.
- [2] 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. *J. Web Semant.* 7 (2009), 154–165.
- [3] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowl. Based Syst.* 46 (2013), 109–132.
- [4] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. (2019), 1803–1813.
- [5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *SIGKDD 2016*. 815–824.
- [6] Michael Edwards and Xianghua Xie. 2016. Graph Convolutional Neural Network. In *BMVC 2016*.
- [7] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *ACL 2016*.
- [8] Junhua He, Hankz Hankui Zhuo, and Jarvan Law. 2017. Distributed-Representation Based Hybrid Recommender System with Short Item Descriptions. *arXiv:cs.LG/1703.04854*
- [9] Jin Huang, Zhaochun Ren, Wayne Xin Zhao, Gaole He, Ji-Rong Wen, and Daxiang Dong. 2019. Taxonomy-Aware Multi-Hop Reasoning Networks for Sequential Recommendation. In *WSDM 2019*. 573–581.
- [10] Jin Huang, Wayne Xin Zhao, Hong-Jian Dou, Ji-Rong Wen, and Edward Y. Chang. 2018. Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks. In *SIGIR 2018*. 505–514.
- [11] Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An Information Retrieval Approach to Short Text Conversation. *ArXiv abs/1408.6988* (2014).
- [12] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP 2014*. 1746–1751.
- [13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015*.
- [14] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR 2017*.
- [15] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*. 426–434.
- [16] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *WSDM 20*. 304–312.
- [17] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2015. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL-HLT*.
- [18] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. (2018), 9748–9758.
- [19] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep Conversational Recommender in Travel. *ArXiv abs/1907.00710* (2019).
- [20] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. *Lecture Notes in Computer Science* (2018), 593–607.
- [21] Iulian Serban, Alessandro Sordani, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*.
- [22] Iulian Vlad Serban, Alessandro Sordani, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*.
- [23] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI 2017*. 4444–4451.
- [24] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *ICLR 2018*.
- [25] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. 2020. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. (2020).
- [26] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *SIGIR 2018*. 235–244.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. , 5998–6008 pages.
- [28] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep Graph Infomax.
- [29] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. *CoRR abs/1506.05869* (2015).
- [30] Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *SIGKDD 2011*. 448–456.
- [31] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation. In *AAAI*. 5329–5336.
- [32] Yuehua Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *ACL*.
- [33] Yi-Ting Yeh and Yun-Nung Chen. 2019. QAINfomax: Learning Robust Question Answering System by Mutual Information Maximization. (2019), 3368–3373.
- [34] Xiaoying Zhang, Hong Xie, Hang Li, and John C. S. Lui. 2019. Toward Building Conversational Recommender Systems: A Contextual Bandit Approach. *arXiv:cs.LG/1906.01219*
- [35] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *CIKM 2018*. 177–186.
- [36] Wayne Xin Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *SIGKDD*. 1935–1944.
- [37] Wayne Xin Zhao, Gaole He, Kunlin Yang, Hong-Jian Dou, Jin Huang, Siqi Ouyang, and Ji-Rong Wen. 2019. KB4Rec: A Data Set for Linking Knowledge Bases with Recommender Systems. *Data Intelligence* 1, 2 (2019), 121–136.
- [38] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*.
- [39] Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised Context Rewriting for Open Domain Conversation. In *EMNLP-IJCNLP 2019*. 1834–1844.
- [40] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view Response Selection for Human-Computer Conversation. In *EMNLP*.